Andreas König Andreas Dengel
Knut Hinkelmann Koichi Kise
Robert J. Howlett Lakhmi C. Jain (Eds.)

LNAI 6882

Knowledge-Based and Intelligent Information and Engineering Systems

15th International Conference, KES 2011
Kaiserslautern, Germany, September 2011
Proceedings, Part II

2 Part II

Springer

Lecture Notes in Artificial Intelligence 6882

Subseries of Lecture Notes in Computer Science

Andreas König Andreas Dengel
Knut Hinkelmann Koichi Kise
Robert J. Howlett Lakhmi C. Jain (Eds.)

Knowlege-Based and Intelligent Information and Engineering Systems

15th International Conference, KES 2011
Kaiserslautern, Germany, September 12-14, 2011
Proceedings, Part II

Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Andreas König
University of Kaiserslautern, Germany
E-mail: koenig@eit.uni-kl.de

Andreas Dengel
DFKI and University of Kaiserslautern, Germany
E-mail: andreas.dengel@dfki.de

Knut Hinkelmann
University of Applied Sciences Northwestern Switzerland, Olten, Switzerland
E-mail: knut.hinkelmann@fhnw.ch

Koichi Kise
Osaka Prefecture University, Osaka, Japan
E-mail: kise@cs.osakafu-u.ac.jp

Robert J. Howlett
KES International, Shoreham-by-sea, UK
E-mail: rjhowlett@kesinternational.org

Lakhmi C. Jain
University of South Australia, Adelaide, SA, Australia
E-mail: lakhmi.jain@unisa.edu.au

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-23862-8 e-ISBN 978-3-642-23863-5
DOI 10.1007/978-3-642-23863-5
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011935629

CR Subject Classification (1998): I.2, H.4, H.3, I.4-5, H.5, C.2, H.2.8

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The 15^{th} International Conference on Knowledge-Based and Intelligent Information and Engineering Systems was held during September 12–14, 2011 in Kaiserslautern, Germany. The conference was hosted by the University of Kaiserslautern and the German Research Center for Artificial Intelligence (DFKI) GmbH, Germany, and KES International.

KES 2011 provided a scientific forum for the presentation of the results of high-quality international research including recent results of large-scale projects, new exciting techniques, and models, as well as innovative solutions in challenging application fields. The conference attracted contributions from 32 countries and 5 continents: Australia, Canada, China, Colombia, Croatia, Czech Republic, Finland, France, Germany, Greece, Indonesia, Iran, Italy, Japan, Jordan, Korea, Latvia, Malaysia, Mexico, Norway, Poland, Romania, Russia, Spain, Sweden, Switzerland, Taiwan, Thailand, Tunisia, Turkey, UK, and USA.

The conference consisted of 6 keynote talks, 9 general tracks and 25 invited sessions and workshops, on the advance and application of knowledge-based and intelligent systems and related areas. The distinguished keynote speakers were:

Ansgar Bernardi
German Research Center for Artificial Intelligence, Kaiserslautern, Germany
"Growing Together: Opening the Way for Comprehensive Public–Private Knowledge Management"

Knut Manske
Vice President SAP Research, SAP AG, Darmstadt, Germany
"Future Urban Management: Towards Best Managed Cities"

Nikhil R. Pal
Indian Statistical Institute, Calcutta, India
"Selection of Useful Sensors/Features with Controlled Redundancy Using Neural Networks"

Peter Schütt
Leader Software Strategy & Knowledge Management, Executive Engagement Manager, IBM Software Group Germany
"Knowledge Sharing in Enterprise Networks"

Ulrich Reimer
Institute for Information and Process Management University of Applied Sciences St. Gallen, Switzerland
"(Meta-) Modeling of Process-Oriented Information Systems"

Keiji Yamada
General Research Manager, C&C innovation Laboratories, NEC Corporation
Professor, Nara Institute of Science and Technology

"Symbiotic System as a New Social Infrastructure Based on Intelligent Interaction Among the Society, Human Beings, and Information Systems"

Overall 244 oral presentations, complemented by focused lab tours at the organizing institutions, provided excellent opportunities for the presentation of intriguing new research results and vivid discussion on these, paving the way to efficient knowledge transfer and the incubation of new ideas and concepts.

As in the previous years, extended versions of selected papers were considered for publication in follow-up journal publications.

We would like to acknowledge the contribution of the Track Chairs, Invited Sessions Chairs, all members of the Program Committee and external reviewers for coordinating and monitoring the review process. We are grateful to the editorial team of Springer led by Alfred Hofmann. Our sincere gratitude goes to all participants and the authors of the submitted papers.

September 2011

Andreas Dengel
Andreas König
Koichi Kise
Knut Hinkelmann
Robert Howlett
Lakhmi Jain

Organization

KES 2011 was hosted and organized by the Chair's Knowledge-Based Systems, Computer Science department, and Integrated Sensor Systems, Electrical and Computer Engineering department at the University of Kaiserslautern, the German Research Center for Artificial Intelligence (DFKI) GmbH, Germany, and KES International. The conference was held at the University of Kaiserslautern, September 12–14, 2011.

Executive Committee

General Co-chairs

Andreas Dengel	University of Kaiserslautern and DFKI GmbH, Germany
Andreas König	University of Kaiserslautern, Germany
Lakhmi Jain	University of South Australia, Australia

Executive Chair

Robert Howlett	Bournemouth University, UK

Program Co-chairs

Knut Hinkelmann	University of Applied Sciences Northwestern Switzerland, Switzerland
Koichi Kise	Osaka Prefecture University, Japan

Organizing Committee Chair

Stefan Zinsmeister	DFKI GmbH, Germany

Organizing Committee

KES Operations Manager

Peter Cushion	KES International, UK

KES Systems Support

Shaun Lee	KES International, UK

ISE Support Staff

Abhaya Chandra Kammara	University of Kaiserslautern, Germany
Shubhmoy Kumar	University of Kaiserslautern, Germany

Track Chairs

Bruno Apolloni	University of Milan, Italy
Floriana Esposito	University of Bari, Italy
Anne Håkansson	Stockholm University, Sweden
Ron Hartung	Franklyn University, USA
Honghai Liu	University of Portsmouth, UK
Heiko Maus	DFKI GmbH, Germany
Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland
Andreas Nuernberger	University of Magdeburg, Germany
Tuan Pham	University of New South Wales, Australia
Toyohide Watanabe	Nagoya University, Japan

Invited Session Chairs

The Second International Workshop on Natural Language Visualization

Minhua Ma	The Glasgow School of Art, UK
Bob Coyne	Columbia University, USA

Workshop on Seamless Integration of Semantic Technologies in Computer-Supported Office Work (SISTCOW)

Oleg Rostanin	DFKI GmbH, Germany
Simon Scerri	University of Ireland, Galway, Ireland
Benedikt Schmidt	SAP Research, Germany

Innovations in Chance Discovery

Akinori Abe	University of Tokyo, Japan
Yukio Ohsawa	The University of Tokyo, Japan

Computational Intelligence Methods to Benefit Society

Valentina Balas	Aurel Vlaicu University of Arad, Romania
Lakhmi C. Jain	University of South Australia, Australia

Knowledge-Based Interface Systems (I)

Yuji Iwahori	Chubu University, Japan
Naohiro Ishii	Aichi Institute of Technology, Japan

Advances in Theory and Application of Hybrid Intelligent Systems

Lakhmi C. Jain	University of South Australia, Australia
CP Lim	Universiti Sains Malaysia, Malaysia

Recent Trends in Knowledge Engineering, Smart Systems and Their Applications

Cesar Sanin	University of Newcastle, Australia
Carlos Toro	VICOMTech, Spain

Data Mining and Service Science for Innovation

Katsutoshi Yada	Kansai University, Japan

Methods and Techniques of Artificial and Computational Intelligence in Economics, Finance and Decision Making

Marina Resta	DIEM sezione di Matematica Finanziaria, Italy

Human-Oriented Learning Technology and Learning Support Environment

Toyohide Watanabe	Nagoya University, Japan
Tomoko Kojiri	Nagoya University, Japan

Human Activity Support in Knowledge Society

Toyohide Watanabe	Nagoya University, Japan
Takeshi Ushiama	Kyushu University, Japan

Design of Social Intelligence and Creativity Environment

Toyohide Watanabe	Nagoya University, Japan
Naoto Mukai	Tokyo University of Science, Japan

Knowledge Engineering Applications in Process Systems and Plant Operations

Kazuhiro Takeda	Shizuoka University, Japan
Takashi Hamaguchi	Nagoya Institute of Technology, Japan
Tetsuo Fuchino	Tokyo Institute of Technology, Japan

Knowledge - Based Interface Systems (II)

Yoshinori Adachi	Chubu University, Japan
Nobuhiro Inuzuka	Nagoya Institute of Technology, Japan

Emergent Intelligent Technologies in Multimedia Information Processing (IMIP)

Giovanna Castellano	University of Bari, Italy
Maria Alessandra Torsello	University of Bari, Italy

Time Series Prediction Based on Fuzzy and Neural Networks

Minvydas Ragulskis	Kaunas University of Technology, Lithuania

Management Technologies from the Perspective of Kansei Engineering and Emotion

Junzo Watada	Waseda University, Japan
Hisao Shiizuka	Kogakuin University, Japan
Taki Kanda	Bunri University of Hospitality, Japan

Knowledge-Based Systems for e-Business

Kazuhiko Tsuda	University of Tsukuba, Japan
Nubuo Suzuki	KDDI Corporation, Japan

Reasoning Based Intelligent Systems (RIS)

Kazumi Nakamatsu	University of Hyogo, Japan
Jair Minoro Abe	University of Sao Paulo, Brazil

Skill Acquisition and Ubiquitous Human–Computer Interaction

Hirokazu Taki	Wakayama University, Japan
Masato Soga	Wakayama University, Japan

International Session on Sustainable Information Systems

Anne Håkansson	KTH, Sweden
Jason J. Jung	Yeungnam University , Korea
Costin Badica	University of Craiova, Romania

Intelligent Network and Service

Jun Munemori	Wakayama University, Japan
Takaya Yuizono	Japan Advanced Institute Science and Technology, Japan

Advances in Theory and Application of Multi-Agent Systems

Bala M. Balachandran	University of Canberra, Australia
Dharmendra Sharma	University of Canberra, Australia

Advanced Design Techniques for Adaptive Hardware and Systems

Sorin Hintea	Technical University of Cluj-Napoca, Romania
Hernando Fernández-Canque	Glasgow Caledonian University, UK
Gabriel Oltean	Technical University of Cluj-Napoca, Romania

Advanced Knowledge-Based Systems

Alfredo Cuzzocrea	ICAR-CNR, University of Calabria, Italy

Computational Intelligence for Fault Diagnosis and Prognosis

Beatrice Lazzerini	University of Pisa, Italy
Marco Cococcioni	University of Pisa, Italy
Sara Lioba Volpi	University of Pisa, Italy

Multiple Classifiers and Hybrid Learning Paradigms

Edmondo Trentin	University of Siena, Italy
Friedhelm Schwenker	University of Ulm, Germany

Soft Computing Techniques and Their Intelligent Utilizations

Norio Baba	Osaka Kyoiku University, Japan
Kunihiro Yamada	Tokai University, Japan

Document Analysis and Knowledge Science

Seiichi Uchida	Kyushu University, Japan
Marcus Liwicki	DFKI GmbH, Germany
Koichi Kise	Osaka Prefecture University, Japan

Model-Based Computing for Innovative Engineering

Klaus Schneider	University of Kaiserslautern, Germany
Norbert Wehn	University of Kaiserslautern, Germany

Immunity-Based Systems

Yoshiteru Ishida	Toyohashi University of Technology, Japan
Andreas König	University of Kaisers-lautern, Germany

Program Committee

Akinori Abe	University of Tokyo, Japan
Jair Minoro Abe	University of Sao Paulo, Brazil
Canicious Abeynayake	DSTO, Australia
Yoshinori Adachi	Chubu University, Japan

Benjamin Adrian	German Research Center for Artificial Intelligence (DFKI), Germany
Plamen Angelov	Lancaster University, UK
Ahmad Taher Azar	Modern Science and Arts University (MSA), Egypt
Norio Baba	Osaka Kyoiku University, Japan
Costin Badica	University of Craiova , Romania
Bala Balachandran	University of Canberra, Australia
Valentina Balas	Aurel Vlaicu University of Arad, Romania
Vivek Bannore	University of South Australia, Australia
Adrian S. Barb	Penn State University, USA
Ansgar Bernardi	German Research Center for Artificial Intelligence (DFKI), Germany
Monica Bianchini	University of Siena, Italy
Isabelle Bichindaritz	University of Washington, USA
Veselka Boeva	Technical University of Sofia, Bulgaria
Christopher Buckingham	Aston University, UK
Giovanna Castellano	University of Bari, Italy
Barbara Catania	Università degli Studi di Genova, Italy
Michele Ceccarelli	University of Sannio, Italy
Javaan Chahl	DSTO, Australia
Stephan Chalup	The University of Newcastle, Australia
Chien-Fu Cheng	Tamkang University, Taiwan
Kai Cheng	Brunel University, UK
Benny Cheung	Honk Kong Polytechnic University, Hong Kong
Marco Cococcioni	University of Pisa, Italy
Bob Coyne	Columbia University, USA
Paolo Crippa	Università Politecnica delle Marche, Italy
Mary (Missy) Cummings	Massachusetts Institute of Technology, USA
Alfredo Cuzzocrea	ICAR-CNR & University of Calabria , Italy
Ernesto Damiani	Università degli Studi di Milano, Italy
Stamatia Dasiopoulou	Informatics and Telematics Institute, Greece
Martine De Cock	University of Washington Tacoma, USA
Philippe De Wilde	Heriot-Watt University, UK
Argyris Dentsoras	University of Patras, Greece
Liya Ding	Macau University of Science and Technology, Hong Kong
Richard J. Duro	Universidade da Coruña, Spain
Schahram Dustdar	Vienna University of Technology, Austria
Isao Echizen	National Institute of Informatics, Japan
Tapio Elomaa	Tampere University of Technology, Finland
Hernando Fernandez-Canque	Glasgow Caledonian University, UK
Ana Fernandez-Vilas	University of Vigo, Spain
Arthur Filippidis	DSTO, Australia
Tetsuo Fuchino	Tokyo Institute of Technology, Japan

Junbin Charles Gao	Sturt University, Australia
Petia Georgieva	University of Aveiro, Portugal
Daniela Godoy	UNICEN University, Argentina
Bernard Grabot	LGP-ENIT, France
Manuel Graña Romay	Universidad del Pais Vasco, Spain
Christos Grecos	University of West Scotland, UK
Anne Hakånsson	KTH, Sweden
Takashi Hamaguchi	Nagoya Institute of Technology, Japan
Alex Hariz	University of South Australia, Australia
Mohamed Hassan	Cairo University, Egypt
Richard Hill	University of Derby, UK
Sorin Hintea	Technical University of Cluj-Napoca, Romania
Dawn Holmes	University of California, USA
Katsuhiro Honda	Osaka Prefecture University, Japan
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Eyke Hullermeier	Philipps-Universität Marburg, Germany
Nikhil Ichalkaranje	University of Mumbai, India
Nobuhiro Inuzuka	Nagoya Institute of Technology, Japan
Naohiro Ishii	Aichi Institute of Technology, Japan
Takayuki Ito	Massachusetts Institute of Technology, USA
Yuji Iwahori	Chubu University, Japan
Norbert Jastroch	MET Communications GmbH, Germany
Richard Jensen	Aberystwyth University, UK
Andrew Jones	Cardiff University, UK
Jason J. Jung	Yeungnam University, Korea
Taki Kanda	Bunri University of Hospitality, Japan
Anastasia Kastania	Athens University of Economics and Business, Greece
Hideki Katagiri	Hiroshima University, Japan
Koichi Kise	Osaka Prefecture University, Japan
In-Young Ko	KAIST, Korea
Vassilis S. Kodogiannis	University of Westminster, UK
Tomoko Kojiri	Nagoya University, Japan
Amit Konar	Jadavpur University, India
Ivan Koychev	University of Sofia, Bulgaria
Halina Kwasnicka	Wroclaw University of Technology, Poland
C.K. Kwong	The Hong Kong Polytechnic University, Hong Kong
Beatrice Lazzerini	University of Pisa, Italy
Dah-Jye Lee	Brigham Young University, USA
CP Lim	Universiti Sains Malaysia, Malaysia
Tsung-Chih Lin	Feng-Chia University, Taiwan
James Liu	The Hong Kong Polytechnic University, Hong Kong
Lei Liu	Beijing University of Technology, China

Marcus Liwicki	German Research Center for Artificial Intelligence (DFKI), Germany
Ignac Lovrek	University of Zagreb, Croatia
Jie Lu	University of Technology, Sydney, Australia
Minhua Eunice Ma	University of Derby, UK
Ilias Maglogiannis	University of Central Greece, Greece
Nadia Magnenat-Thalmann	University of Geneva, Switzerland
Dario Malchiodi	Università degli Studi di Milano, Italy
Milko T. Marinov	University of Ruse, Bulgaria
Mia Markey	The University of Texas at Austin, USA
Maja Matijasevic	University of Zagreb, Croatia
Rashid Mehmood	School of Engineering, Swansea, UK
Stefania Montani	Università del Piemonte Orientale, Italy
Ramón Moreno Jimenez	Universidad del Pais Vasco, Spain
Naoto Mukai	Tokyo University of Science, Japan
Christine Mumford	Cardiff University, UK
Jun Munemori	Wakayama University, Japan
Hirofumi Nagashino	The University of Tokushima, Japan
Kazumi Nakamatsu	University of Hyogo, Japan
Zorica Nedic	University of South Australia, Australia
Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland
Vesa A. Niskanen	University of Helsinki, Finland
Lidia Ogiela	AGH & University of Science and Technology, Poland
Yukio Ohsawa	The University of Tokyo, Japan
Gabriel Oltean	Technical University of Cluj-Napoca, Romania
Vasile Palade	Oxford University, UK
Gabriella Pasi	Università degli Studi di Milano Bicocca, Italy
Kunal Patel	Ingenuity Systems, USA
Jose Pazos-Arias	University of Vigo, Spain
Carlos Pedrinaci	The Open University, UK
Alfredo Petrosino	Università di Napoli Parthenope, Italy
Dilip Pratihar	Indian Institute of Technology, India
Goran D. Putnik	University of Minho, Portugal
Minvydas Ragulskis	Kaunas University of Technology, Lithuania
Elisabeth Rakus-Andersson	Blekinge Institute of Technology, Sweden
Nancy Reed	University of Hawaii , USA
Paolo Remagnino	Kingston University, UK
Marina Resta	DIEM sezione di Matematica Finanziaria, Italy
Oleg Rostanin	German Research Center for Artificial Intelligence (DFKI), Germany
Asit Saha	Central State University, USA
Ziad Salem	Aleppo University, Syria
Cesar Sanin	University of Newcastle, Australia
Carlo Sansone	Università di Napoli Federico II, Italy

Mika Sato-Ilic	University of Tsukuba, Japan
Simon Scerri	University of Ireland Galway, Ireland
Benedikt Schmidt	SAP Research, Germany
Klaus Schneider	University of Kaiserslautern, Germany
Steven Schockaert	Ghent University, Belgium
Friedhelm Schwenker	University of Ulm, Germany
Udo Seiffert	Fraunhofer Institute IFF Magdeburg, Germany
Dharmendra Sharma	University of Canberra, Australia
Hisao Shiizuka	Kogakuin University, Japan
Christos Sioutis	DSTO, Australia
Masato Soga	Wakayama University, Japan
Margarita Sordo	Harvard University, USA
Anthony Soroka	Cardiff University, UK
Myra Spiliopoulou	Otto-von-Guericke-Universität, Germany
Dipti Srinivasan	National University of Singapore, Singapore
Jadranka Sunde	DSTO, Australia
Nobuo Suzuki	KDDI Corporation , Japan
Edward Szczerbicki	The University of Newcastle, Australia
Kazuhiro Takeda	Shizuoka University, Japan
Hirokazu Taki	Wakayama University, Japan
Tatiana Tambouratzis	University of Piraeus, Greece
Pavel Tichy	Rockwell Automation Research Centre, Czech Republic
Peter Tino	The University of Birmingham, UK
Carlos Toro	VICOMTech, Spain
Maria Torsello	University of Bari, Italy
Edmondo Trentin	University of Siena, Italy
George A. Tsihrintzis	University of Piraeus, Greece
Kazuhiko Tsuda	University of Tsukuba, Japan
Jeffrey Tweedale	University of South Australia, Australia
Seiichi Uchida	Kyushu University, Japan
Eiji Uchino	Yamaguchi University, Japan
Taketoshi Ushiama	Kyushu University, Japan
Sunil Vadera	University of Salford, UK
Annamaria Varkonyi Koczy	Obuda University, Hungary
István Vassányi	University of Pannonia, Hungary
Alfredo Vellido	Universitat Politècnica de Catalunya, Spain
Juan D. Velásquez	University of Chile, Chile
Maria Virvou	University of Piraeus, Greece
Sara Volpi	University of Pisa, Italy
Junzo Watada	Waseda University, Japan
Toyohide Watanabe	Nagoya University, Japan
Rosina Weber	The iSchool at Drexel, USA
Norbert Wehn	University of Kaiserslautern, Germany
Richard J. White	Cardiff University, UK

M. Howard Williams	Heriot-Watt University, UK
Katsutoshi Yada	Kansai University, Japan
Kunihiro Yamada	Tokai University, Japan
Zijiang Yang	York University, Canada
Hiroyuki Yoshida	Harvard Medical School, USA
Jane You	The Hong Kong Polytechnic University, Hong Kong
Takaya Yuizono	JAIST, Japan
Cecilia Zanni-Merk	LGeCo - INSA de Strasbourg, France

Sponsoring Institutions

Center for Computational and Mathematical Modeling $(CM)^2$, University of Kaiserslautern, Germany

German Research Center for Artificial Intelligence (DFKI) GmbH, Kaiserslautern, Germany

Institute of Integrated Sensor Systems, University of Kaiserslautern, Germany

Table of Contents – Part II

Web Intelligence, Text and Multimedia Mining and Retrieval

Autonomous and Adaptive Identification of Topics in Unstructured Text 1
Louis Massey

Outlier-Based Approaches for Intrinsic and External Plagiarism Detection 11
Gabriel Oberreuter, Gaston L'Huillier, Sebastián A. Ríos, and Juan D. Velásquez

An Extended Method for Finding Related Web Pages with Focused Crawling Techniques 21
Kazutaka Furuse, Hiroaki Ohmura, Hanxiong Chen, and Hiroyuki Kitagawa

Development of Multilingual Interview-Sheet Composition System to Support Multilingual Communication in Medical Field 31
Taku Fukushima, Takashi Yoshino, and Aguri Shigeno

User Modeling-Based Spatial Web Personalization 41
Hadjouni Myriam, Baazaoui Hajer, Aufaure Marie Aude, and Ben Ghezala Henda

Context-Aware Website Personalization 51
Daniela Wolff, Marc Schaaf, Stella Gatziu Grivas, and Uwe Leimstoll

Node-First Causal Network Extraction for Trend Analysis Based on Web Mining 63
Hideki Kawai, Katsumi Tanaka, Kazuo Kunieda, and Keiji Yamada

Consumer Behavior Analysis from Buzz Marketing Sites over Time Series Concept Graphs 73
Tetsuji Kuboyama, Takako Hashimoto, and Yukari Shirota

Fuzzy Image Labeling by Partially Supervised Shape Clustering 84
G. Castellano, A.M. Fanelli, and M.A. Torsello

Intelligent Tutoring Systems and E-Learning Environments

Intelligent E-Learning System for Training Power Systems Operators . . . 94
Liliana Argotte, Yasmin Hernandez, and G. Arroyo-Figueroa

Interaction Based on Contribution Awareness in Collaborative Learning 104
Yuki Hayashi, Tomoko Kojiri, and Toyohide Watanabe

The MATHESIS Semantic Authoring Framework: Ontology-Driven Knowledge Engineering for ITS Authoring 114
Dimitrios Sklavakis and Ioannis Refanidis

An Intelligent Tutoring System Architecture for Competency-Based Learning 124
Miguel Badaracco and Luis Martínez

A Majority Density Approach for Developing Testing and Diagnostic Systems 134
Dechawut Wanichsan, Patcharin Panjaburee, Parames Laosinchai, and Sasithorn Chookaew

Exploiting Learners' Tendencies for Detecting English Determiner Errors 144
Ryo Nagata and Atsuo Kawai

Analysis of Students' Learning Activities through Quantifying Time-Series Comments 154
Kazumasa Goda and Tsunenori Mine

Back-Review Support Method for Presentation Rehearsal Support System 165
Ryo Okamoto and Akihiro Kashihara

Other / Misc. Intelligent Systems Topics

Multiple Hypothesis Testing and Quasi Essential Graph for Comparing Two Sets of Bayesian Networks 176
Hoai-Tuong Nguyen, Philippe Leray, and Gérard Ramstein

Financial Performance Analysis of European Banks Using a Fuzzified Self-Organizing Map 186
Peter Sarlin and Tomas Eklund

Self-Organizing Map in Process Visualization 196
Miki Sirola and Jaakko Talonen

Kalman Filter vs. Particle Filter in Improving K-NN Indoor Positioning 203
Jaegeol Yim, Jinseog Kim, Gyeyoung Lee, and Kyubark Shim

On Possibility of Conditional Invariant Detection 214
Hani Fouladgar, Behrouz Minaei-Bidgoli, and Hamid Parvin

On Temporal Gödel-Gentzen Translation 225
Norihiro Kamide

A Decidable First-Order Logic for Medical Reasoning 235
Norihiro Kamide

Interpolation Theorems for Some Extended Description Logics 246
Norihiro Kamide

Investigating Attachment Behavior of Nodes during Evolution of a Complex Social Network: A Case of a Scientific Collaboration Network 256
Alireza Abbasi and Liaquat Hossain

Activity Recognition in Healthcare Monitoring Systems Using Wireless Sensor Networks 265
Rodica-Elena Doran and Ioana-Iuliana Farkas

Defining Events as a Foundation of an Event Notification Middleware for the Cloud Ecosystem 275
Rolf Sauter, Alex Stratz, Stella Gatziu Grivas, Marc Schaaf, and Arne Koschel

Effective Content-Based Music Retrieval with Pattern-Based Relevance Feedback 285
Ja-Hwung Su, Tzu-Shiang Hung, Chun-Jen Lee, Chung-Li Lu, Wei-Lun Chang, and Vincent S. Tseng

Meta-context: Putting Context-Awareness into Context 296
Ramón Hervás, Jesús Fontecha, Vladimir Villarreal, and Jose Bravo

Negotiation in Electronic Commerce: A Study in the Latin-American Market 306
Osvaldo Cairo, Juan Gabriel Olarte, and Fernando Rivera

Adjusted Case-Based Software Effort Estimation Using Bees Optimization Algorithm 315
Mohammad Azzeh

A New Honeybee Optimization for Constraint Reasoning: Case of Max-CSPs 325
Ines Methlouthi and Sadok Bouamama

Hybrid Virtual Sensor Based on RBFN or SVR Compared for an Embedded Application 335
Kuncup Iswandy and Andreas König

Skyline Adaptive Fuzzy Query 345
Wei Yan, Cecilia Zanni-Merk, and Francois Rousselot

Improved Travel Time Prediction Algorithms for Intelligent Transportation Systems 355
Nihad K. Chowdhury and Carson K.-S. Leung

Methods and Techniques of Artificial and Computational Intelligence in Economics, Finance and Decision Making

An Effective Utilization of Many Neural Networks for Improving the Traditional Technical Analysis in the Stock Market 366
Norio Baba, Kokutan Liu, Lee Chen Han, Takao Mitsuda, Kou Ro, and Kou Ninn

A New Framework for Assets Selection Based on Dimensions Reduction Techniques 372
Marina Resta

Forecasting Stock Price Based on Fuzzy Time-Series with Entropy-Based Discretization Partitioning 382
Bo-Tsuen Chen, Mu-Yen Chen, Hsiu-Sen Chiang, and Chia-Chen Chen

On the Use of Feed-Forward Neural Networks to Discriminate between Models in Financial and Insurance Risk Frameworks 392
Enrico di Bella

Workshop on Seamless Integration of Semantic Technologies in Computer-Supported Office Work (SISTCOW)

Interaction Data Management 402
Benedikt Schmidt and Eicke Godehardt

SWord: Semantic Annotations Revisited 410
Oleg Rostanin and Passant al Agroudy

Using Suffix Arrays for Efficiently Recognition of Named Entities in Large Scale 420
Benjamin Adrian and Sven Schwarz

Extracting Personal Concepts from Users' Emails to Initialize Their Personal Information Models 430
Sven Schwarz, Frank Marmann, and Heiko Maus

Innovations in Chance Discovery

A Diagrammatic Approach to Discovering Chances in Team Relationships 440
Ruediger Oehlmann and Balpreet Gill

Ontology-Based Knowledge Navigation Platform for Intelligent Manufacturing 447
Reiko Fujiwara, Akira Kitamura, and Kouji Mutoh

Finding Top-N Chance Patterns with KeyGraph®-Based Importance . . . 457
Yoshiaki Okubo, Makoto Haraguchi, and Sachio Hirokawa

Quantitative Evaluation Method of Criticism in Value Creating Conversation 469
Yoko Nishihara and Yukio Ohsawa

Chance Discovery and Unification in Linear Modal Logic 478
Vladimir V. Rybakov

From Epistemic Luck to Chance-Seeking: The Role of Cognitive Niche Construction 486
Lorenzo Magnani and Emanuele Bardone

Relation between Chance Discovery and Black Swan Awareness 495
Akinori Abe

Advanced Knowledge-Based Systems

On Kernel Information Propagation for Tag Clustering in Social Annotation Systems 505
Guandong Xu, Yu Zong, Rong Pan, Peter Dolog, and Ping Jin

An Efficient Itemset Mining Approach for Data Streams 515
Elena Baralis, Tania Cerquitelli, Silvia Chiusano, Alberto Grand, and Luigi Grimaudo

The Representation of Inconsistent Knowledge in Advanced Knowledge Based Systems 524
Mark Burgin and Kees (C.N.J.) de Vey Mestdagh

Contextual Ontology Module Learning from Web Snippets and Past User Queries 538
Nesrine Ben Mustapha, Marie-Aude Aufaure, Hajer Baazaoui Zghal, and Henda Ben Ghezala

Inconsistency-Tolerant Integrity Checking Based on Inconsistency Metrics 548
Hendrik Decker

OLAP over Continuous Domains via Density-Based Hierarchical Clustering 559
Michelangelo Ceci, Alfredo Cuzzocrea, and Donato Malerba

Non-separable Transforms for Clustering Trajectories 571
Alfredo Cuzzocrea and Elio Masciari

Recent Trends in Knowledge Engineering, Smart Systems, and Their Applications

Knowledge Discovery about Web Performance with Geostatistical Turning Bands Method 581
Leszek Borzemski and Anna Kamińska-Chuchmała

Context Change Detection for Resource Allocation in Service-Oriented Systems 591
Piotr Rygielski and Jakub M. Tomczak

A Multi-Agent Approach for Engineering Design Knowledge Modelling 601
Ricardo Mejía-Gutiérrez, Alejandro Cálad-Álvarez, and Santiago Ruiz-Arenas

An Architecture for the Semantic Enhancement of Clinical Decision Support Systems 611
Eider Sanchez, Carlos Toro, Eduardo Carrasco, Gloria Bueno, Carlos Parra, Patricia Bonachela, Manuel Graña, and Frank Guijarro

An Approach to Measure Quality of Knowledge in the e-Decisional Community 621
Leonardo Mancilla-Amaya, Cesar Sanín, and Edward Szczerbicki

Application of Decisional DNA in Web Data Mining 631
Peng Wang, Cesar Sanín, and Edward Szczerbicki

A Concept for Comprehensive Knowledge Management System 640
Bartosz Kucharski and Edward Szczerbicki

The Role and Concept of Sub-models in the Smart Fuzzy Model of the Internet Mortgage Market 650
Aleksander Orłowski and Edward Szczerbicki

Knowledge Management Challenges in Collaborative Design of a Virtual Call Centre 657
Marcin Sikorski, Igor Garnik, Bohdan Ludwiszewski, and Jan Wyrwiński

Decisional DNA Applied to Digital TV 667
Haoxi Zhang, Cesar Sanín, and Edward Szczerbicki

Measurement of the Development of a Learning IT Organization Supported by a Model of Knowledge Acquisition and Processing 677
Cezary Orłowski and Tomasz Sitek

Prediction Based Handovers for Wireless Networks Resources Management 687
Piotr Rygielski, Paweł Świątek, Krzysztof Juszczyszyn, and Adam Grzech

Author Index 697

Autonomous and Adaptive Identification of Topics in Unstructured Text

Louis Massey

Department of Mathematics and Computer Science,
Royal Military College, Kingston, Canada, K7K 7B4
massey@rmc.ca

Abstract. Existing topic identification techniques must tackle an important problem: they depend on human intervention, thus incurring major preparation costs and lacking operational flexibility when facing novelty. To resolve this issue, we propose an adaptable and autonomous algorithm that discovers topics in unstructured text documents. The algorithm is based on principles that differ from existing natural language processing and artificial intelligence techniques. These principles involve the retrieval, activation and decay of general-purpose lexical knowledge, inspired by how the brain may process information when someone reads. The algorithm handles words sequentially in a single document, contrary to the usual corpus-based bag-of-words approach. Empirical results demonstrate the potential of the new algorithm.

Keywords: Text Analysis, Information Retrieval, Knowledge Management, Topics Identification, Text Mining, Text Clustering, Document Classification, Topics Modeling.

1 Introduction

With the continually increasing amount of human knowledge available as electronic text, the design of algorithms capable of determining what a text document is about has been an important research area for years. Despite important progress, computers still have major problems making sense of text. On one hand, the problem originates from the ambiguous nature of text and computers inability to deal with this ambiguity. On the other hand, knowing what a text document is about requires large amounts of difficult and costly to assemble knowledge. The efforts currently deployed to build the Semantic Web illustrate the scale of human intervention required to acquire the knowledge enabling intelligent search and access to text information on the Internet.

We investigate the task of finding the main topics present in a document. The word topics is taken in its intuitive sense of what a text is about, as identified by a few summarizing and evocative keywords. These keywords are not necessarily present in the text under analysis. We particularly focus on making the topic identification task deal with text ambiguity in an autonomous and adaptive manner. We contribute new fundamental principles that achieve this aim. The principles differ from existing techniques and are inspired by the way humans may be processing information when

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 1–10, 2011.

they read. We named these principles ReAD, which stands for REtrieval, Activation and Decay. The main idea is that as words are read sequentially, they activate regions of the brain that contain information related to each word. This activation augments as more words accessing the same regions are read, but also decays when regions are rarely accessed. At the end of the document, the strongest activated regions indicate what the text was about.

The innovation stems from the ReAD principles that allow for the handling of text ambiguity to determine topics in an adaptive and autonomous manner. Indeed, the approach we propose avoids both the issues of human intervention and of dependence on inflexible corpus-based training that exist with current techniques. We demonstrate an algorithmic implementation based on the ReAD principles and show experimentally that the algorithm can eliminate the problem of ambiguity and capture the general meaning – the topics - of a text document. The algorithm emulates the ReAD cognitive process using only lexical information stored in a dictionary.

The paper is organized as follows: section 2 contains an overview of related work and identifies the problems with existing topics identification methods. In section 3, we introduce the ReAD principles for topics identification, while in section 4 an algorithmic implementation of these principles is described. In section 5, we report on empirical evaluations of the algorithm and discuss the results obtained.

2 Related Work

The topics of documents are often sought in co-occurrence information found in text corpora. Applications of this idea can be found in methods like Latent Dirichlet Allocation (LDA) [1] and Latent Semantic Analysis (LSA) [2, 3]. The problems with these methods include a dependence on the availability of large domain corpora usually assembled at high cost by humans and a lack of adaptability since topics are determined based on a static snapshot of the data present in the corpus during training.

Other techniques to identify the topics of documents based on a large set of documents are document classification [4, 5] and text clustering [6, 7]. Document classification involves the application of supervised machine learning techniques to settle on a set of rules acquired from the recognition and generalization of patterns in training samples. There are two main problems with classification. First, human experts must initially expand much effort to acquire and prepare a set of training documents labeled with the correct topics. Second, once the classifier has been learned, they often do not fare well in the face of novelty (for example, given a new topic) and must be retrained. Contrary to document classification, text clustering operates without training: it aims at grouping documents based on the similarity of their content. Clustering is hence more autonomous and adaptable than classification, but is generally less reliable [8].

An additional problem with existing topic identification techniques is that documents are usually represented in computer memory as high dimensional vectors using the bag-of-words model of text [9]. The value of each numerical component is based on statistical information obtained from the overall vocabulary from a training text collection, using weighting schemes such as TFIDF [10]. Under this representation, the compositional construction of meaning from the order of words

within a particular document is ignored and terms considered unrelated. However, sentences use word order and related terms to unify text and provide information to the reader about the main topics of the document [11]. Some researchers have investigated the use of knowledge to enrich document vectors semantically [12,13], partly to address this problem. This solution however does not address the critical loss of document word order and furthermore does nothing to resolve issues of low autonomy and adaptability arising from the dependence on a training set to build the bag-of-words representation.

Another approach is to harness natural language processing (NLP) techniques to adequately determine the underlying meaning of text. NLP can also possibly be exploited to improve the effectiveness of classification and clustering [14]. NLP techniques aim at deriving meaning of a particular document by performing, among other steps, syntactic and semantic analysis. The former requires that the proper set of grammatical rules be coded to isolate the syntactic structure of the document sentences. The latter is based on large quantities of encyclopedic knowledge [15] to infer the meaning of text. The preparation of syntactic rules and of encyclopedic knowledge is a costly and lengthy endeavor. For this reason, traditional NLP tends to work only in restricted domains and does not scale well to real-life applications.

Also related to our work is automated semantic tagging and cross-referencing [16, 17]. Again, there is a strong dependence on human intervention and knowledge acquisition. Keyword extraction (also known as term recognition or automatic indexing) [18-20] is another research area that aims at finding topics. In this case, one aims at extracting keywords from documents usually with some form of frequency-based measure of word importance within a document or across a corpus [10, 21-22]. One thus obtains the final product (keywords) for a search or text mining task, or alternatively the keywords extracted can be used as features for further processing such as supervised learning or clustering. Either way, frequency-based information has its limits in its ability to identify the most discriminant features for learning or the most evocative keywords for direct consumption. For this reason, additional sources of evidences are investigated by many researchers (for e.g., [12-13], [23-24]). These include ontologies or other knowledge bases that provide background knowledge to reason about or at least establish basic relationships between words and the concepts they represent; domain corpora for extracting statistical information deemed to carry information on conceptual relationships between words; and, NLP to establish the syntactic structure of the text and determine its semantic nature. Clearly, the same problems of costly human intervention and lack of adaptability arise again.

The specific problem we aim at solving here can be summarized as follows: existing techniques to identify topics depend on human interventions to acquire knowledge and to handcraft training sets. The training set is used to build a bag-of-word representation of documents, based on the assumption that the documents in it are representative of future circumstances. The need to handcraft knowledge makes existing topics identification systems costly to develop, while the dependence on training sets makes them little adaptive to new situations. Our objective is to solve these problems with an approach based on fundamentally different principles than those used by existing methods.

3 The ReAD Principles

To address the issues just described, we propose an approach that differs from traditional keyword extraction, corpus-based co-occurrence analysis, classification, clustering and NLP. The approach is intuitively inspired by how humans may determine the topics of a text document when they read [25, 26]. The general idea is as follows: When one reads a document, it can be expected that each word read sequentially will trigger multiple neural activations corresponding to memory areas of the brain associated with the word in question. Over time, neural activation decays, unless subsequent words incrementally activate the same regions. At the end of the document, the concepts associated with the most activated regions in the reader's brain constitute the topics of the document.

The fundamentally important principles at play are first that words read sequentially cause the retrieval and activation of knowledge items associated with each word. Second, that accrued activation of overlapping memory areas where word-related knowledge is stored accumulates over time, while memory areas infrequently accessed decay. Third, that a convergence onto particular areas will take place as words after words in the text focus onto a few common knowledge items. The hypothesis is that in the end, the interplay of activation and decay will cause just a few areas to be discriminately more activated than others, thus identifying the predominant knowledge items, which can then be interpreted as the main concepts – the topics – present in the text. We call these fundamental principles ReAD for Retreival (of stored knowledge), Activation and Decay.

Assuming usage of existing general-purpose knowledge, an algorithm based on the ReAD principles has the advantage of eliminating many problems common with existing approaches. First, it does not depend on costly and laborious knowledge acquisition efforts. Second, it does not rely on training documents statistics and is able to determine a document's topics in isolation from other documents. Third, the algorithm works in a single pass over the text data, which could benefit real-time applications such as social networks mining and streaming newsfeed analysis (for instance in the context of business intelligence). And lastly, the algorithm abandons the vector representation of documents to take into account word order and the compositional effect of words towards meaning.

4 Algorithmic Implementation

An algorithm based on the ReAD principles is shown in Fig 1. The main steps corresponding to the ReAD principles are as follows: step 3.2.2 is where knowledge related to the word currently under consideration (w_j) is retrieved; steps 3.3.1.1 and 3.3.3.2 are respectively where the activation is incremented and decayed. The knowledge source emulating the brain for retrieval can be any existing source of lexical information (e.g., a dictionary) and doesn't depend on particular representation formalism. The only requirement is that it be able to return a list of words describing or defining the word w_j being queried. Here, as a surrogate for knowledge stored in the human brain, we access WordNet [27], a database that, among other features, describes words and their multiple senses as sets of synonyms. Many researchers

(e.g., [12]) have exploited WordNet as a knowledge source to support text analysis tasks. The way we use WordNet in this work is unique in two ways. First, only unstructured and non-disambiguated lexical information - the list of words for all definitions - is exploited to emulate brain knowledge related to words in the text; and second, the information from WordNet is not used to semantically augment a bag-of-words representation but is rather available directly as candidate topic labels. More precisely, given a word extracted sequentially from the text, the algorithm retrieves the set of synonyms for nouns, as well as a short definition and a sentence showing a sample usage. This is done for all senses of a word that can be interpreted as a noun. There is no attempt at selecting the correct sense as is common in traditional natural language processing with word sense disambiguation techniques [28]. Convergence to meaning and disambiguation are by-products of activation and decay. This is an important attribute of the ReAD principles.

Words retrieved from WordNet are called items to distinguish them from words in the text. Items are filtered to remove words that are deemed useless or too general to precisely identify topics (e.g., thing, entity, time, etc). The list of knowledge items $T_j = \{t_1, t_2, \ldots, t_n\}$ contains items retrieved for word w_j, that is, the non-unique words found in all WordNet senses. Knowledge items emulate the specific regions of the brain that are activated when a word is read. An association between each word w_j and its list of knowledge items T_j is kept in the words table W (step 3.2.4), whereas the various parameters related to the computation of activation of each item are kept in the topics table T (step 3.2.6).

The activation is computed based on a modified version of TFIDF [10]. TFIDF is a common measure of word importance in information retrieval, but we use it here to measure the importance of items related to words within a single document. There is therefore no corpus statistics involved, only word related items statistics within an individual document. The activation α_i of item t_i is the product $\alpha_i = tf_i \times idf_i$ where:

$tf_i = q_i / Q$ (originally in information retrieval, *tf* denotes term frequency, but here it is item frequency) q_i is the number of times item i is retrieved from the knowledge source, and Q is the total number of items retrieved that are not stop words, for all words in the current document.

$idf_i = \log(V/v_i)$ (inverse document frequency) V is the total number of words that are not stop words in the current document and v_i is the number of words that trigger retrieval of item i from the knowledge source.

In the algorithm of Fig 1, the values of Q, V, q_i and v_i used in the calculation of activation are updated in the sub-steps of 3.2.6. The computation of activation is delayed until all items for a word have been seen (at step 3.3.3.1.1) because an item may occur more than once in the set retrieved for a given word and thus may have its q_j value incremented repeatedly. It would therefore be pointless to calculate activation before having collected all counts for a word. The incremental establishment of Q, V, q_i and v_i allows for online processing of words within a document. There are other ways to calculate activation, such as simply counting each item occurrence. We used a TFIDF-inspired approach because it appeared to be a judicious choice for the information retrieval related task of topic identification. We will investigate alternatives in future work.

```
1. inputs: text, knowledge source, list of stop words, max retention
time τmax, decay rate γ and number of topics M.
2.  V=0, Q=0, clear items table T and words table W
3.  while there are words in the text:
3.1   get next word wj
3.2   if wj is not in the stop list:
3.2.1   V++
3.2.2   retrieve information about wj from the knowledge
          source: a list of non unique items Tj={t1,t2,…, tn}
3.2.3   remove stop words from Tj
3.2.4   if wj is not in W: store (wj, Tj, τj = τmax) in words
          table W
3.2.5   else: reset τj = τmax for word wj in words table W
3.2.6   for each ti in Tj
3.2.6.1     Q++
3.2.6.2     if ti is already in T: qi ++  and for first occurrence
              of ti in Tj: vi ++
3.2.6.3     else : qi =1, vi =1 and store (ti, qi, vi) in T

3.3   for each word k in W
3.3.1     if  wk != wj  (not the word just read)
3.3.1.1       τk --
3.3.2     Tk = list of items associated with wk in W
3.3.3     for each item i in list Tk
3.3.3.1       if τk > 0
3.3.3.1.1           αi = αi + ( qi / Q * log(V/vi) )
3.3.3.2       else
3.3.3.2.1           αi = αi - αi / (τmax * γ)
4.  All words have been processed: sort the items in T and output the M
items with highest activation
```

Fig. 1. An algorithm based on the ReAD principles

Words extracted from the document are stored in a table emulating human short-term memory (STM), which is table W in the algorithm (step 3.2.4). A fundamental idea is that STM has a limited capacity, as is the case with humans [29]. Because of limited STM capacity, a word previously read will eventually be pushed out of STM when a new word is extracted from the text. STM can be interpreted as a sliding window considering a few words of the text at a time. The size of the window can be pre-determined or computed. In the algorithm of Fig 1, STM capacity is defined by the maximum retention time parameter, τ_{max}. When a new word is read from the document, retention time for that word is set to this maximal value. The τ_k parameter associated with each word k (or τ_j associated with word j depending on which algorithm loop we are in: 3.2 for initialization or 3.3 for updates) is used to determine when a word is being pushed out of STM. Each time a new word is read, the window slides to the right to include the next word while the : oldest word is expulsed from STM. The size of the window determines the persistence of un-decaying activation for items associated with a specific word. Once a word looses the focus of attention granted by the window, that is, when it is pushed out of STM, the activation of items associated with the word starts to decay (step 3.3.3.2.1). There are other decay formulae possible. At this point we have only evaluated the linear decay shown in the algorithm.

The role of decay is to help distinguish between relevant and non-relevant items. Activation can be seen as a process of amplification of relevant items while decay plays the role of a filter to eliminate semantically unimportant items. γ is the decay rate, a larger value meaning a slower decay. Although an item may be undergoing decay because its associated words have lost the focus of attention, if another word is read that activates this same item, the activation will also accumulate. As a direct consequence, in step 3.3, one can observe that an item's activation changes as many times as there are words associated with that item. Besides, an item's activation may be increased due to one word still being in STM and decreased due to another word because it is not (based on the value of the τ_k parameter associated with each word k).

The last words of the documents cannot decay entirely and are therefore advantaged. The solution we have implemented in the current implementation is to ignore any item fully activated due to the presence of a word in the last window before the end of the document, but other options are also possible and need to be investigated. This action is omitted from the algorithm of Fig. 1. There may be other variants to the algorithm presented. For example, one might be to allow for cascading activations, where retrieved items recursively propagate activation to other items in a way that might be expected in neural networks. As well, the selection of items for output could be modified in various ways, such as for instance with an activation threshold instead of selecting the M most activated ones. These will be tested in future work. Finally, one might argue that the algorithm could be simplified by eliminating everything related to STM and decay. At this point we have not tested this alternative. However, we must point out that such an alternative implementation does not correspond to the inspiration of the human model of reading as embodies in the ReAD principles, in which neural regions activation is maintained through their association with words in STM and decay occurs over time.

5 Empirical Evaluation and Discussion

To evaluate the quality of the topics produced by our algorithm, we conducted two experiments. First, we processed the documents in the Reuter benchmark collection [30] with our implementation of the ReAD principles, with clustering algorithms and using keywords extracted directly from the documents with TFIDF. Only one topic per document was retained (i.e. M=1 in the algorithm) with ReAD. We tested two text clustering algorithms, k-means as a baseline and state-of-the-art spherical k-means [32]. The quality of the results obtained was evaluated with F1, which is a common quality measure in text classification and clustering. F1 results are in the range [0, 1] with 1 being the best quality. Due to space limitation, we refer the reader to [31] for details on the F1 quality metric and its use. Results are shown in table 1, with ReAD having achieved the best F1 quality.

Table 1. Results

Technique	**F1**
ReAD	0.33
Spherical k-means clustering	0.29
K-Means clustering	0.23
TFIDF	0.16

For the second experiment, we turned to human assessments. This answers the important question of whether the topics generated by the program are actually intelligible and evocative of the document content to a human user. According to the independent human assessors, 36% of the topics found by the ReAD algorithm were judged to be acceptable to perfect.

The human assessment scores supplement the F1 quality evaluations, confirming that the algorithmic implementation of the ReAD principles can establish the semantic content of text documents. The level of success is still relatively low but it is comparable to clustering techniques. It is important to note that WordNet is an imperfect replacement for the richness of knowledge found in an actual human brain, as specified in the ReAD principles. Using the limited form of knowledge present in WordNet has the advantage of demonstrating a baseline of what can be achieved, so one can imagine the possibilities with better, richer knowledge such as what can be found on the Web. Nevertheless, it appears quite an accomplishment to obtain a correct identification of topics in 36% of the cases merely with activation and decay of items obtained from a general-purpose lexical source like Wordnet. There is, after all, no special purpose knowledge handcrafting and no traditional semantic or even syntactic analysis. This is also achieved without a large corpus for training, so that each new document can be processed independently and novelty can thus be handled accordingly. Hence, the ReAD algorithm provides a major advantage over existing techniques, namely autonomy and adaptability. Moreover, there are a variety of improvements to be explored, the technique introduced here being in its infancy. Notably, about 30% of words were not found in WordNet and no attempt has been made to exploit words other than nouns. An interesting question that thus needs to be looked into is the effect of using information on all words. For instance, a more complete source of knowledge such as the World Wide Web or Wikipedia could be exploited, as others have done with other techniques [33, 34]. Hence, the web itself could potentially be exploited to make sense of itself. This would be an exciting endeavour in the context of the Semantic Web [35]. We are currently conducting more comprehensive evaluations, examining different variants and parameterization of the algorithm, comparing with other techniques and performing more user assessments.

6 Conclusions

We presented an autonomous and adaptable approach that eliminates the problem of ambiguity and captures the general meaning of a text document. The fundamental ReAD principles behind the approach are: first, the retrieval and activation, for each word read sequentially from the text, of a set of items – simply other words – from a general-purpose, domain independent knowledge source. Second, the decay of infrequent items activation and incremental augmentation for those that occur repeatedly. Third, the convergence over time, as words are read, onto a few discriminately activated items that represent the main concepts discussed in the text, or in other words, its topics.

The principles are different from existing text mining techniques and are inspired by the way humans may be processing information when they read. The innovative

nature of the principles avoids computationally complex NLP and the issue of lack of autonomy due to human intervention. As well, it eliminates the dependence of large text corpora, allowing for the processing of single text in isolation and offering adaptive processing in the face of novelty. The algorithm performs computational determination of the general semantic nature of text – its thematic or conceptual content, or in other words, its topics - with general-purpose lexical knowledge. The approach abandons the standard vector and bag-of-word representation, rather harnessing the order and interdependence of words to compute meaning, and this without conventional syntactic and semantic analysis, without task specific knowledge acquisition, and without training.

References

1. Blei, D., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research (3), 993–1022 (2003)
2. Landauer, T.K., Dumais, S.T.: Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review 104(2), 211–240 (1997)
3. McNamara, D.S.: Computational methods to extract meaning from text and advance theories of human cognition. Topics in Cognitive Science 3(1), 3–17 (2011)
4. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. 34, 1 (2002)
5. Qi, X., Davison, B.D.: Web page classification: Features and algorithms. ACM Comput. Surv. 41, 2 (2009)
6. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. 31, 3 (1999)
7. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, NY (2006)
8. Massey, L.: On the quality of ART1 text clustering. Neural Networks 16, 5–6 (2003)
9. Salton, G., Lesk, M.E.: Computer evaluation of indexing and text processing. J. ACM 15, 1 (1968)
10. Spärck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. of Doc. 28, 1 (1972)
11. Halliday, M.A.K., Hasan, R.: Cohesion in English. Longman Pub. Group, NY (1976)
12. Hotho, A., Staab, S., Stumme, G.: Wordnet improves text document clustering. In: Proceedings of Semantic Web Workshop, the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, NY (2003)
13. Hu, J., Fang, L., Cao, Y., Zeng, H., Li, H., Yang, Q., Chen, Z.: Enhancing text clustering by leveraging Wikipedia semantics. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in information Retrieval, pp. 179–186. ACM, NY (2008)
14. Scott, S., Matwin, S.: Feature engineering for text classification. In: Proceedings of 16th International Conference on Machine Learning, pp. 379–388 (1999)
15. Lenat, D.B.: CYC: A Large-Scale Investment in Knowledge Infrastructure. Commun. ACM 38, 11 (1995)
16. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008), pp. 509–518. ACM, New York (2008)

17. Kim, H.L., Scerri, S., Breslin, J.G., Decker, S., Kim, H.G.: The state of the art in tag ontologies: a semantic model for tagging and folksonomies. In: Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications (DCMI 2008), Dublin Core Metadata Initiative, pp. 128–137 (2008)
18. Turney, P.D.: Learning algorithms for keyphrase extraction. Information Retrieval 2(4), 303–336 (2000)
19. Velardi, P., Navigli, R., D'Amadio, P.: Mining the Web to Create Specialized Glossaries. IEEE Intelligent Systems 23(5), 18–25 (2008)
20. Wong, W., Liu, W., Bennamoun, M.: A probabilistic framework for automatic term recognition. Intelligent Data Analysis 13(4), 499–539 (2009)
21. Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development 1(4), 390 (1957)
22. Cabre-Castellvi, T., Estopa, R., Vivaldi-Palatresi, J.: Automatic term detection: A review of current systems. In: Bourigault, D., Jacquemin, C., L'Homme, M.C. (eds.) Recent Advances in Computational Terminology. John Benjamins, Amsterdam (2001)
23. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In: Proc. of the 2003 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 216–223. Association for Computational Linguistics, Morristown (2003)
24. Milne, D.N., Witten, I.H., Nichols, D.M.: A knowledge-based search engine powered by wikipedia. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, pp. 445–454 (2007)
25. Jarvella, R.J.: Syntactic processing of connected speech. J. Verb. Learn. Verb. Behav. 10 (1971)
26. Just, M.A., Carpenter, P.A.: A capacity theory of comprehension: Individual differences in working memory. Psychol. Rev. 99 (1992)
27. Fellbaum, C.: WordNet: An Electronic Lexical Database (1998)
28. Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. 41, 2 (2009)
29. Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol. Rev. 63 (1956)
30. Lewis, D.D.: Reuters-21578 Distribution 1.0, `http://www.daviddlewis.com/resources/testcollections/reuters21578` (last retrieved April 22, 2010)
31. Massey, L.: Evaluating and Comparing Text Clustering Results. In: Proceedings of 2005 IASTED International Conference on Computational Intelligence (2005)
32. Dhillon, I.S., Modha, D.M.: Concept Decompositions for Large Sparse Text Data using Clustering. Mach. Learn. 42, 1 (2001)
33. Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In: Proceedings of the 21st National Conference on Artificial intelligence, pp. 1301–1306 (2006)
34. Gabrilovich, E., Broder, A., Fontoura, M., Joshi, A., Josifovski, V., Riedel, L., Zhang, T.: Classifying search queries using the Web as a source of knowledge. ACM Trans. Web 3, 2 (2009)
35. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Sci. Am. 284, 5 (2001)

Outlier-Based Approaches for Intrinsic and External Plagiarism Detection

Gabriel Oberreuter, Gaston L'Huillier, Sebastián A. Ríos, and Juan D. Velásquez

Web Intelligence Consortium Chile Research Centre
Department of Industrial Engineering
University of Chile
goberreu@ing.uchile.cl, {glhuilli,srios,jvelasqu}@dii.uchile.cl

Abstract. Plagiarism detection, one of the main problems that educational institutions have been dealing with since the massification of Internet, can be considered as a classification problem using both self-based information and text processing algorithms whose computational complexity is intractable without using space search reduction algorithms. First, self-based information algorithms treat plagiarism detection as an outlier detection problem for which the classifier must decide plagiarism using only the text in a given document. Then, external plagiarism detection uses text matching algorithms where it is fundamental to reduce the matching space with text search space reduction techniques, which can be represented as another outlier detection problem. The main contribution of this work is the inclusion of text outlier detection methodologies to enhance both intrinsic and external plagiarism detection. Results shows that our approach is highly competitive with respect to the leading research teams in plagiarism detection.

Keywords: Text Classification, Outlier Detection, Search Space Reduction, External Plagiarism Detection, Intrinsic Plagiarism Detection.

1 Introduction

Plagiarism in academia is rising and multiple authors have worked to describe this phenomena [6,9]. As commented by Hunt in [6], "Internet Plagiarism" is referred sometimes as a cataclysmic consequence of the "Information Technology revolution", as it proves to be a big problem in academia. In [9], plagiarism is analyzed from various perspectives and considered as a problem that is growing bigger over time. To tackle this problem, the most common approach so far is to detect plagiarism using automated algorithms based on rules and string matching algorithms.

Two main strategies for plagiarism detection have been considered by researchers [10]: Intrinsic and external plagiarism detection. Intrinsic plagiarism detection aims at discovering plagiarism by examining only the input document, deciding whether parts of the input document are not from the same author.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 11–20, 2011.

External plagiarism detection is the approach where suspicious documents are compared against a set of possible references. From exact document copy, to paraphrasing, different levels of plagiarism techniques can been used in several contexts [16].

The main contribution of this work is the usage of outlier detection techniques on text-based data to enhance two plagiarism detection strategies, one for intrinsic plagiarism detection using deviation parameters with respect of the writing style of a given document, and another one to reduce the search space for external plagiarism detection based on the generation of segments of n-gram for approximated plagiarism decision where unrelated documents are discarded efficiently.

This paper is structured as follows: In Section 2 an overview of intrinsic and external plagiarism detection algorithms is presented. Then, in Section 3 the proposed plagiarism detection methods are introduced. Afterwards, in Section 4, the experimental setup and evaluation performance criteria are described. In Section 5 results are discussed. Finally, in Section 6 the main conclusions are presented.

2 Related Work

According to Schleimer et al. [12], copy prevention and detection methods can be combined to reduce plagiarism. While copy detection methods can only minimize it, prevention methods can fully eliminate it and decrease it. Notwithstanding this fact, prevention methods need the whole society to take part, thus its solution is non trivial. Copy or plagiarism detection methods tackle different levels, from simple manual comparison to complex automatic algorithms [11,10].

2.1 Intrinsic Plagiarism Detection

When comparing texts against a reference set of possible sources, comes the complication of choosing the right set of documents to compare. And now more than ever, with the possibilities that Internet bring to plagiarists, this task becomes more complicated to achieve. For this, the writing style can be analyzed within the document and an examination for incongruities can be done. The complexity and style of each text can be analyzed based on certain parameters such as text statistics, syntactic features, part-of-speech features, closed-class word sets, and structural features [16]. Whose main idea is to define a criterium to determine if the style has changed enough to indicate plagiarism.

Stamatatos [14] presented a new method for intrinsic plagiarism detection. As described by it's author, this approach attempts to quantify the style variation within a document using character n-gram profiles and a style change function based on an appropriate dissimilarity measure originally proposed for author identification. Style profiles are first constructed using a sliding window. For the construction of those profiles the author proposed the use of character n-grams. These n-grams are used for getting information on the writer's style.

The method then analyzes changes on the profiles to determine if a change is significative enough to indicate another's author style.

Other approaches have been proposed, such as presented by Seaward & Matwin [13] introduce Kolmogorov Complexity measures as a way of extracting structural information from texts for Intrinsic Plagiarism Detection. They experiment with complexity features based on the Lempel-Ziv compression algorithm for detecting style shifts within a single document, thus revealing possible plagiarized passages.

2.2 External Plagiarism Detection

In terms of external plagiarism detection algorithms, the use of n-grams have shown to give some flexibility to the detection task, as reworded text fragments could still be detected [8]. Other approaches focus on solving the plagiarism detection problem as a traditional classification problem from the machine learning community [1,4]. Bao et al. in [1], proposed to use a Semantic Sequence Kernel (SSK), and then using it into a traditional Support Vector Machines (SVMs) formulation based on the Structural Risk Minimization (SRM) principle from statistical learning theory [15], where the general objective is finding out the optimal classification hyperplane for the binary classification problem (plagiarized, not plagiarized).

In [7] the authors introduced their model for automatic external plagiarism detection. It consist of two main phases; the first is to build the index of the documents, while in the second the similarities are computed. This approach uses word n-grams, with n ranging from 4 to 6, and takes into account the number of matches of those n-grams between the suspicious documents and the source documents for computing the detections. The algorithm have the authors won the first place at the PAN@2010 competition [10].

2.3 Outlier Identification Approaches and Plagiarism Detection

As described in [5], an outlier is an observation which deviates from other observations as to become suspicious that was generated by a different statistical process. In general terms, outlier detection can be classified in proximity approaches, such as distance-based or density-based, and model-based approaches, such as statistical tests, depth-based, and deviation-based methodologies [5]. In plagiarism detection, deviation-based methodologies have been previously used for intrinsic plagiarism detection [14], and distance-based for probability distribution approaches for external plagiarism [2].

3 Proposed Methods

In this section, the main contribution of our work is described. In the first place, a search space reduction algorithm using outlier detection techniques for external plagiarism detection is presented. Then, an intrinsic plagiarism detection algorithm is proposed based on variance of n-gram content on sliding windows over the whole document.

Let us introduce some concepts. In the following, let $\mathcal{V}$ a vector of words that defines the vocabulary to be used. We will refer to a word w, as a basic unit of discrete data, indexed by $\{1, ..., |\mathcal{V}|\}$. A document d is a sequence of S words ($|d| = S$) defined by $\mathbf{w} = (w^1, ..., w^S)$, where w^s represents the s^{th} word in the message. Finally, a corpus is defined by a collection of $\mathcal{D}$ documents denoted by $\mathcal{C} = (\mathbf{w}_1, ..., \mathbf{w}_{|\mathcal{D}|})$.

3.1 Distance-Based Outlier Detection for External Plagiarism Detection

For a corpus $\mathcal{C} = \{\mathcal{D}_{\text{source}}, \mathcal{D}_{\text{suspicious}}\}$, the idea is to find all plagiarized documents in the suspicious partition, using as search space the source partition. In general terms, the algorithm first reduces the search space by using an approximated search of segments of n-grams, and then within selected pairs of documents, using an exhaustive search algorithm, finds the offset and its length.

First, for each document d_i, after stop-words removal, a set t_i of n-grams with the structure $(w_i, w_{i+1} \ldots, w_{i+n}), \forall i \geq 1, n \leq S$ must be created. Then, to compute the difference between these document vector's, groups κ_i of k n-grams are created. The basic idea is to test the closeness between a pair of documents using a distance-based outlier detection approach, which firstly indicates whether chunks of words between n-gram representations of the document have at least θ_κ exact matches, and then, checks if all n-grams of both documents have at least θ_t matches, as shown in Algorithm 1.

Algorithm 1. Approximate comparison between two documents

Require: $\kappa_i, \kappa_j, \mathbf{t}_i, \mathbf{t}_j, \theta_\kappa, \theta_t$
1: **if** sMatch$(\kappa_i, \kappa_j, s \geq \theta_\kappa)$ **then**
2: **if** sMatch$(\mathbf{t}_i, \mathbf{t}_j, s \geq \theta_t)$ **then**
3: return true
4: **end if**
5: **else**
6: return false
7: **end if**

As presented in Algorithm 1, once documents d_i and d_j are processed in n-grams and segments of k n-grams, t_i, t_j and κ_i, κ_j respectively, a set of conditions are evaluated in order to set the relation that document d_i has with document d_j, that is, if they are somehow related (algorithm 1 returns true), or if it is not worthy to keep finding further relationships (Algorithm 1 returns false). In this sense, this is an approximated finding procedure that considers both n-grams and their k segments to decide if there is enough information to classify as plagiarism or not, and using the distance function sMatch, which checks for thresholds θ_κ and θ_t.

Condition sMatch($\kappa_i, \kappa_j, s \geq \theta_\kappa$) states that at least θ_κ n-grams must match in between segments κ_i and κ_j. If this is hold, the next condition sMatch($\mathbf{t}_i$,$\mathbf{t}_j$, $s \geq \theta_2$) is associated to find whether at least θ_t n-grams matches between $\mathbf{t}_i$ and $\mathbf{t}_j$.

After reducing search space, it is possible now to go into a further algorithm for finding the needed offset and its length. More details on offset and length finding algorithm, please refer to Oberreuter et al. in [8], which is intentionally omitted by authors due to lack of space.

3.2 Intrinsic Plagiarism Detection

Using uni-grams and without removing stop-words, a frequency-based algorithm to test self-similarity of document is proposed. First, a frequency vector $\mathbf{v}$ is built for all words on a given document. Then, the complete document is clusterized creating groups $\mathcal{C}$. As a first approach, these groups or segments $c \in \mathcal{C}$ are created using a sliding window of length m over the complete document. Afterwards, for each segment $c \in \mathcal{C}$, a new frequency vector v_c is computed, which is used in further steps to compare whether a segment is deviated with respect to the footprint of the complete document. This is performed by using the Algorithm 2.

Algorithm 2. Intrinsic plagiarism evaluation

Require: $\mathcal{C}, \mathbf{v}, m, \delta$
for $c \in \mathcal{C}$ **do**
  $d_c \leftarrow 0$
  build v_c using term frequencies on segment c
  for word $w \in v_c$ **do**
    $d_c \leftarrow d_c + \frac{|freq(w,\mathbf{v}) - freq(w,v_c)|}{|freq(w,\mathbf{v}) + freq(w,v_c)|}$
  end for
end for
style $\leftarrow \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} d_c$
for $c \in \mathcal{C}$ **do**
  if $d_c < style - \delta$ **then**
    Mark segment c as outlier and potential plagiarized passage.
  end if
end for

As presented in Algorithm 2, the general footprint or style of the document is represented by the average of all differences computed for each segment and the complete document. Finally, all segments are classified according to its distance with respect to the document's style. As an example, in Figure 1, a graphical representation of this evaluation is presented.

In this case, the average value (represented by yellow line), is compared against the style function, which is roughly computed by the difference on the frequency of words between vectors $\mathbf{v}$ and $v_c, \forall c \in \mathcal{C}$. In any case that the style function is lower than the average value minus δ, the segment is classified as suspicious.

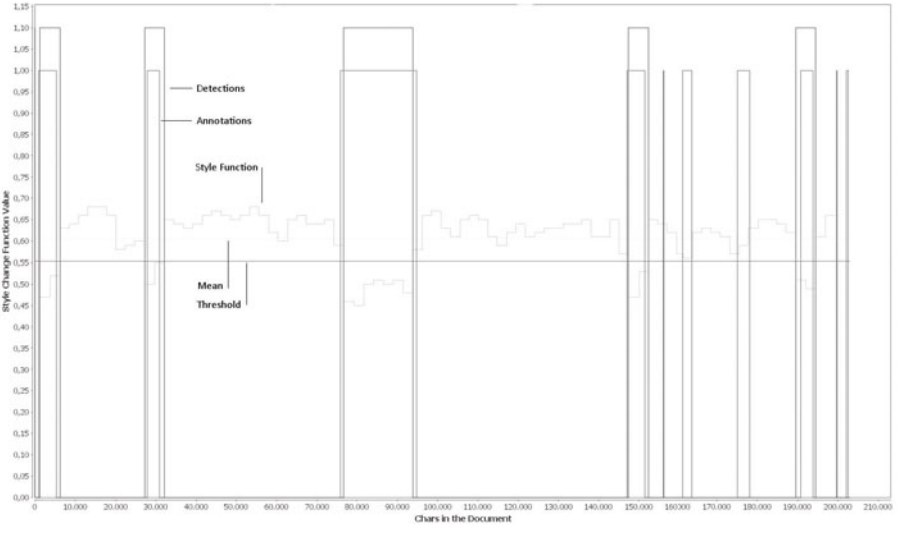


Fig. 1. Intrinsic plagiarism detection example

In this example, real plagiarized annotations are presented by red areas, and all classified passages are presented by blue areas, indicating that for all 10 cases of real plagiarized annotations, the proposed method achieved to classify 5 of them, without mistakes.

4 Experiments

In this section, the experimental setup and the evaluation criteria is presented. For external plagiarism detection evaluation, the PAN@2010 plagiarism detection corpus [10] was used, which considers a set of 11,148 source documents and another set of 15,925 suspicious documents, with 68,558 plagiarism cases. For the evaluation of intrinsic plagiarism detection, the PAN@2009 plagiarism detection corpus [11] was used, which considers a set of 6,183 suspicious documents.

4.1 Evaluation Criteria

As described in [10,11], let S be the set of all plagiarized passages, R the set of all detections made by a given plagiarism detection algorithm, and S_R a subset of S whose detections are presented in R. Let the function $|\cdot| : \mathcal{P}(\mathcal{V}) \rightarrow \mathbb{R}^1$ which states the number of chars in a given string generated from vocabulary $\mathcal{V}$. The evaluation metrics proposed for plagiarism detection are described as follows:

$$\text{Recall} = \frac{1}{|S|} \sum_{i=1}^{|S|} \left(\frac{\#\text{ detected chars of } s_i}{|s_i|} \right) \tag{1}$$

$$\text{Precision} = \frac{1}{|R|}\sum_{i=1}^{|R|}\left(\frac{\#\text{ plagiarized chars of } r_i}{|r_i|}\right) \tag{2}$$

$$\text{Granularity} = \frac{1}{|S_R|}\sum_{i=1}^{|S_R|}(\#\text{ of detections of } s_i \in R) \tag{3}$$

$$\text{Overall score} = \frac{\text{F-measure}}{log_2(1+\text{granularity})} \tag{4}$$

where F-measure is the harmonic mean of precision and recall. The Granularity was introduced in order to quantify the number of detections of a given plagiarized passage. If the model detects the passage more than once, the score gets penalized.

4.2 Intrinsic Plagiarism Detection Experimental Setup

For intrinsic plagiarism detection, evaluation was conducted by using the PAN@2009 intrinsic competition, for a detailed description of benchmark algorithms, please refer to Potthast et al. [11].

4.3 External Plagiarism Detection Experimental Setup

In this case, the evaluation setup is characterized by all performance metrics described in Section 4.1, and considering as benchmark all contestants of the PAN@2010 plagiarism detection competition[1]. For more information on external plagiarism detection benchmark algorithms, please refer to Potthast et al. [10].

5 Results and Discussions

5.1 External Plagiarism Detection

As shown in Table 1, the best results for the retrieval task were achieved by Kasprzak and Brandejs approach [7]. The overall score was 0.80, and their method achieved good results at the three metrics: precision, recall and granularity. The next top results show similar characteristics, being well balanced in the three metrics. Our proposed model took fifth place, with an overall score of 0.61, precision of 0.85 and recall of 0.48. The granularity of the top performers were all close to 1.

5.2 Intrinsic Plagiarism Detection

The results for the intrinsic task are shown in Table 2. The results are based on the quality of the detection, which only considers the information on each document itself. The second best can be considered as the baseline, as it classified almost every segment as plagiarized [11]. This lead to a reduced precision, thus

[1] http://pan.webis.de/ [last accessed 01-03-2010].

Table 1. Results for ranking, overall score, F-measure, precision, recall, granularity, and name of lead developer [10]

Rank	Overall Score	F-Measure	Precision	Recall	Granularity	Lead developer
1	0.80	0.80	0.94	0.69	1.00	Kasprzak et al.
2	0.71	0.74	0.91	0.63	1.07	Zou et al.
3	0.69	0.77	0.84	0.71	1.15	Muhr et al.
4	0.62	0.63	0.91	0.48	1.02	Grozea et al.
5	0.61	0.61	0.85	0.48	1.01	**Proposed Model**
6	0.59	0.59	0.85	0.45	1.00	Torrejón et al.
7	0.52	0.53	0.73	0.41	1.00	Pereira et al.
8	0.51	0.52	0.78	0.39	1.02	Palkovskii et al.
9	0.44	0.45	0.96	0.29	1.01	Sobha et al.
10	0.26	0.39	0.51	0.32	1.87	Gottron et al.
11	0.22	0.38	0.93	0.24	2.23	Micol et al.
12	0.21	0.23	0.18	0.30	1.07	Costa-jussá et al.
13	0.21	0.24	0.40	0.17	1.21	Nawab et al.
14	0.20	0.22	0.50	0.14	1.15	Gupta et al.
15	0.14	0.40	0.91	0.26	6.78	Vania et al.
16	0.06	0.09	0.13	0.07	2.24	Suárez et al.
17	0.02	0.09	0.35	0.05	17.31	Alzahrani et al.
18	0.00	0.00	0.60	0.00	8.68	Iftene et al.

Table 2. Results for rank, overall score, F-measure, precision, recall, and granularity for each algorithm presented in section 4

Rank	Overall Score	F-Measure	Precision	Recall	Granularity	Lead Developer
1	0.2462	0.3086	0.2321	0.4607	1.3839	Stamatatos (2009)
2	0.1955	0.1956	0.1091	**0.9437**	1.0007	Hagbi and Koppel (2009)
3	0.1766	0.2286	0.1968	0.2724	1.4524	Muhr et al. (2009)
4	0.1219	0.1750	0.1036	0.5630	1.7049	Seaward and Matwin (2009)
	0.3457	**0.3458**	**0.3897**	0.3109	**1.0006**	**Proposed Model**

obtaining an overall score of 0.1955. The winner was Stamatatos approach ([14]), with a recall of 0.4607, precision of 0.2321 and granularity of 1.3839. His method achieved a good combination of precision and recall, and a not top performer granularity. Our proposed method gets an overall score of 0.3457, greater that any other approach, with a positive difference of 0.0995 with the winner's approach. Our model gets the best result at F-measure, precision and granularity.

6 Conclusions and Future Work

In this work we have introduced two new models for outliers classification applied to plagiarism in digital documents. In the intrinsic plagiarism detection task, our model uses information only from the given document, and select those segments from the text that deviate significantly from the general style. The algorithm

achieves remarkable results, being the best at precision and overall score, using as a benchmark other approaches from PAN@2009 intrinsic plagiarism competition. Also, the algorithm does not utilize language-dependent features such as stopwords, and is simple and straightforward. The model's effectiveness is to be studied in languages other than English, for which as future work, it would be interesting to study other applications, e.g. extracting different topics in a given document, or for author identification.

For the external plagiarism detection task, the proposed model remains competitive against other approaches, obtaining the fifth place in the PAN@2010 external plagiarism detection competition. The task of classifying whether a document is plagiarized or not, and comparing it against a set of possible sources, remains to be a compelling task, as it would be interesting to include Web search for the retrieval of additional source candidates.

Acknowledgment. Authors would like to thank continuous support of "Instituto Sistemas Complejos de Ingeniería" (ICM: P-05-004- F, CONICYT: FBO16; www.isci.cl); and FONDEF project (DO8I-1015) entitled, DOCODE: Document Copy Detection (www.docode.cl).

References

1. Bao, J.-P., Shen, J.-Y., Liu, X.-D., Liu, H.-Y., Zhang, X.-D.: Semantic sequence kin: A method of document copy detection. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 529–538. Springer, Heidelberg (2004)
2. Barrón-Cedeño, A., Rosso, P., Benedí, J.-M.: Reducing the plagiarism detection search space on the basis of the kullback-leibler distance. In: Gelbukh, A. (ed.) CICLing 2009. LNCS, vol. 5449, pp. 523–534. Springer, Heidelberg (2009)
3. Braschler, M., Harman, D., Pianta, E. (eds.): CLEF 2010 LABs and Workshops, Notebook Papers, Padua, Italy (September 22-23, 2010)
4. Chow, T.W.S., Rahman, M.K.M.: Multilayer som with tree-structured data for efficient document retrieval and plagiarism detection. Trans. Neur. Netw. 20(9), 1385–1402 (2009)
5. Hawkins, D.: Identification of Outliers. Chapman and Hall, London (1980)
6. Hunt, R.: Let's hear it for internet plagiarism. Teaching Learning Bridges 2(3), 2–5 (2003)
7. Kasprzak, J., Brandejs, M.: Improving the reliability of the plagiarism detection system - lab report for pan at clef 2010. In: Braschler, et al. (eds.) [3] (2010)
8. Oberreuter, G., L'Huillier, G., Ríos, S.A., Velásquez, J.D.: Fastdocode: Finding approximated segments of n-grams for document copy detection - lab report for pan at clef 2010. In: Braschler, et al. (eds.) [3] (2010)
9. Park, C.: In other (people's) words: plagiarism by university students – literature and lessons. Assessment and Evaluation in Higher Education (5), 471–488 (2003)
10. Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd international competition on plagiarism detection. In: Braschler, M., Harman, D. (eds.) Notebook Papers of CLEF 2010 LABs and Workshops, Padua, Italy (September 22-23, 2010)

11. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st international competition on plagiarism detection. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009), pp. 1–9. CEUR-WS.org (September 2009)
12. Schleimer, S., Wilkerson, D.S., Aiken, A.: Winnowing: local algorithms for document fingerprinting. In: SIGMOD 2003: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pp. 76–85. ACM, New York (2003)
13. Seaward, L., Matwin, S.: Intrinsic plagiarism detection using complexity analysis. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009), pp. 56–61. CEUR-WS.org (September 2009)
14. Stamatatos, E.: Intrinsic plagiarism detection using character n-gram profiles. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009), pp. 38–46. CEUR-WS.org (September 2009)
15. Vapnik, V.N.: The Nature of Statistical Learning Theory (Information Science and Statistics). Springer, Heidelberg (1999)
16. Eissen, S.M.z., Stein, B., Kulig, M.: Plagiarism detection without reference collections. In: Decker, R., Lenz, H.-J. (eds.) GfKl. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 359–366. Springer, Heidelberg (2006)

An Extended Method for Finding Related Web Pages with Focused Crawling Techniques

Kazutaka Furuse, Hiroaki Ohmura, Hanxiong Chen, and Hiroyuki Kitagawa

Department of Computer Science,
Graduate School of Systems and Information Engineering,
University of Tsukuba

Abstract. This paper proposes an extended mechanism for efficiently finding related web pages, which is constructed by introducing some focused crawling techniques.

One of the successful methods for finding related web pages is Kleinberg's HITS algorithm, and this method determines web pages which are related to a set of given web pages by calculating the hub and authority scores. Although this method is effective for extracting fine related web pages, it has a limitation that it only concerns the web pages which are directly connected to the given web pages for the score calculation.

The proposed method of this paper extends the HITS algorithm by enlarging neighborhood graph used for the score calculation. By navigating links forward and backward, pages which are not directly connected to the given web pages are included in the neighborhood graph. Since the navigation is done by using the focused crawling techniques, the proposed method effectively collects promising pages which contribute to improve accuracy of the scores. Moreover, unrelated pages are filtered out for avoiding topic drift in the course of the navigation. Consequently, the proposed method successfully finds related pages, since scores are calculated with adequately extended neighborhood graphs. The effectiveness and the efficiency of the proposed method is confirmed by the results of experiments performed with real data sets.

1 Introduction

In recent years, the World Wide Web is continuously and rapidly growing, and the number of web pages is increasing at a staggering rate of speed. Consequently, users are frequently in face of difficulty to yield useful information from billions of pages in the Web.

Many types of research work have been done to resolve this kind of problems. One of such research work is the construction of mechanisms for finding related web pages. The mechanisms of this type extract some web pages from the web, which are related to a set of given web pages. For example, when a user gives a set of known web pages of a football team he/she likes, the mechanisms find some web pages related to the team in a short time.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 21–30, 2011.

The HITS (Hyperlink-Induced Topic Search) algorithm proposed by Kleinberg[5] is one of successful methods which can be used for finding related web pages by utilizing link analysis techniques. In this method, web pages related to a set of given pages are determined by calculating the hub and authority scores. In HITS, pages with high authority scores are considered that they are authoritative pages (pages including useful information), and pages with high hub scores are considered that they are good hub pages (pages including useful links to authoritative pages). The authority score and hub score are mutually calculated, since they affect each other.

Although the HITS method is effective method, it has a mechanistic limitation when it is used for extracting fine related web pages. That is, it only calculates the scores of web pages which are directly connected to the given web pages as candidates. Therefore, it can only extract related pages which have links from/to the pages given by the users.

In this paper, we propose an effective method for finding web pages, which are related to, but not-necessarily directly connected to, the given web pages. The proposed method of this paper extends the HITS algorithm by navigating links forward and backward for including candidate pages which are not directly connected to the given web pages. Since the navigation is done by using the focused crawling techniques for avoiding to navigate hopeless links, the proposed method effectively finds promising candidate pages. In this paper, we also show the effectiveness and the efficiency of the proposed method with the results of experiments performed with real data sets.

The remainder of this paper is organized as follows. In Section 2, we reviews the HITS algorithm and the focused crawling techniques. Section 3 explains the approach and the details of the proposed method, and Section 4 gives the results of the experiments. In Section 5, we conclude this paper.

2 Related Work

2.1 HITS

Kleinberg's HITS[5] is a method for scoring web pages by using the notions of hub and authority. This method was firstly implemented in the CLEVER search engine[3, 5], and the calculated scores are used in the system for ranking search results.

In this method, it is considered that web pages can be categorized to two different types: authoritative pages and hub pages. Authoritative pages are considered that they include useful information, and hub pages are considered that they include useful links to authoritative pages.

The authority scores and the hub scores are calculated as follows. Given a web graph (V, E), where V is a set of vertices (web pages) and E is a set of edges (links from a web page to another web page). Let $R \subseteq V$ be a set of web pages given by a user. This set is called the *root set*. First, HITS computes a *base set* $B \subseteq V$ which is defined as:

$$B = R \cup \bigcup_{u \in R} \{v \mid v \in V \wedge (u, v) \in E\} \cup \bigcup_{v \in R} S_n(\{u \mid u \in V \wedge (u, v) \in E\}),$$

where $S_n(X)$ denotes a set of n elements randomly sampled from the set X (n is typically set to be below 100). Then, the neighborhood graph (B, N), where $N = \{(u, v) \mid u \in B \wedge v \in B \wedge (u, v) \in E\}$, is constructed.

For each web page p in the neighborhood graph, the authority score $A(p)$ and the hub score $H(p)$ are mutually calculated as follows.

$$A(p) = \sum_{q \in \{q \mid (q,p) \in N\}} H(q)$$

$$H(p) = \sum_{r \in \{r \mid (p,r) \in N\}} A(r)$$

With this algorithm, the set of pages which have high authority scores becomes the result of related pages extraction (together with the pages with high hub scores, if needed).

Since this method calculates scores only for the pages in the neighborhood graph, it cannot extract any pages which are not directly linked with the initially given pages (pages in the root set R). The proposed method overcomes this disadvantage by enlarging the neighborhood graph using focused crawling tecuniques.

2.2 Focused Crawling

Focused crawling is a technique for efficiently collecting web pages, which are related to a given topic, by selectively navigating links [2] [8]. Unlike generic crawlers (also known as *bots* or *spiders*) used for search engines, which navigate links in depth-first or breadth-first manner, focused crawlers preferentially select promising links to navigate. For determining which links are promising, they usually use supervised learning techniques such as *Bayesian filters* and *support vector machines* [6].

Typical algorithm of the focused crawling is shown in Figure 1. In general, focused crawlers use a *priority queue* F, in which URLs are stored in the descending order of the priority scores. The priority scores are usually calculated by some supervised learning mechanism. F is also called *frontier*.

In the procedure of the crawling, the URL which has the highest score is dequeued from the priority queue. After fetching the page of the URL, outgoing links in the page are extracted, and the URLs of the links are enqueued to the priority queue if they are not visited yet. If some criteria holds (e.g., when N pages are fetched since the last reordering process), then URLs in the queue are reordered in the descending order of the priority score.

In this manner, focused crawlers preferentially select promising links. Consequently, it enables us to collect web pages related to specific topics with much smaller number of link navigations than generic crawlers which navigate in the depth-first or breadth-first manner.

```
Inputs:
s: the URL of the seed page
n: the number of pages to be crawled

Outputs:
P: the set of crawled pages

Algorithm:
  F ← {s}
  P ← ∅
  while |P| < n ∧ F ≠ ∅ do
    u ← dequeue a URL from F
    p ← fetch the page u
    P ← P ∪ {p}
    L ← extract out-going links in p
    for all q such that q ∈ L ∧ q ∉ P do
      enqueue q to F
    end for
    if reordering criteria holds then
      reorder URLs in F according to the priority score
    end if
  end while
```

Fig. 1. Typical Algorithm of Focused Crawling

3 Finding Related Pages with Focused Crawling Techniques

3.1 Approach

As described in the previous section, the HITS algorithm only calculates scores of pages directly linked with the initially given pages (the root set). In other words, it is impossible to extract any related web pages which are distant from the given pages. To resolve this problem, the proposed method of this paper extends this method by increasing the number of pages for which scores are calculated. In the proposed method, this is achieved by enlarging the neighborhood graph using the focused crawling techniques.

Difference between original HITS and the proposed method is illustrated in Figure 2. In original HITS, scores are calculated with a base set, which are constructed by adding directly connected pages to the given root set. In the proposed method, on the other hand, the neighborhood graph is constructed by including related pages found by the focused crawling procedure and excluding unrelated pages filtered by Bayesian filters. The scores are calculated with this enlarged neighborhood graph.

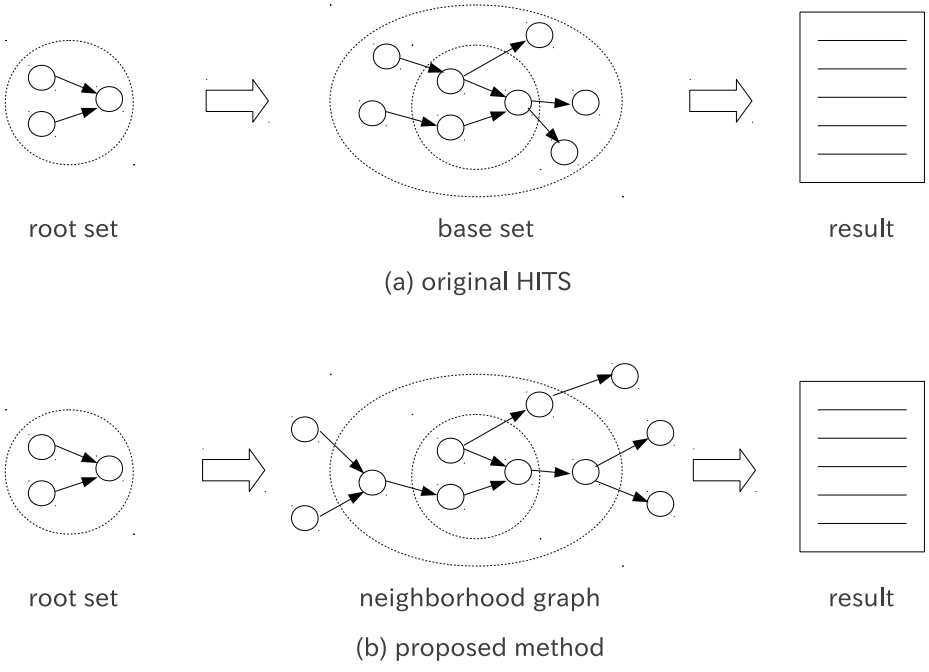


Fig. 2. Comparison of Original HITS and the Proposed Method

Figure 3 illustrates how neighborhood graphs are constructed in the proposed method. Firstly, the initial neighborhood graph (surrounded by dashed line) is constructed in the same way as the HITS algorithm, and authority and hub scores are calculated (a). Then, by using the focused crawling technique, we navigate and fetch the page which has the highest authority score (b). Next, the page of the highest hub score is fetched (c). In this way, we fetch pages of high authority scores and pages of high hub scores alternately. After fetching some pages, we enlarge the neighborhood graph, and recalculate the authority and hub scores (d). We continue this procedure until the size of the neighborhood graph becomes a predefined size.

Since the proposed method uses the authority and hub scores for the priority to select the links to be navigated, it is likely that the promising pages (pages which might have high authority/hub scores) are fetched preferentially. Thanks to this, it is possible to efficiently collect related pages which are distant from the pages in the given root set.

One problem arising when we construct the neighborhood graphs is *topic drift*, that is, pages obtained by following forward/backward links might not contain any text of related topics. This problem is getting more serious in the proposed method, since it follows much links than the original HITS. To overcome this problem, we use the Bayesian filters, which are built from web pages in the given root sets, to exclude unrelated pages in the course of the focused crawling.

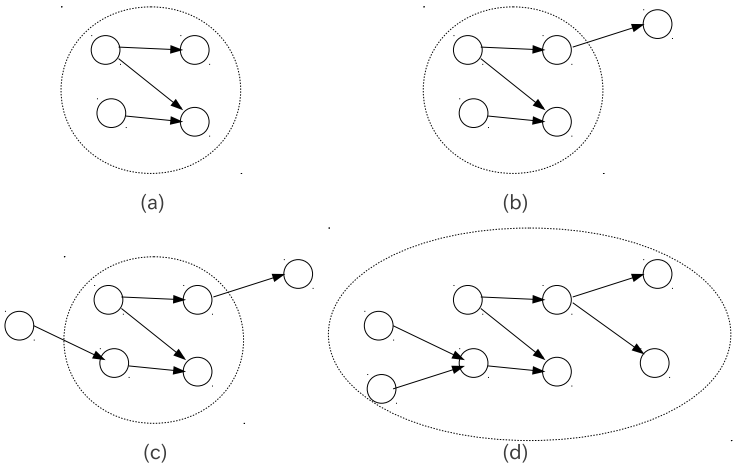


Fig. 3. Construction of a Neighborhood Graph

3.2 Details of the Algorithm

In this subsection, we describe the details of the algorithm of the proposed method. The method takes the root set (set of initially given web pages) as input, and extracts some pages related to the root set.

First of all, the base set B and links of the neighborhood graph N are constructed in the same manner as the HITS algorithm. For using the focused crawling techniques, we use two priority queues, F_a and F_h. Pages in F_a and F_h are sorted in the descending orders of the authority scores and the hub scores, respectively. These queues are initialized by the neighborhood graph (B, N).

After the initialization procedure, the proposed method enlarges the neighborhood graph by performing focused crawling process according to the order of pages in the priority queues F_a and F_h. First, a page of the highest authority score is dequeued from F_a, and fetched. Then, *in-coming* links to the page are extracted (for this, we need a mechanism for obtaining backward links from the web, which are provided as a set of APIs by some search engines). Since the dequeued page has a high authority score, it is very possible that the pages linking to the dequeued page have high hub scores. Therefore, they are enqueued to the priority queue F_h, in which pages are sorted in the descending order of the hub scores, unless they are filtered out by the Bayesian filters (as described later).

In the similar way, the page of the highest hub score in F_h is dequeued and fetched, and *out-going* links from the page are extracted. The extracted links are enqueued to F_a, since the pages are linked from the page which has high hub score. The processes of dequeuing from the F_a and F_h are performed alternately.

After fetching a certain number of pages (defined as p in Figure 4), the reorganization of the neighborhood graph and the recalculations of authority and hub scores are performed.

The algorithm is terminated when the neighborhood graph is enlarged to the predefined size, and outputs pages which have top n authority scores as the results.

As mentioned earlier, one serious problem that may occur in this method is that it may include some unrelated web pages in the neighborhood graphs. For example, this can happen when we reach at a page which collects wide range of links (e.g, the pages in the Open Directory Project[1]). To overcome this

Inputs:
R: the set of given pages (root set)
n: the number of pages to be extracted
f: the final size of the neighborhood graph
p: the recalculating period

Outputs:
P: the set of extracted pages (related pages)

Algorithm:

$B \leftarrow$ construct the base set from R
$N \leftarrow$ collect links in the base set B
$F_a \leftarrow \{v \mid (u, v) \in N\}$ in the descending order of authority scores
$F_h \leftarrow \{u \mid (u, v) \in N\}$ in the descending order of hub scores
while $|B| < f \wedge F_a \neq \emptyset \wedge F_h \neq \emptyset$ **do**
  { fetch an authoritative page }
  $u_a \leftarrow$ dequeue a URL from F_a
  $p_a \leftarrow$ fetch the page u_a
  $B \leftarrow B \cup \{p_a\}$
  $L_a \leftarrow$ extract in-coming links in p_a
  for all q_a such that $q_a \in L_a \wedge q_a \notin B$ **do**
    enqueue q_a to F_h unless it is filtered out by the Bayesian filter
  end for
  { fetch a hub page }
  $u_h \leftarrow$ dequeue a URL from F_h
  $p_h \leftarrow$ fetch the page u_h
  $B \leftarrow B \cup \{p_h\}$
  $L_h \leftarrow$ extract out-going links in p_h
  for all q_h such that $q_h \in L_h \wedge q_h \notin B$ **do**
    enqueue q_h to F_a unless it is filtered out by the Bayesian filter
  end for
  { reorganize the neighborhood graph }
  if p pages are fetched after the last reorganization **then**
    recalculate authority and hub scores in B
    reorder pages in F_a and F_h by the recalculated scores
  end if
end while
$P \leftarrow$ pages in B which have top n authority scores

Fig. 4. Algorithm of the Proposed Method

problem, we use the Bayesian filters to exclude unrelated pages in the process of enlarging the neighborhood graphs. The filters we use in the proposed method is straightforwardly constructed from texts in web pages of the root sets (as positive examples) and texts in the randomly sampled web pages (as negative examples).

4 Experiments

For confirming the effectiveness of the proposed method, we performed preliminary experiments with real data sets and compared it with the original HITS algorithm. The experiments have been implemented in the Ruby programming language, and performed on the Intel AMD Phenom II box running under Ubuntu Linux 10.04. For evaluating the appropriateness of the results, we invited ten evaluators to participate in our survey to judge whether the result pages output from the original HITS and the proposed method are related to the predefined topics or not. The parameters used in the experiments are listed in Table 1.

We manually constructed three sets of web pages of topics: "apples (of fruits)", "ruby", and "native american". We use these sets for the input (root sets) of the original HITS algorithm and the proposed method. The invited evaluators rated top 10 pages output from the methods, according to the relatedness to the predefined topics.

We use two performance measures for evaluating the effectiveness and quality of the methods: average and $AP@k$.

Let s_{ij} be the score determined by the evaluator i for the topic j, and let n be the number of pages output by the evaluated methods. The average score of the output pages for the topic j is calculated as follows.

$$average(j) = \frac{1}{n}\sum_{i=1}^{n} s_{ij}$$

$AP@k$, the average precision at result cut-off value k (which is originally presented in [4]), is defined as follows,

$$P@k = \frac{1}{k}\sum_{i=1}^{k} rel(i)$$

$$AP@k = \frac{\sum_{i=1}^{k} rel(i)P@i}{\sum_{i=1}^{n} rel(i)}$$

Table 1. Values of the Parameters used in the Experiments

parameter	value	description
n	10	the number of pages to be extracted
f	4000	the final size of the neighborhood graph
p	128	the recalculating period

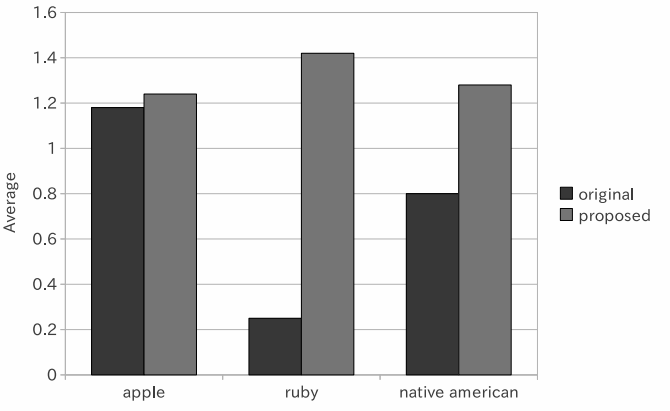


Fig. 5. Average of Ratings

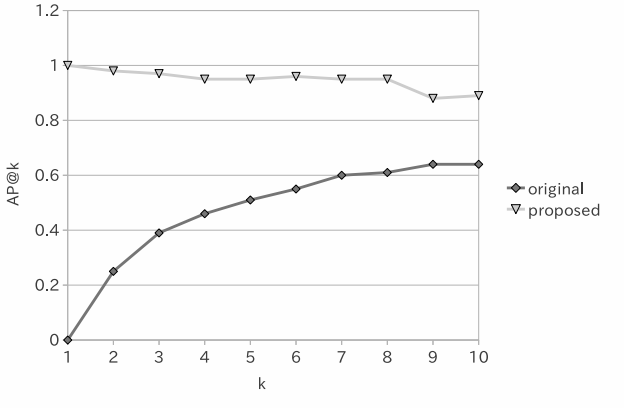


Fig. 6. $AP@k$ of Ratings

where $rel(i) = 1$ when the result page i is rated as a related page by the evaluators, otherwise $rel(i) = 0$. Therefore, $AP@k$ is the average of the fraction of relevant results among the k highest-ranked results.

The results of the experiments are shown in Figure 5 and Figure 6. In both of these results, we confirm that the proposed method outperforms the original HITS for finding related web pages from the set of given pages of topics.

5 Conclusions

This paper describes the extended method for effectively finding web pages related to a given set of web pages. The proposed method enlarges the neighborhood graphs by using focused crawling techniques, and enables to find related

pages which are distant from the initially given pages. This overcomes the disadvantage of the original HITS algorithm.

As future work, we plan to carry out more large scale experiments and evaluate the results. We also plan to incorporate the adaptive focused crawling techniques[7] to improve the proposed method.

References

1. The Open Directory Project, http://www.dmoz.org/
2. Chakrabarti, S., van den Berg, M., Dom, B.: Focused crawling: a new approach to topic-specific web resource discovery. In: Proceedings of the Eighth International Conference on World Wide Web, WWW 1999, pp. 1623–1640. Elsevier North-Holland, Inc., New York (1999)
3. Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Kleinberg, D.G.J.: Automatic resource compilation by analyzing hyperlink structure and associated text. In: Proceedings of the Seventh International Conference on World Wide Web 7, WWW 7, pp. 65–74. Elsevier Science Publishers B. V., Amsterdam (1998)
4. Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2000, pp. 41–48 (2000)
5. Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46, 604–632 (1999)
6. Liu, B.: Web Data Mining — Exploring Hyperlinks, Contents, and Usage Data. Springer, Heidelberg (2007)
7. Micarelli, A., Gasparetti, F.: Adaptive focused crawling. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 231–262. Springer, Heidelberg (2007)
8. Olston, C., Najork, M.: Web crawling. Foundations and Trends in Information Retrieval 4, 175–246 (2010)

Development of Multilingual Interview-Sheet Composition System to Support Multilingual Communication in Medical Field

Taku Fukushima[1], Takashi Yoshino[2], and Aguri Shigeno[3]

[1] Graduate School of Systems Engineering, Wakayama University,
930 Sakaedani, Wakayama, Japan
fukushima@yoslab.net
[2] Faculty of Systems Engineering, Wakayama University,
930 Sakaedani, Wakayama, Japan
yoshino@sys.wakayama-u.ac.jp
http://www.wakayama-u.ac.jp/~yoshino/
[3] Center for Multicultural Society Kyoto,
143 Manjuji-cho, Shimogyo-ku, Kyoto, Japan
aguri@tabunka-kyoto.org

Abstract. Recently, the number of foreign residents and foreign visitors in Japan has been increasing every year. Consequently, the opportunities for communication amongst people whose native languages differ are increasing. In healthcare facilities, a paper-based multilingual interview sheet is used to facilitate communication between medical workers and foreign patients. However, this interview sheet has been found to be inadequate for such purposes. Moreover, Japanese medical workers find it difficult to understand the different languages written on a paper-based interview sheet. To resolve this problem, we have developed a multilingual interview-sheet composition system that uses parallel texts and machine translation. This system can convey essential patient information to a medical worker during consultation. The contributions of this study are as follows: (1) We have proposed a multilingual interview-sheet composition system that can be used for communication between medical workers and foreign patients. We have developed this system using both parallel texts and machine translation. (2) We showed that a patient is able to create a multilingual interview sheet using a parallel corpus and machine translation. (3) Because it may be difficult to indicate an affected area of the body using words alone, we suggest that affected areas be indicated by the user using a human-body image for more accurate communication.

Keywords: interview sheet, parallel text, machine translation.

1 Introduction

Recently, worldwide globalization has helped to increase communication among people with different native languages. However, it is difficult to learn many

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 31–40, 2011.

languages. A language barrier is created when individuals attempt to communicate in their respective native languages; and this barrier prevents the individuals from communicating[1,2,3]. Therefore, several attempts have been made to overcome this language barrier. For example, Language Grid[4,5] is an infrastructure that combines machine translation engines, parallel corpora, and so on.

In the medical field, imperfect communication can have detrimental effects. That is to say, even a slightly poor communication can result in medical errors. In particular, Japanese medical workers are not likely to communicate sufficiently with foreign patients, which might lead to an increase in the number of medical errors. Presently, medical interpreters accompany foreign patients to bridge the communication gap. However, a dearth of medical interpreters suggests that an improvement in the current situation is unlikely.

A multilingual medical reception support system named M^3, which is based on parallel texts for foreign patients and uses information technology, is used in the medical field[6]. M^3 supports a dialog and inquiry mechanism for medical support using parallel texts. M^3 operates from a pre-prepared set of symptoms; therefore, it can only respond to those exact symptoms. Thus, under this system, it is difficult for foreign patients to communicate a symptom that is not present in the system or be specific about a symptom.

Generally, patients fill in an interview sheet on their first visit to a medical facility. Most interview sheets provide an area for a free description of the problem. However, many medical facilities do not provide a multilingual interview sheet, although such a sheet is available on the Web[1] as a PDF file. We believe that this multilingual interview sheet has the following problems.

1. Patients have to select a symptom in advance. For this reason, it is difficult for patients to communicate their symptoms completely and specifically.
2. If a foreign patient writes his/her symptoms in their native language, it may be difficult for the medical workers to understand what was written.
3. There are different requirements between medical facilities in the content and style of the interview sheets; however, the multilingual PDF interview sheet cannot be changed.

To resolve these problems, we have developed a Web-based multilingual interview-sheet composition system. This system uses a parallel corpus and machine translation to input a patient's symptoms in detail. Moreover, it can support customization of the interview sheet.

2 Design of System

This section explains the design of our multilingual interview-sheet composition system. This system is a Web-based system developed using PHP and JavaScript.

[1] http://www.k-i-a.or.jp/medical/

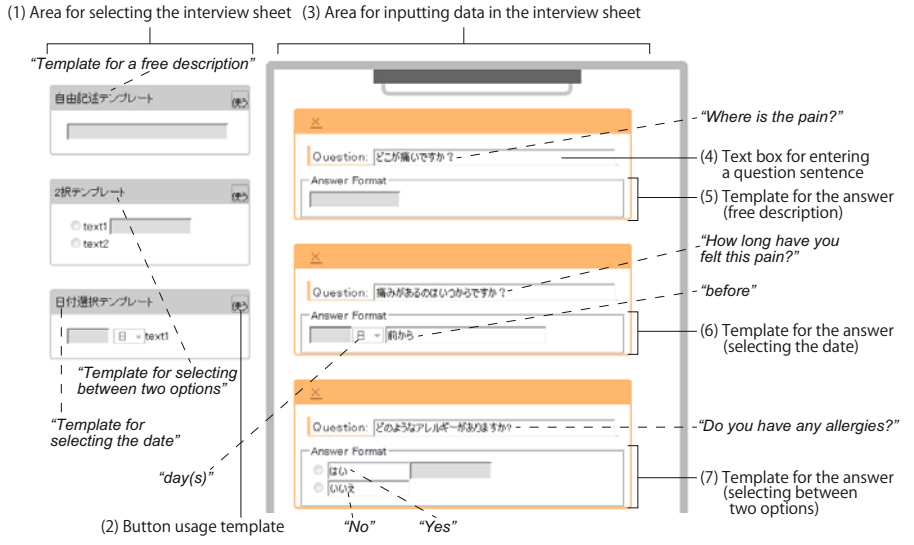


Fig. 1. Screenshot of function of creating interview sheet

2.1 Function for Creating Interview Sheet

This subsection describes the function for creating the interview sheet. It can be used to create multilingual interview sheets for different medical agencies. Medical workers can use this function.

Figure 1 shows a screenshot of the function. The function has two areas: an area for selecting the type of interview sheet templates (Figure 1-(1)) and an area for inputting data into the interview sheet (Figure 1-(3)).

The area for selecting interview sheet templates (Figure 1-(1)) shows template catalogs of interview sheets. The templates have a text area for miscellaneous descriptions, a radio button form, and so on. The medical worker tasked with creating the interview sheet selects the required template and then clicks on the usage button, as shown in 1-(2). The selected template is then shown in the area for creating the interview sheet (Figure 1-(3)).

The medical worker describes the question items as they create the interview sheet (Figure 1-(3)). The question items include a text box for entering the question (Figure 1-(4)) and a template for providing the answer (Figure 1-(5), 1-(6), and 1-(7)). The medical worker has to write in the text box to input a question item (Figure 1-(4)), answer item in the date format (Figure 1-(6)), and answer item with a choice of two options (Figure 1-(7)).

To allow multilingualization, the following protocol is proposed.

1. The system searches for parallel texts by using the data provided.
2. The user selects the appropriate text as displayed on the interview sheet.

This function can use only parallel texts. Therefore, it can be used to create a correct multilingual interview sheet containing the medical worker's language

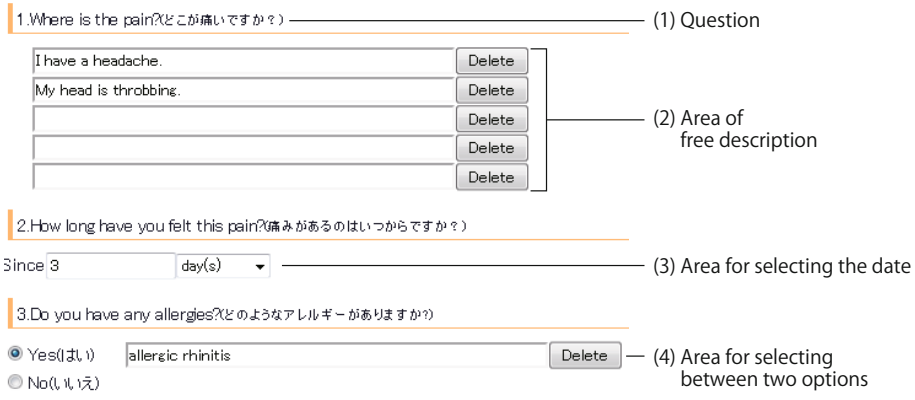


Fig. 2. Screenshot of function for inputting data into interview sheet

and the foreign patient's language. The method for selecting the parallel text is explained in section 2.2. Note that the system also enables question items to be reordered and deleted.

2.2 Function for Entering Data into Interview Sheet

This subsection describes the function for entering data into the interview sheet. We believe that this function would most often be used in the reception area of a medical facility.

Foreign patients use this function to input data to the interview sheet created by the function previously described.

Figure 2 shows a screenshot of this function. The question items of Figure 2 correspond to those shown in Figure 1. However, the language used in Figure 1 is the native language of the medical workers (in this example, Japanese) and the language used in Figure 2 is the native language of the foreign patient (in this example, Chinese).

The foreign patient enters his/her symptoms in the appropriate areas. The answer input areas show the answer format created in section 2.1. Figure 1-(6) and Figure 2-(2), Figure 1-(7) and Figure 2-(3), and Figure 1-(8) and Figure 2-(4) are all corresponding entries.

The system shows a translation display area when the foreign patient clicks in the text area, as shown in Figure 2-(2), Figure 2-(4), etc. Figure 3 shows a screenshot of the translation display. If the foreign patient clicks in the text box of Figure 3-(1), the system displays Figure 3. Figure 3-(2) shows similar parallel texts in the parallel corpus and the translations made by the machine translator.

Figure 4 shows the translation concept for this system. The translation in this system uses the parallel texts and machine translator. (However, the function for creating an interview sheet uses only the parallel texts.)

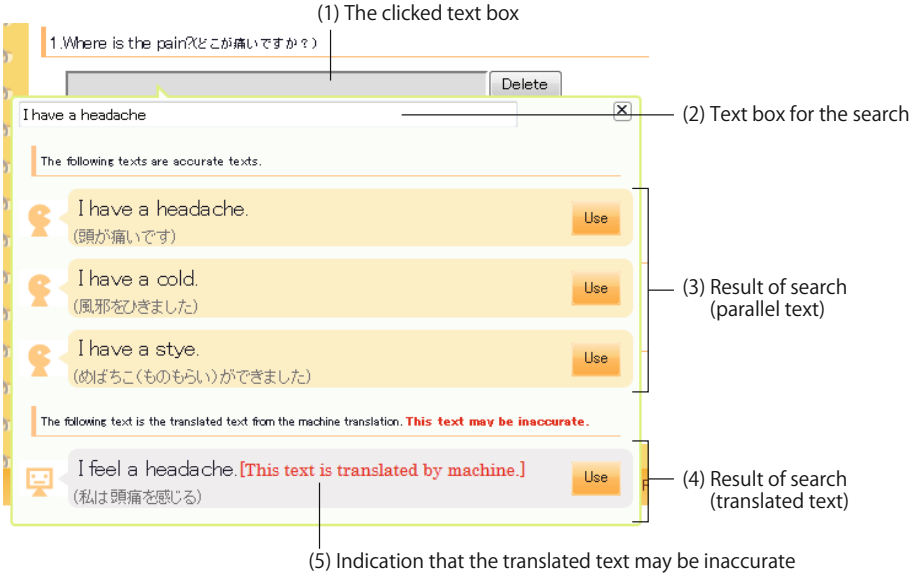


Fig. 3. Screenshot of function for selecting translation

Figure 3-(3) shows parallel texts that result from the search. Three texts are provided in this example, and Figure 3-(4) shows the translated text. However, the translated text may be incorrect, as shown in Figure 3-(5). Therefore, Figure 3-(4) also shows the reverse translation of the text, called the back translation [7].

Finally, when the submit button is clicked, the parallel text or the translated text is shown in the text box in Figure 3-(1). In this way, the patient is able to create a multilingual interview sheet.

When the foreign patient fills out the interview sheet and clicks the button "Prepare a medical interview sheet," this system create a PDF type of multilingual interview sheet. The created multilingual interview sheet displays the medical worker's language and the foreign patient's language.

2.3 Parallel Corpus and Machine Translation

This subsection explains the parallel corpus and machine translation that is used in this system.

We have developed a multilingual parallel-text sharing system named TackPad to collect and share parallel texts [8]. TackPad can collect parallel texts in the following nine languages: Japanese, English, Chinese, Korean, Portuguese, Spanish, Vietnamese, Thai, and Bahasa Indonesia. TackPad has about 7000 example sentences and is a Web-based system. The parallel texts of TackPad are evaluated for accuracy. TackPad users can evaluate the parallel texts that they exchange with each other. The parallel corpus used by this system is created by TackPad, and the machine translation is provided by Language Grid [4].

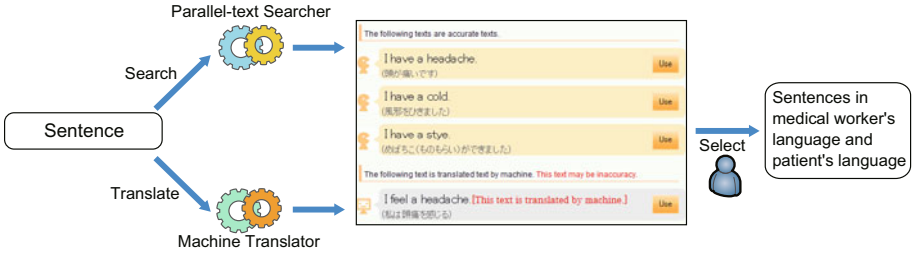


Fig. 4. Concept of translation for this system

3 Trial Experiment

3.1 Experiment on System's Usability

Experiments were performed regarding the practical use of this system. These experiments did not investigate the function for creating an interview sheet. We will investigate this function another time.

Trial users (playing the role of foreign patients) consisted of ten Chinese foreign students. The system displayed the medical worker's language and the foreign patient's language. In other words, it was assumed that the users of this system were unable to understand any Japanese at all. However, because the Chinese foreign students chosen could in fact understand Japanese, the system language was set to be Chinese only.

The following steps describe the overall flow of the experiment.

1. Explain the experiment
   The users entered data but were encouraged not to disclose secret personal information.
2. Filling in the interview sheets on paper and using the system (first turn)
   The users filled in the interview sheets both on paper and using the developed system, in Chinese. We requested that the same symptoms be written in both cases. The remaining two groups performed the tasks in the reverse order.
3. First questionnaire
   We requested that they complete a questionnaire about the paper and the developed systems.
4. Filling in the interview sheets on paper and using the system (second turn)
   We requested that the users then enter different symptoms than in the first turn.
5. Last questionnaire
   We requested that they fill out one more questionnaire regarding the paper and developed systems.

We assumed that during the first turn, the users represented foreign patients who had never operated the system before, and that during the second turn, they represented accustomed foreign patients.

Table 1. Questions used in the interview sheet for the experiment

	Question	Format of the question
1	What symptoms do you have?	Free description
2	In what region of the body are your symptoms?	Free description
3	How long have your symptoms persisted?	Input number and pull down menu
4	Do you have any food or drug allergies?	Radio button and free description
5	Are you taking any drugs now?	Radio button and free description

In the experiment, these questions were displayed in Chinese.

Table 1 shows the question items of the interview sheet used in the experiment. These questions were only displayed in Chinese.

The parallel texts of TackPad included the question items. Therefore, if parallel texts included a "?" character, they were not displayed. There were 314 parallel texts in the experiment.

3.2 Evaluation of Data Entered

This subsection explains how the data entered in section 3.1 were evaluated. Six students (three Japanese and three Chinese) were used as evaluators. We selected Chinese students who had passed Level 1 of the Japanese-Language Proficiency Test or had equivalent proficiency [2,3]. A different set of people was chosen compared to the users in section 3.1, especially the Chinese users.

We evaluated the accuracy of their translated messages by using the adequacy evaluation method developed by Walker [9]. In this method, a five-point scale is used to evaluate the translation accuracy according to the following grades: 5 = All; 4 = Most; 3 = Much; 2 = Little; and 1 = None.

If the average result was 4 or under, we judged the translation to be inaccurate.

4 Considerations

This section explains the experimental considerations of section 3.

4.1 Percentage of Parallel Texts and Translated Text

This section discusses the parallel texts and the translated texts provided by the machine translation. Table 2 shows the number of parallel texts and the number of translated texts used in the experiment. We found that users tended to use

[2] http://www.jlpt.jp/e/

[3] Level 1 of the Japanese-Language Proficiency Test is defined as follows: "One is able to read writings with logical complexity and/or abstract writings on a variety of topics, such as newspaper editorials and critiques, and comprehend both their structures and contents".

Table 2. The number of parallel texts and machine translations used

	First experiment	Second experiment	Sum	Rate
Machine translation	39	47	86	81.1%
Parallel-text	5	15	20	18.9%
Sum	44	62	106	100.0%

The data represent the number of sentences.

the translated text. In this experiment, users tended to use the sentences for the type of answer. However, TackPad has more parallel texts for the type of question than for the type of answer. For this reason, we think users tended not to use the parallel texts.

We believe that the inputted machine translation text will be used again. Therefore, we will connect this system to TackPad. The following steps describe the overall flow of the method.

1. This system sends only inputted machine translation texts to TackPad. (This system does NOT send translated text. There is a possibility that this will be inaccurate.)
2. Users of TackPad create and evaluate multilingual parallel texts based on the inputted text.

4.2 Accuracy of the Translation of the Inputted Texts

This section discusses the translation accuracy between the parallel texts and the translated texts. Table 3 shows the percentage of inaccurate sentences between the Japanese and Chinese languages. In Table 3, the machine-translated text was inaccurate 14% of the time.

One such example of an inaccurate translation involved “腿” (Chinese) and “足” (Japanese). “腿” means “from the crotch of foot to the ankle,” whereas “足” means “from the crotch of foot to the tiptoe.” Therefore, the translated text (“腿” and “足”) was inaccurate. It was noted that the words used to describe various regions of the body could differ slightly in meaning across different languages, which could lead to a false diagnosis.

We think that this is a serious problem in the medical field. We propose a method involving the depiction of the human body on a chart. In this method, the patient annotates the affected area of the body on the chart. Using this method, we think that the likelihood of a false diagnosis would decrease.

4.3 Actions of Users

This section discusses the actions of the users of the system. Table 4 shows the time required to complete the interview sheets on paper and using our system. We discovered that the interview sheet on paper took less time than our system. We believe that the inputting and translation of the text caused this difference.

Table 3. The percentage of inaccurate sentences between Japanese and Chinese translations

	Number of used	Inaccurate	Rate of inaccurate
Machine translation	86	12	14.0%
Parallel text	20	1	5.0%
Sum	106	13	12.3%

The numbers in this table are the number of sentences.

Table 4. Time to write interview sheets on paper and using the system

	First experiment		Second experiment	
	Paper	System	Paper	System
Average	154.7	293.4	112.0	201.8
Standard deviation	60.04	203.25	48.75	82.78
Minimum	66	150	30	82
Max	258	820	197	341

The numbers in this table are seconds.

However, the difference in time was shorter on the second turn than the first turn. We therefore believe that accustomed users can use the developed system quickly. However, more studies are required across a wider range of users.

5 Conclusion

In this study, we developed a multilingual interview-sheet composition system and evaluated its usefulness.

The contributions of this paper are as follows:

1. We have proposed a multilingual interview-sheet composition system that can be used for communication between medical workers and foreign patients. We have developed this system using both parallel texts and machine translation.
2. We showed that a patient is able to create a multilingual interview sheet using a parallel corpus and machine translation.
3. Because it may be difficult to indicate an affected area of the body using words alone, we suggest that affected areas be indicated by the user using a human-body image for more accurate communication.

We can work together with TackPad to ensure that the inputted machine translation texts are available for multilingual parallel texts. The functions for inputting data and creating interview sheets should be evaluated across a wider variety of users.

Acknowledgment. This work was supported by SCOPE (Strategic Information and Communications R&D Promotion Programme) of Ministry of Internal Affairs and Communications.

References

1. Takano, Y., Noda, A.: A temporary decline of thinking ability during foreign language processing. Journal of Cross-Cultural Psychology 24, 445–462 (1993)
2. Aiken, M., Hwang, C., Paolillo, J., Lu, L.: A group decision support system for the Asian Pacific rim. Journal of International Information Management 3(2), 1–13 (1994)
3. Kim, K.-J., Bonk, C.J.: Cross-Cultural Comparisons of Online Collaboration. Journal of Computer Mediated Communication 8(1) (2002)
4. Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration. In: IEEE/IPSJ Symposium on Applications and the Internet (SAINT 2006), pp. 96–100 (2006)
5. Sakai, S., Gotou, M., Tanaka, M., Inaba, R., Murakami, Y., Yoshino, T., Hayashi, Y., Kitamura, Y., Yumiko Mori6, T.T., Naya, Y., Shigeno, A., Matsubara, S., Ishida, T.: Language Grid Association: Action Research on Supporting the Multicultural Society. In: International Conference on Informatics Education and Research for Knowledge-Circulating Society (ICKS 2008), pp. 55–60 (2008)
6. Miyabe, M., Fujii, K., Shigenobu, T., Yoshino, T.: Parallel-text Based Support System for Intercultural Communication at Medical Receptions. In: Ishida, T., R. Fussell, S., T. J. M. Vossen, P. (eds.) IWIC 2007. LNCS, vol. 4568, pp. 182–192. Springer, Heidelberg (2007)
7. Miyabe, M., Yoshino, T., Shigenobu, T.: Effects of Undertaking Translation Repair using Back Translation. In: Proceedings of the 2009 ACM International Workshop on Intercultural Collaboration (IWIC 2009), pp. 33–40 (2009)
8. Yoshino, T., Fukushima, T., Miyabe, M., Shigeno, A.: A Web-based Multilingual Parallel Corpus Collection System for the Medical Field. In: Proceedings of the 2009 ACM International Workshop on Intercultural Collaboration (IWIC 2009), pp. 321–324 (2009)
9. Walker, K., Bamba, M., Miller, D., Ma, X., Cieri, C., Doddington, G.: Multiple-Translation Arabic (MTA) Part 1, Data Consortium, Philadelphia (2003)

User Modeling-Based Spatial Web Personalization

Hadjouni Myriam[1], Baazaoui Hajer[1],
Aufaure Marie Aude[2], and Ben Ghezala Henda[1]

[1] Riadi-Gdl Laboratory, ENSI Tunis, Campus of Manouba La Manouba, Tunisia
{myriam.hadjouni,hajer.baazaouizghal,
henda.benghezala}@riadi.rnu.tn
[2] MAS Laboratory, SAP Business Objects Chair, Ecole Centrale Chatenay-Malabry,
marie-aude.aufaure@ecp.fr

Abstract. Web personalization can be seen as an interdisciplinary field whose objective is to facilitate the interaction between web content and user's needs. It includes per definition several research domains from social to information sciences. The personalized search focuses on integrating the user context and needs in the information retrieval process. In this context, this paper presents a user modeling based spatial web personalization system that constructs a users' model network. The idea is to provide user with personalized results based on his model and on the neighbour users' models.

Keywords: web personalization, user modeling.

1 Introduction

The continued expansion of web documents number and type has contributed to the difficulty, for web users, to find relevant resources that respond to their requests. In fact, data retrieval from the Web has become a non-straightforward task. As finding pertinent data is becoming difficult, information retrieval (IR) systems have to be personalized. This means personalize the results set displayed, and analyse the fundamental differences between users [16]. In fact, using personalized systems address the "user lack of pertinent returned data" problem by providing an adaptive and intelligent Human-Computer-Interaction and by the improvement of information systems usability. The main interest of these systems when extracting information is the use of a process that considers end-users interests and preferences [19]. In this context, we aim to present an information retrieval system (IRS) for Web personalization that is based on users' network construction and on user modeling. The proposed system construct a multidimensional user model and aims to provide personalized results that best meet users' needs (1) according to their preferences and (2) considering the preferences of the other system users'. In fact, this system relies upon the construction of a user models based graph. Our assumption is that when searching for a relevant document, the retrieval system should use, in addition to specific user needs and previous searches, knowledge on the other users. The paper is organized as follows: the second section covers an overview of the state of the art.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 41–50, 2011.

The third section presents the proposed system and its components. In Section 4, we present the experimentation. We conclude the presented work in the last section.

2 Related Works

The main interest of personalized systems when extracting information is the use of a process that considers end-users interests and preferences. The personalization process requires representing, accessing and storing the users' related information [14]. We present in this section an overview of web personalization and user modeling.

2.1 Web Personalization

Existent personalization approaches have contributed to the improvement of information systems use. However these approaches have weaknesses and limitations. In fact, as the content of Web pages is being more complex and inter-related, classical personalization systems need to be improved. This has led to an interest on integrating semantic knowledge in the personalization and retrieval processes [17]. The personalization process has been enriched at the semantic level, based on user modeling [14] and on log files analysis [7]. Several approaches, like the collaborative ones, do not consider some specific users preferences when they represent a minority in a given group. In another hand, content based approaches facilitate items retrieval by proposing some alternatives and recommended items similar to the one the user is visiting. However these approaches focus only on the user's actual and temporary needs and cannot highlight the items that are related to the current query results. Other approaches try to determinate the interests of each user but they are limited by their items model that does not describe the differences between items properties. This lack of semantic description of the items decreases the quality of personalization since similarities and dissimilarities between items can't be measured accurately.

Another effect of the Web evolution is the use of geo-referenced entities that are more and more present in Web resources [20]. Usual Web personalization approaches, such as collaborative filtering or content-based filtering, are not tailored to the complexity and to the semantic of spatial information available on the Web. However, spatial information personalization on the Web must consider spatial properties and relationships found in Web documents. Additional spatial properties derived from demographics, interests and preferences in space should be considered for the design of Web systems. The design of spatial Web applications, as the design of Web applications, requires a user model and associated user preference elicitation mechanisms. In this context, a personalization engine combining spatial and semantic criteria as well as a user interface enriched with spatial components has been proposed by [10]. To provide relevant results to the user, [21] explores semantic similarity and spatial proximity measures as well as relevance ranking functions on the behalf of the user. Semantic similarity is the evaluation of semantic links existing between two concepts [15]. In [11], Larson and Frontiera introduced a classification algorithm for measuring the spatial proximity between two regions.

2.2 User Modeling

In the context of providing more accurate search results to the user, personalization systems and personalized information retrieval systems integrate the user model in their working process. In fact, the end user is represented by a model which contains necessary information to make personalization. This collection of data about the users is often referred as user models. Hook [6] introduced a user model oriented toward the integration of "knowledge about the user explicitly or implicitly coded, used by the system to improve interactions". The easiest way to model the user is to save what he knows or does not know. This implies to represent and to store its most relevant characteristics that are in the context of the application domain. Acquisition of the user data, in an information retrieval process, is used to perform query reformulation [9]. [12] proposed an approach based on user categories and profiles inference. Other approaches also consider social-based filtering [13] and collaborative filtering [20]. These techniques are based on relationships inferred from users' profiles. Another representation of the user model is based on using ontology [18].

3 User Modeling-Based Spatial Web Personalization

Introducing the user within the personalization process involve its modeling [8]. In fact, the reason of quality retrieval lack is the use of user model without context. Users can have general, recurrent and stable preferences and usually, systems use a subset of this data. In fact, these systems usually use explicit user preferences formulation. We aim, with our proposition, to have a user model mainly constructed with implicit data. This data is dependent only with user search context and navigations. And as most of current search systems do not integrate the spatial data, we also use user spatial data combined with documents spatial data to enrich the user model. Adding to this, we take benefit of a users' models network construction based on inter-nodes similarities. These similarities are based on semantic and spatial data contained in users' models.

As our main objective is to give the end user with personalized search result, the system (SyRIPs) first builds users models. It connects then these models to each other to construct a users' models network. The assumption is that when searching for a relevant document, the search system should use, in addition to specific user needs and previous searches, knowledge on the other users. In fact, the search personalization is achieved by re-ranking search results returned to the user with respect to its query and using the user model. The overall process supporting the personalization system is presented in the next subsection. Regarding our objectives, we built the system upon three main components: (1) the user model construction component, (2) the network construction component and (3) the user-system interaction component.

The user model is based on an implicit interaction with the user: implicit because the user isn't directly asked to give opinion, and interactive because we use the navigation to measure its interest to a given entity. These measurements are based on:

- The similarities that could exist between attributes and entities of interest.
- The deduction of user interests from all its navigation.
- The calculation of the pertinence of a supposed spatial move.

In SyRIPs, we also consider that when a user is interested on a spatial zone, he aims to move there. In fact, we think that before going to one place, a user may search information on it.

3.1 The Users' Models' Network Description

The models' network topology is hierarchical and is layered (*cf.* Figure 1): the first layer is composed of the users' models nodes and the second is the based on the unknown users' stereotypes. The edges relying nodes are calculated using semantics and spatial distances calculation [5]. Choosing these two distances is based on the assumption that a user usually performs a text search and needs spatial information. In the network representation, to each arc connecting two nodes *x* and *y* is associated a set of values *w (x, y)* (spatial and semantic). We represent it using an n*n adjacency matrix, with n the number of users' models nodes.

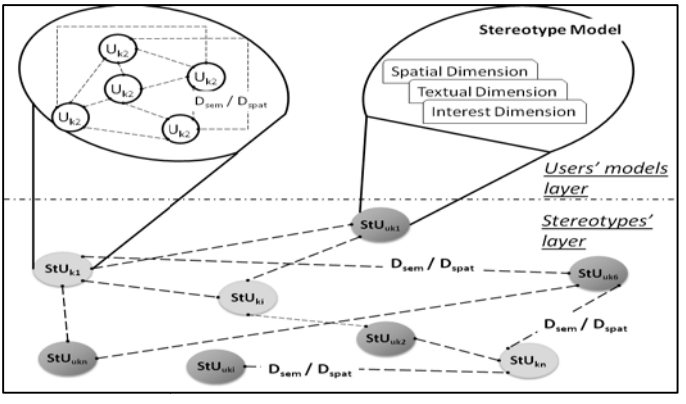


Fig. 1. The users' models network

3.2 The Proposed User Model Dimensions

The user model proposed is based on the use of interactive and implicit data recorded from user interactions with the system. The model is a multidimensional one and is composed of the dimension bellow:

The Textual Dimension represents user textual searches. This dimension is closely related to the interest dimension. In fact, the latter is mainly constructed using data from the textual dimension. The textual data is represented by vector that contain the term written by the user and the frequency of use. The user navigation history is also contained in this dimension. In fact, to each query is assigned the URLs clicked and the time spent on the document. This feedback is used in the implicit calculation of the user interest.

The Interest Dimension is the representation of user interests. This dimension is constructed using the textual dimension data and global domain ontology[1]. It is represented by a dynamic preference ontology that is constructed while user asking and navigating. The concepts weight is the implicit user interest. The interest degree I_e of a user u towards a visited item x, considering only the semantic aspect of x is calculated using the duration of the user visits to x, the total duration, the number of times the user visited x and finally the total visits number [5]. Ontology construction is based on the use of two semantic sources: The ODP Ontology[2] and the WordNet[3] data dictionary. The ontology learning technique and construction used were proposed in [1]. This technique was developed for the ontology OntoCoSemWeb[4].

The Spatial Dimension contains spatial weighted user interests. The weight is the affected value corresponding to the implicit user interest on the spatial data calculated using the formula presented in [5]. This data is collected from different inputs: the textual query, the selection of a position on a map-in the user page- and navigations within the results displayed. Spatial data is represented by entities, concepts and attributes. The spatial entities are classified into concepts.

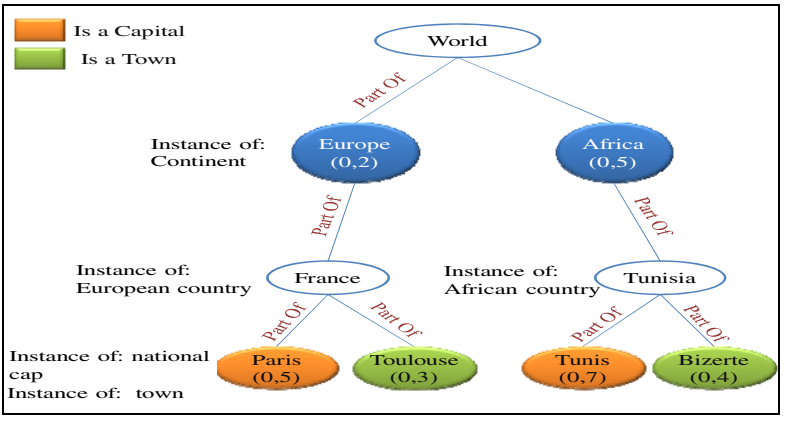


Fig. 2. A visualization of our spatial locations ontology

These concepts belong to a specified domain of interest. Each spatial concept is defined by attributes used to describe the entities belonging to it. For example if the domain of interest is Tourism: a spatial concept should be Hostel, the corresponding entities are stars number, location and location attributes near a beach, Nice[5], and so on. To enrich this dimension, we also use the spatial data of WordNet. An example for locations representation within this dimension is given by figure 2. It is an example of ontology of spatial locations. This spatial information is used firstly for

[1] WordNet, Url: http://wordnet.princeton.edu/
[2] http://www.dmoz.org/
[3] http://wordnet.princeton.edu/. Url: http://wordnet.princeton.edu/wordnet/download/
[4] This work and the referred one are integrated in the same bilateral project: DGRST-INRIA-STIC.
[5] Nice: A town in France.

the query reformulation. In fact, we use the weighted concepts to enrich the query when the user asks for some spatial location or concept existing in his spatial dimension. When it does not exist, we use real geographic distances between places to improve the spatial target search zone. Secondly, spatial information is used when the neighbour node is searched. In fact we use spatial proximity between the network nodes to extract the suitable nodes that allow result search amelioration with implicit collaboration.

The User Personal data Dimension is concerned with users registered to the system. It contains basically the user login and password. We added to this data the IP address and, with the user permission, the use of cookies.

3.4 The User Model Construction Process

The user model proposed is based on the use of interactive and implicit data recorded from user interactions with the system. This data collection is called implicit because the user isn't directly asked to give information on his preferences, and interactive because we use navigation to measure its interest to a given entity.

The first interaction of the user with the system and evolves with next searches and navigations. We consider the two cases: the user is logged and has its own profile and a user is not logged or not known by the system. For the first case, the correspondent model is activated and used and update with current navigations.

If we consider M_u *(t0=0)* this user model at activation time, and having the current user search – implicit information about his preference at that moment and it could give a spatial information if the user locates its area of interest in the map or in the query–, we will have at the next navigation (and in the same current session):

$$M_u(t_1)=M_u(t_0)+Z_i+Sim(x)+I_e(u,x) \quad (1)$$

where $t_1=t_0+\delta$

Z_i The user target search zone, it is the spatial search zone

Sim(x) Similarity that could exist between attributes of the search results and the entity of interest

$xI_e(u,x)$ The deduction of user interests.

4 Experimentation and Results Interpretation

To perform experimentation and evaluation a prototype supporting the proposition was developed [3]. Our objective is to evaluate the search results according to the users' needs. First, the experimental users were asked to notify their corresponding relevant results and then precision was calculated. Google Api was integrated to perform web search using the Google index, and then to compare results with Google's one. The evaluation protocol [2] was designed to tune the experimentation parameters and then to evaluate the effectiveness of our system's personalized search. It is based on two stages: a learning process and the experimentation. The learning process, which aims to construct the necessary data about the user, begins with inquiring users' queries on the basis of the experimental topics (three defined ones and

the fourth is depending on own users independent preferences). Interactions and navigations users are stored. Then, a subset of n-1 queries is extracted from those collected in the learning process. The n-1 queries represent a subset of test to perform with the n^{th} remaining query. From collected queries, the users' models are constructed and learned using, for each query, a subset of relevant documents. For the experimentation purpose, we asked users to check the answers that they consider as relevant. After, we evaluate queries and results corresponding to the user, using the experimental user notification to relevant documents.

The data used for our experimentations is a set of indexed documents. Our domain interest in this experimentation is the tourist one. We defined four scenarios and constructed our users query sets. Experiments were made with the collaboration of students and academic collaborators who are from different interest domains (such as literature, science, low…).

The experiments results presented in this paper concern the whole personalizing process. Stereotypes (for unknown users) construction and evaluation were presented in [4]. For these stereotypes construction and evaluation, Gallois lattices were used and allows concluding that (1) stereotypes constructed correspond to the experimental scenarios and (2) users corresponding to the free scenario are independent and do not belong to any stereotype.

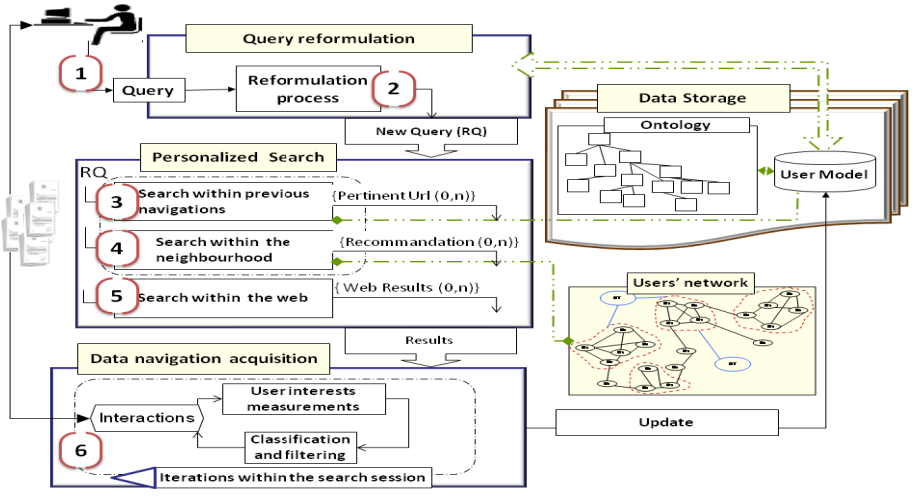


Fig. 3. The scenario process

4.1 The Scenario Process

The scenario (presented in figure 3) goes as follows. A user *Toto* is fond of all kinds of luxurious hotels that are based in Europe. His implicit preferences include five stars hotels, 4 stars restaurants and family rooms, among others.

Toto starts a search session with the query *q* "hotel sun fun". The system will first expand the query with interest information from the user model. The query q becomes "hotel sun fun luxury Europe".

This new query is then executed to search within previous user navigation to give the preferred results, and then to find the most correspondent neighbour user and finally, to retrieve other results from the web that will be ranked after the personalized results. And while user navigating throws the results set, its preferences and interest degrees are updated and this will have an impact on next searches within the same session.

4.2 Experimental Results

We attempt to achieve through our experiments the objectives of evaluating the effectiveness of our system over the various real users' requests. The number of pertinent documents used is 15. First results presented are the comparison of precision values between personalized and classic search, in Figure 4.

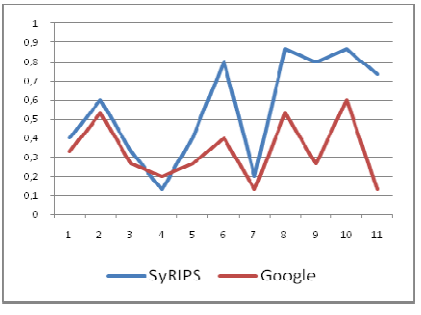


Fig. 4. Comparison of the P@5 between Google an SyRIPs

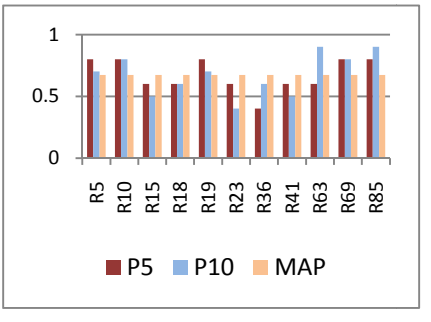


Fig. 5. Search results quality measurements after personalization

We notice that the precision at five of personalized search is better that the classic research, especially when the number of requests grow within the system.

The sheet in fig 5 presents search results quality measurements with personalization. We calculate the P@5, P@10 and the Mean Average Precision for a set of 11 queries. We can see in this sheet that the p@10 presents more accurate results. Moreover, we see in the table 1 that the performance improvement is better for P@10 or P@15 documents than that at P@30 documents. This amelioration can be explained by the fact that there is less non pertinent when considering the first 10 or 15 results. We can also conclude from this table that the user's model, including spatial dimension ameliorate the system precision.

Table 1. Average precision at top n documents

	AP@5	AP@10	AP@15	AP@30
SyRIPs	0.418	0.536	0.490	0.342
Google	0.290	0.4	0.272	0.248
Improvement	12.7%	13.7%	21.8%	9.4%

5 Conclusions

We began in this paper with an overview of the state of the art in personalized information retrieval and user modeling fields. Then we presented our proposal of a user modeling-based spatial web personalization system. The proposed system includes the creation of a users' models based network and the implicit construction of multidimensional user model. The user dimensions contain spatial users' data and implicit interest and preferences. Using this proposition, we developed a client-side web search system on top of a popular search engine. We conducted the experimentation and evaluation phase which involves the construction of users' models stereotypes network and the personalization process. In this paper, we presented the whole personalization system results. Evaluation shows that the system improves search accuracy over the web. The proposed system thus can improve web search performance without any additional effort from the user. Next experiments step will concern the users' models network. As the network is constructed considering implicit data and with evolutions over the search sessions, it will enhance the personalized search results.

References

[1] Baazaoui, H., Aufaure, M.A., Ben Mustapha, N.: Extraction of ontologies from web pages: conceptual modeling and tourism. Journal of internet Technologies (2007)

[2] Daoud, M., Tamine-Lechani, L., Boughanem, M., Chebaro, B.: A session based personalized search using an ontological user profile. In: Proceedings of the 2009 ACM Symposium on Applied Computing, SAC (2009)

[3] Hadjouni, M., Baazaoui, H., Aufaure, M., Claramunt, C., Ben Ghezala, H.: Towards personalized spatial web architecture. In: Workshop Semantic Web meets Geospatial Applications, Held in Conjunction with International Conference on Geographic Information Science (2008)

[4] Hadjouni, M., Baazaoui, H., Aufaure, M.A., Ben Ghezala, H.: Vers un système d'information pour la personnalisation sur le web basé sur la modélisation de l'utilisateur. Ingénierie des connaissances (2009)

[5] Hadjouni, M., Haddad, M., Baazaoui, H., Aufaure, M., Ben Ghezala, H.: Personalized information retrieval approach. Web Information Systems Modeling. In: Conjunction with the 21st International Conference on Advanced Information Systems: CAiSE (2009)

[6] Hook, K.: A Glass Box Approach to Adaptive Hypermedia. PhD thesis. Stockholm University (1996)

[7] Jrad, Z., Aufaure, M., Hadjouni, M.: A contextual user model for web personalization. In: Weske, M., Hacid, M.-S., Godart, C. (eds.) WISE Workshops 2007. LNCS, vol. 4832, pp. 350–361. Springer, Heidelberg (2007)

[8] Kobsa, A.: User modeling and user-adapted interaction. User Modeling and User-Adapted Interaction 15(1-2), 185–190 (2005)

[9] Koutrika, G., Ioannidis, Y.: A unified user-profile framework for query disambiguation and personalization. In: Workshop on New Technologies for Personalized Information Access, Held in Conjunction with the 10th International User Modeling (2005)

[10] Kuhn, W.: Handling data spatially: Spatializing user interfaces. In: Proceedings of 7th International Symposium on Spatial Data Handling, SDH 1996, Advances in GIS Research II, vol. 2, pp. 13B.1–13B. 23. IGU (1996)
[11] Larson, R., Frontiera, P.: Spatial ranking methods for geographic information retrieval (gir) in digital libraries. In: Heery, R., Lyon, L. (eds.) ECDL 2004. LNCS, vol. 3232, pp. 45–56. Springer, Heidelberg (2004)
[12] Liu, F., Yu, C., Meng, W.: Personalized web search by mapping user queries to categories. In: CIKM 2002: Proceedings of the Eleventh International Conference on Information and Knowledge Management, pp. 558–565. ACM, New York (2002)
[13] Mladenic, D.: Text-learning and related intelligent agents: A survey. IEEE Intelligent Systems 14(4), 44–54 (1999)
[14] Razmerita, L.: Modeling behavior of users in adaptive and semantic enhanced information systems: The role of a user ontology. In: Adaptive Hypermedia and Adaptive Web-Based Systems 2008, Hannover, Germany (August 2008)
[15] Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11, 95–130 (1999)
[16] Rieh, S.: Investigating web searching behavior in home environments. In: Todd, R.J. (ed.) Proceedings of the 66th ASIST Annual Meeting, vol. 40, pp. 255–264. Information Today, Medford, NJ (2003)
[17] Schouten, K., Ruijgrok, P., Borsje, J., Frasincar, F., Levering, L., Hogenboom, F.: A semantic web-based approach for personalizing news. In: Proceedings of the 2010 ACM Symposium on Applied Computing (2010), pp. 854–861. ACM, New York (2010)
[18] Sosnovsky, S., Dicheva, D.: Ontological technologies for user modeling. Int. J. Metadata Semant. Ontologies 5(1), 32–71 (2010)
[19] Vallet, D., Castells, P., Fernández, M., Mylonas, P., Avrithis, Y.: Personalized content retrieval in context using ontological knowledge. IEEE Trans. Circuits Syst. Video Techn. 17(3), 336–346 (2007)
[20] Winter, S., Tomko, M.: Translating the Web Semantics of Georeferences, pp. 297–333. Idea Publishing, Hershey (2006)
[21] Yang, Y.: Towards spatial web personalization. PhD thesis, Ecole Nationale Supérieure d'Arts et Métiers (2006)

Context-Aware Website Personalization

Daniela Wolff, Marc Schaaf, Stella Gatziu Grivas, and Uwe Leimstoll

University of Applied Sciences Northwestern Switzerland
{daniela.wolff,marc.schaaf,
stella.gatziugrivas,uwe.leimstoll}@fhnw.ch

Abstract. Today there is a need in on-line shops to support the visitors of a web page appropriately by analyzing their current situation. In this paper we introduce a model which supports the identification of the content the user is interested in and the shopping strategy of the current session. We use context information extracted from enterprise data, content data, the current and the historical behaviour of the user. This allows us to learn more about the interests and needs of the user. We monitor and analyze this content information at runtime and use it for the adaption of the web site during the user's navigation.

Keywords: Semantic Technology, Personalization, Event processing architecture, Context.

1 Introduction

Sales people are quite successful in understanding their customers' needs and requirements and providing them with relevant information and offers. As stated by Moe, some shoppers (please note that in the following we use the notions of shoppers/users/visitors of a web page in the same manner) appear to be very focused on looking for a specific product and sales people may assist them to find what they are looking for. Other shoppers are "window shoppers". Sales people have the possibility to ignore them and let them continue or try to stimulate a purchase in the appropriate manner. This behaviour, however, cannot be easily transferred to the electronic commerce [16].

As electronic channels have the potential to become an important alternative avenue for enterprises to reach their customers and thus to generate sales [17] the ability to recognize the customer and to provide the relevant information and offers is essential.

The ability to customize each individual user's experience of electronic content is defined as personalization [15]. Usually, for the provision of individual information the user has to be identified either by cookies or login. The identification allows the gathering of user related data and their storing them in user profiles. These can be used to provide individual content or functions such as transaction histories, recommendations, and personal product catalogues. Adaptive hypermedia systems for instance build user models containing goals, preferences and knowledge about each individual user to provide personalized web sites [5]. Due to the restrictions of user profiles this approach does not support people who visit a web site for the first time,

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 51–62, 2011.

users who cannot clearly be identified, and customers who change their shopping behaviour. Other approaches find similar users (Collaborative Filtering) or similar products (Content-based Filtering). These so called recommender systems belong to the most well-known examples of personalization functions. If a user has selected a product, the system provides a list of products which other people are interested in who bought the same product. However, the current situation of the visitor is not analyzed. There is a need to support the users (visitors of a web page) appropriately after having analyzed their current situation.

To fill this gap we propose to use the context information of the users to analyze their current situation. According to Dey et al. (2000) context is all information, which can be used to characterize the situation of an entity [11]. To determine the relevant context information for the visitors we use the context elements of web pages which can be analyzed while the visitor is navigating through the e-commerce's web site. For instance, the pages can give information about products or other offers. Relationships between products can provide information about the users' preferences and needs. For instance, whether he/she is looking for a cheaper product or searching for information about components of a specific product.

While navigating through the web site the users' clickstream reflects a series of choices made both within a web site, like which page to visit, how long to stay, whether or not to make an online purchase and across web sites (e.g. which products a user visits) [6]. This navigation helps to interpret the user's situation.

During the project Semantic OBDE[1] we developed a model describing the context of a web site user. During run time a rule based system analyzes and interprets the visitor's context. Depending on the situation the web page is adapted. Because this analysis is time consuming but needs to take place in near real time we introduce an event driven architecture based processing system that allows a timely analysis of the clickstreams. This can be seen in contrast to classical database approaches where the relevant information is stored and at some later time retrieved to be processed in a batch. Our contribution consists of a proposal for a context model which supports the analysis of user clickstreams and the interpretation of the user's situation. Additionally we propose an architecture which analyzes, interprets and makes adaptations while the user is navigating through the web site.

This paper is organised as follows: Section 2 provides related works. A scenario which helps to identify the requirements is given in Section 3 followed by the context model in Section 4 and the architecture in Section 5. In Section 6 we conclude and give an overview of the next steps.

2 Related Work

Applications which use the context to automatically provide information and take actions according to the user's present context are called context-aware applications [4]. Usually, the context-awareness is used in the area of ubiquitous, wearable or mobile computing. Adaptive Hypermedia Systems are systems which use parts of the user's context to improve the usability of hypermedia. For the improvement the user's goals, preferences and knowledge are taken into account [5].

[1] This project was partially funded by the KTI project 11628.1 PFES-ES (Semantic OBDE).

The AHA! system is an Open Source software originally developed to support an e-learning on-line course with some user guidance. It maintains the user model and filters content pages and link structures accordingly [8]. It is inspired by AHAM, a reference model for adaptive web applications which bases on a domain model, a user model and a teaching model [9]. Because of its educational background the customization according to a broader interpretation of the context is not tackled.

A combination of a content management system and a context engine is proposed by Belotti et.al.. It allows developers to adapt the content, view, structure and presentation of Web applications according to runtime context information [3]. Hera is another approach which designs adaptive, dynamic Web applications. At runtime the user model is updated and the web application is adapted accordingly. However, the adaptation is based on a user model which does not tackle the broader interpretation of context [22].

Ceri et.al. use a context model [7] in addition to the user profile. They apply their context model on mobile and multichannel web applications, which leads to a context model containing elements like Device, Location, and Activity.

All these approaches consider only parts of the users' context. Analyzing a broader context leads to a better understanding of the present visitor and supports adapting the web site more ideally.

3 Scenario

For the identification of the requirements of an on-line shop, we use a case study of a wine-seller. Currently, the web site of the seller has no implemented personalization features. It only indicates the special offers and the different wine categories (like white or red wine). Selecting a wine category leads to the different countries of origin of the wines. When visitors choose a specific country they get a list of all wines of this country. The list shows a picture of the bottle, some information about the wine, like name, country, winery, and price. Clicking the link “Details” visitors get more detailed information about grape variety or vintage, or comments about taste and drink temperature, and what kind of food the selected wine fits best, as well as information about the winery and ranking. The page offers the opportunity to purchase the wine.

Today, the web site has a lot of disadvantages in supporting the different shopping strategies. The direct buying strategy is likely to result in an immediate purchase. For instance, a user wants to buy a specific red wine, e.g. "Primitivo Selezionato". First, the user has to find the specific wine he wants to purchase by selecting the correct wine category and country. Then, he/she has to search for the wine in a list of all wines offered. It takes several clicks to get to the specific wine (particularly since the web site offers no search function). If users do not know the country of the wine's origin, the list of wines is even longer.

Also, the web site does not support the search and deliberation strategy aiming at acquiring all relevant information to help users to make an optimal choice. Today the user has to find similar wines on his/her own, by selecting every wine and comparing different properties.

Regarding a hedonic browsing strategy, which is dominated by exploratory search behaviour, the page offers a good navigational structure due to the visitor's less focused search and his spending more time on viewing the broader category level pages. However, the inspiration is not supported currently.

Finally, the support of a buying strategy, which requires knowledge building with the objective to increase product and/or marketplace expertise, is marginal. The only way to get information about products is to select every wine and read the detailed information. If there are any updates the user has to find them.

To eliminate these weak points we propose a model describing the context of a web site user. Our model fulfils the following requirements:

- **Context identification** - The context of the user must be identified during runtime. This supports the appropriate adaptation of the web site. For instance, to support the objective of comparing the current wine with other similar wines, the system has to know in which wine the user is currently interested in. Then, the web page can be adapted by showing a list of links referring to similar wines. The more wines are visited by the users, the more refined the list would be.
- **Observation of changing context** - To detect changing interests the context has to be observed while the user is navigating through the web site (especially by using the hedonic browsing strategy).
- **Performing architecture** - For the adaptation of the web site during the navigation of the user the context has to be monitored and analyzed at runtime. Therefore, a performing system is necessary.

The next section describes the context model and how it is used during runtime. , In section 5 some architecture is introduced, which allows to analyze the users' context during runtime.

4 Context Identification

Systems are able to analyze and interpret the user's context only if the context has been made explicit [1]. Web sites provide very valuable context information about what the visitor is interested in. To identify the user's shopping strategy not only the content data but also the behavioural data has to be analyzed. Also demographic information and historical purchase behaviour provide important information about visitors. Van den Poel and Buckinx provide variables to distinguish whether a visitor wants to purchase or not, like the number of days since the last visit, the customer's gender or the fact whether personal information is supplied to the company [21].

Due to the fact, that the kind of offered product influences the usability [23], the enterprise has to be considered as well. Using enterprise architecture, the enterprise data provides information for instance about the goal, products on offer, business rules or processes. This information helps to determine the broader interests (context) of a visitor. This means that to be able to learn more about the user's interests, context information should be extracted from enterprise data, content data, the user's current behaviour and the historical behaviour. Whereas the historical and the current behaviour are data which can be used in several domains, the enterprise and content

data is highly dependent on the domain, and changes from enterprise to enterprise. To be able to analyze the data of a particular enterprise independently, a meta model is necessary. Because ontologies enable the reuse of domain knowledge, support sharing common understanding, separating domain knowledge from the operational knowledge, we use ontologies to describe the context.

To present our approach we focus in this paper on the modelling of the content and current behavioural data.

While the user is navigating through the web site we want to identify the user's interests and behaviour. Therefore, each user's behaviour and interests are analyzed using a "user profile" ontology, containing three concepts which represent the current session: the concept *session* is related to the concepts *interest* and *behaviour*.

For the identification of the interest the content ontology is used. Usually, the content is represented by the offered products. There already exist some general product ontologies, which can be reused (like eClassOWL[2]). So, each enterprise can refine the general one but can also model its own domain ontology. Referring to the wine-seller example, the wine domain would have to be modelled. For the description of the product wine we (re)use the wine ontology offered by W3C[3]. The user profile ontology is related to the concept product. So, to every step a visitor takes the relations are added from the interest to the product he is interested in.

For the analysis of the behaviour, in particular the shopping strategy, we rely on the different page types as proposed by Moe. Figure 1 shows the ontology describing the type of pages. Home (for the home page), information related pages, search pages, category pages and product pages. During the user's visit the percentage is identified, how often he/she visits an informational page, category page or product page. If the user visits informational pages more often, the shopping strategy is more likely to be a knowledge building. If the user visits a lot of category pages, the user seems to follow the hedonic browsing strategy. The distinction of administrative and non-administrative pages helps to differentiate, whether a user is a shopper or whether he/she wants to e.g. update his/her user profile. All types of pages are sub-classes of non-administrative pages. All pages are described as instances of a specific kind of page.

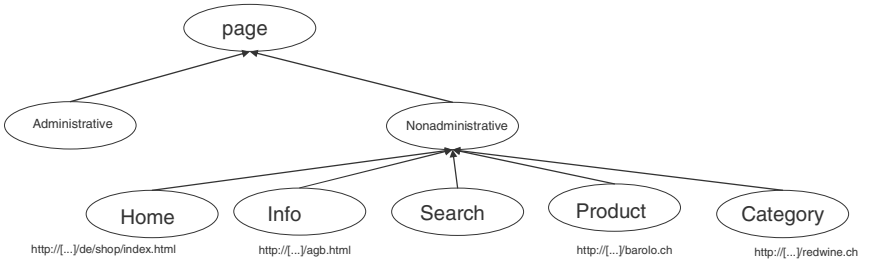


Fig. 1. Page Type Ontology

[2] eClassOWL http://www.heppnetz.de/projects/eclassowl/

[3] Wine Ontology: http://www.w3.org/TR/owl-guide/wine.rdf

Each web page is annotated with the terms specified in the page type ontology.

To combine the page type and the content we link each instance of the page type to the wine instances, which is shown exemplified in Figure 2. Using the object property "containsInformationAbout" indicates the content of each page.

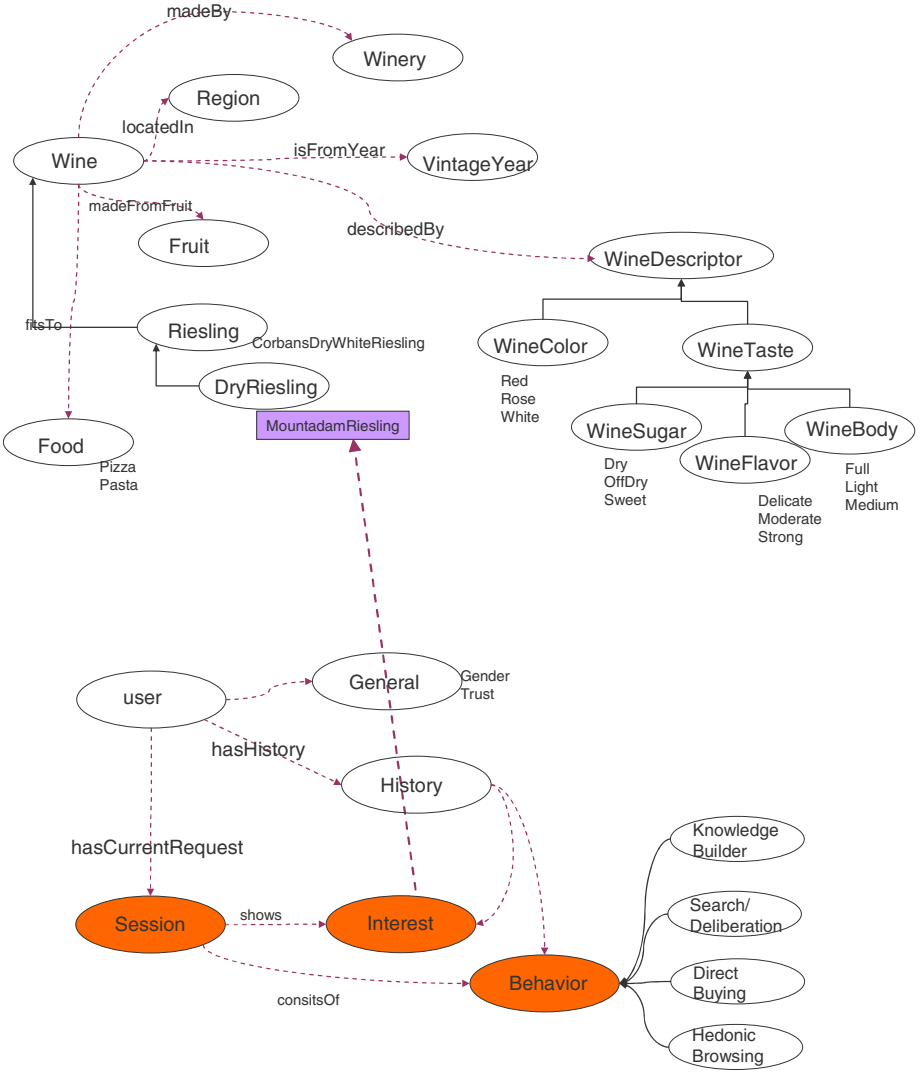


Fig. 2. Excerpt of the Context model

With this model we are able to identify the content the user is interested in and the shopping strategy of the current session. As mentioned before, to get additional information about the user his historical (purchase) behaviour, his demographics and the enterprise data have to be considered (if there is an existing user profile).

During runtime, a copy of the context model is made for each user. These context models are updated with every step the visitors take by adding instances to the fitting concepts. For example, for every product a user is interested in, the concept "interest" is linked to the referring product of the product ontology. His/her behaviour is analyzed accordingly by counting the type of web pages he/she visits. However, this analysis has to be extended to get a more detailed behavioural scheme of the user.

So, the presented model allows interpreting the situation and reacting accordingly. A causal relationship exists between use behaviour and usability, and usability and purchase behaviour [23]. Usability is inherently a subjective phenomenon [19] which is contingent upon both the task for which the system is to be used as well as the target users [13].

There exist two main parts of adaptation: internal adaptation, and external adaptation. Internal adaptation supports the usability. Usability can be associated with the aspects: content, ease of use, promotion, made-for-the-medium, and emotion [2]. Ceri et.al. identified the following internal adaptations: adaptation of content and services delivered by accessed pages, adaptation of navigation, adaptation of whole hypertext structure and adaptation of presentation properties. External adaptation means the context is used to adapt external applications, like newsletter or e-mail-services. For the external adaptation we use OWL-S[4] to describe the external application.

We use rules to combine the context and the (re)actions. On the condition part of the rules the context is defined. This context is analyzed during run time by a rule engine. If a specific context is kept the rules trigger the appropriate actions. Figure 3 shows a simplified rule, combining the context with a service. In this case the rule expresses that if a visitor uses the direct buying strategy and is interested in a specific wine and there is a special offer available for this wine, the visitor should get a newsletter.

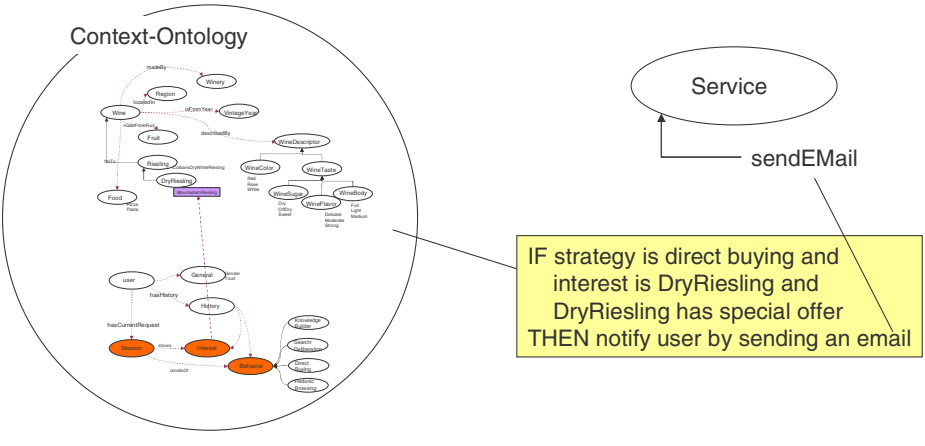


Fig. 3. Combining context with adaptation using rules

[4] OWL-S: http://www.w3.org/Submission/OWL-S/

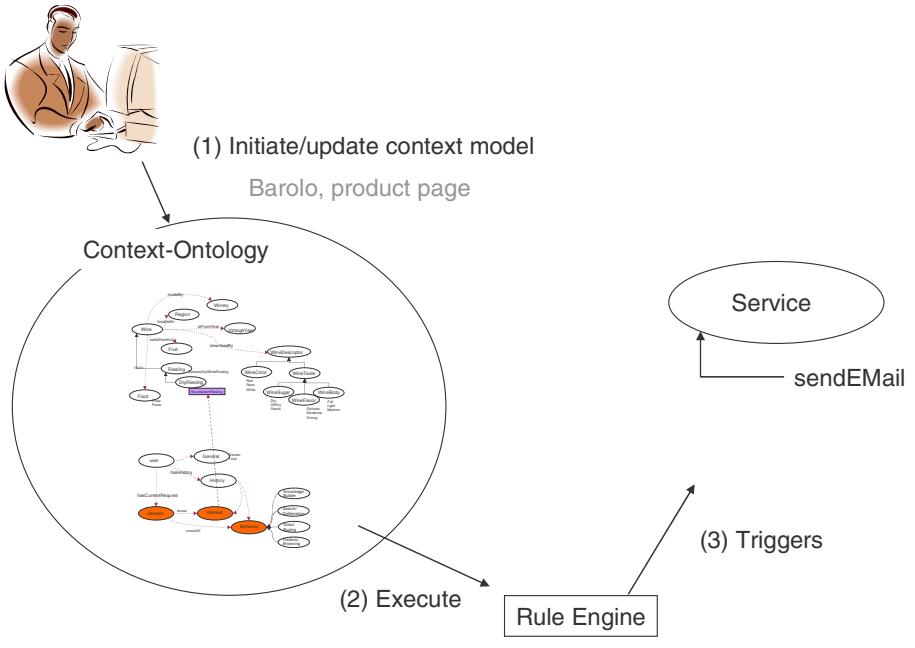


Fig. 4. Runtime steps

As SWRL combines OWL and RuleML we express the rules using SWRL[5]. For the execution of SWRL rules a rule engine was developed during the FIT-project[6] which is described in [12].

As illustrated in a simplified way in Figure 4, every step a visitor takes results in his context model being updated (1). In this case the system retrieves from the web page the annotated meta data: Barolo and "product page". With this information a link from interest to Barolo is executed in the context model and the product page count is increased. After the update of the model the rule engine is executed (2) and the current situation is interpreted. If the page count shows, that product and category pages are often visited, the system recognizes the visitor as a hedonic browser. According to his/her situation the rule engine triggers an additional service called sendEMail (3).

For the internal adaptation we use the conditional inclusion of fragments. So, similar to the AHA-Sytem [10] the web site of the wine-seller is adapted. Each web page contains an <if> tag representing the condition of the rules, e.g. `If (behaviour = „hedonic browsing“)`. If the rule engine returns true, then the operation tag specified in a <block> tag is executed. For instance, if the user is interested in a specific DryRiesling, the web page asks the systems to send a list of

[5] SWRL: http://www.w3.org/Submission/SWRL/

[6] FIT (Fostering self-adaptive e-government service improvement using semantic Technologies) is a project funded by the European Commission within the IST programme, IST-2004-27090.

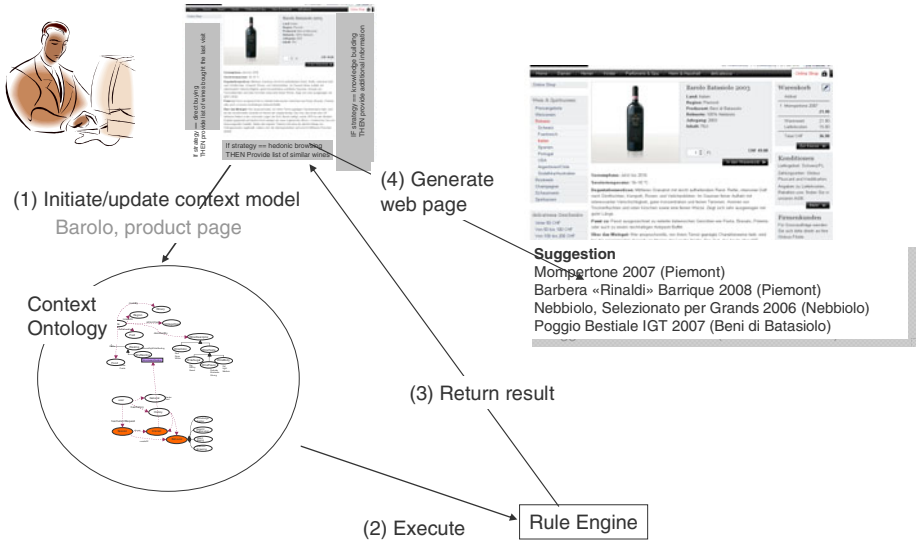


Fig. 5. Runtime steps for the internal adaptation

wines related to the ones the user already requested. Otherwise, if the user is a knowledge builder and is interested in the same wine, he will get a list of information pages about the wine. A simple scenario is shown in Figure 6. The user requests a web site about the Barolo. The context model is updated similar to the external adaptation. Then the rule engine is executed (2). The result is returned (3) and according to the result, the web page is updated (4). In this case a list of similar wines is presented to the user in order to inspire him.

The provided model is flexible and can be adapted to various enterprises. Additionally, rules have the advantage that each enterprise can create its own according to its customers, its products and so on. The first prototype implemented at the wine-seller's uses ca. 40 rules to adapt the web pages.

As already shown, there are a lot of analyzing steps for each visitor and updating steps of the model which have to be executed during run time. This requires a performing architecture, which is described in the next section.

5 Architecture

Our aim is to support the visitor on a web site by generating a profile that we create based on his/her clickstream in his/her current session. Therefore the processes for the profile generation and page adaption have to be executed rapidly. For this reason we defined an architecture that is based on the well proven concepts of event processing systems to allow the active event driven processing of a visitors clickstream without interfering with the content management systems (CMS) performance. An additional challenge we have to deal with is the scalability of the processing system across several systems to allow the processing of parallel clickstreams from various visitors in near real-time.

We define each user interaction with the CMS as an event which is signaled by the CMS. Each signaled event is processed by the filtering and analyzing components and results in updates of the visitor's context model. This model in turn is used by the CMS to personalize the content that is delivered to the visitor with his next request. Our processing architecture is defined by three different processing stages (Figure 6) which reflect the outlined processing flow:

1. **Event Filtering and Normalization:** In a first step, the events that can be signaled by the CMS in different ways must be normalized into the internal event format of our application. Afterwards each event is screened to allow the filtering of irrelevant events like those generated by crawling bots used by search engines to scan the pages of the CMS.
2. **Event Analysis:** The pre-filtered events are then mapped to the corresponding context model which are retrieved from the profile manager or created for new visitors. Based on the context from the user profile, the events are processed as outlined in the previous chapters, to derive the current user's interests and aims which are in turn written back to the user profile to make those results available to other components. Furthermore the analyzer raises an *interestUpdate* event when new knowledge was derived to notify the components of the third processing step.
3. **Action Execution:** The *interestUpdate* is used by the action engine to trigger the evaluation of custom rules which determine whether a concrete action should be invoked once a particular user interest was detected. E.g., sending an e-mail with a special offer to a customer.

Furthermore, we defined different possibilities for the integration of the CMS with the processing system. For the transfer on the user request we aim to support three different approaches to allow an easy and efficient integration with the used CMS: (1) monitoring of the access log files (2) direct integration into the CMS to notify the processing system directly of a users' request, similar to the AHA! system (3) notification by a specialized web application which is linked via a JavaScript fragment into the delivered content of the CMS.

For the realization of the actual page personalization, we provide a simple REST API that can be used by CMS components to obtain information on the current user and thus allow the CMS to personalize the delivered content.

The event driven communication between the different components is realized via a messaging system, namely Apache ActiveMQ, which provides the processing system with a reliable and well proven communication middleware. As the

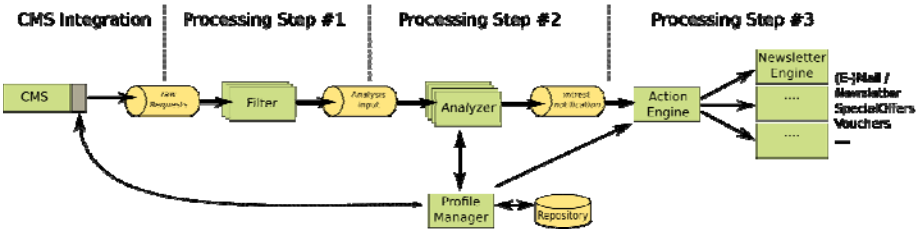


Fig. 1. The stages of the processing chain and the corresponding components

components in the different processing stages are communicating based on self-contained messages/events, the application can easily be scaled horizontally by deploying additional processing components on additional computers. Furthermore, we defined a specialized component concept that eases the development of the several processing components as it abstracts from the underlying communication infrastructure and provides access to the central profile storage of the profile manager.

With regard to the need of the analyzer components to access the context model, we facilitate mechanisms to stick the user sessions to particular analyzer instances. Due to this a concrete user's context model is only used by one particular analyzer for each user session. This allows us to implement efficient caching mechanisms and thus to prevent the possible bottleneck from sharing the context model between the various processing systems.

6 Conclusion

Our current work in the project Semantic OBDE focuses on the development of a model describing the context of a web site for a user after analyzing the user's clickstreams and interpreting the user's situation. This information can be used during the adaption of the web page. One challenge is the scalability of the processing system to allow the processing of parallel clickstreams from various visitors in near real-time. For this we are currently working on an event driven architecture based processing system. Our future work includes the implementation of the mentioned components and the evaluation of the concept based on real word scenarios. Furthermore, we are developing concepts for the integration of the required semantic technologies with the notion of event processing as it is realised for this project. In addition, we currently analyze the behaviour by getting the average number of visited page types. Additional analyses have to be made, to get a more specific user behaviour.

References

[1] Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggles, P.: Towards a better understanding of context and context-awareness. In: Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing, pp. 304–307 (1999)

[2] Agarwal, R., Venkatesh, V.: Assessing a Firm's Web Presence: A Heuristic Evaluation Procedure for the Measurement of Usability. Information Systems Research 13(2), 168–186 (2002)

[3] Belotti, R., Decurtins, C., Grossniklaus, M., Norrie, M.C., Palinginis, A.: Interplay of Content and Context. Journal Web Engineering 4(1), 57–78 (2005)

[4] Brown, P.J.: Triggering Information by Context. Personal Technologies 2(1), 1–9 (1998)

[5] Brusilovsky, P.: Adaptive Hypermedia. User Modeling and User-Adapted Interaction 11, 87–110 (2001)

[6] Bucklin, R.E., Lattin, J.M., Ansari, A., Gupta, S., Bell, D., Coupey, E., Little, J.D.C., Mela, C., Montgomery, A., Steckel, J.: Choice and the Internet: From Clickstream to Research Stream. Marketing Letters 13(3), 245–258 (2002)

[7] Ceri, S., Daniel, F., Matera, M., Facca, F.M.: Model-driven development of context-aware Web applications. ACM Trans. Internet Technol. 7(2) (2007)
[8] De Bra, P., Aerts, A., Berden, B., de Lange, B., Rousseau, B., Santic, T., Smits, D., Stash, N.: AHA! The adaptive hypermedia architecture. In: Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia, pp. 81–84 (2003)
[9] De Bra, P., Houben, G.-J., Wu, H.: AHAM: a Dexter-based reference model for adaptive hypermedia. In: Proceedings of the Tenth ACM Conference on Hypertext and Hypermedia, pp. 147-156 (1999)
[10] De Bra, P., Stash, N.: Multimedia adaptation using AHA! In: ED-MEDIA 2004 Conference, Lugano, Switzerland (2004)
[11] Dey, A.K., Abowd, G.: Towards a Better Understanding of Context and Context-Awareness. In: Proceedings of CHI 2000: Conference on Human Factors in Computing, The Hague, The Netherlands (2000)
[12] Feldkamp, D.: KITFramework: A Framework Architecture For Supporting Knowledge-Intensive Processes. Communications of SIWN 6, 1–7 (2009)
[13] Leverof, A., Paterno, F.: Automatic Support for Usability Evaluation. IEEE Transactions on Software-Engineering 24, 863–887 (1998)
[14] Linden, G., Smith, B., York, J.: Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing 7(1), 76–80 (2003)
[15] McCarthy, J.F.: The virtual world gets physical: Perspectives on personalization. IEEE Internet Computing 5(6), 48–53 (2001)
[16] Moe, W.M.: Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream. Journal of Consumer Psychology 13(1&2), 29–39 (2003)
[17] Moe, W.M., Fader, P.S.: Dynamic Conversion Behavior at E-commerce Sites. Management Science 50(3), 326–335 (2004)
[18] Nielsen, J.: Usability Engineering. Academic Press, San Diego (1993)
[19] Nielsen, J.: Designing Web Usability. New Riders, Indianapolis (2000)
[20] Wu, D., Im, I., Tremain, M., Instone, K., Turoff, M.: A Framework for Classifying Personalization Scheme Used on e-Commerce Websites. Systems Sciences, 12 (2003)
[21] Van den Poel, D., Buckinx, W.: Predicting online-purchasing behaviour. European Journal of Operational Research 166, 557–575 (2005)
[22] Vdovjak, R., Frasincar, F., Houben, G.-J., Barna, P.: Engineering Semantic Web Information Systems in Hera. J. Web Eng. 2(1-2), 3–26 (2003)
[23] Venkatesh, V., Agarwal, R.: Turning visitors into customers: A usability-centric perspective on purchase behavior in electronic channels. Management Science 52(3), 367–382 (2006)

Node-First Causal Network Extraction for Trend Analysis Based on Web Mining

Hideki Kawai[1], Katsumi Tanaka[2],
Kazuo Kunieda[1], and Keiji Yamada[1]

[1] NEC C & C Innovation Research Laboratories,
8916-47, Takayamacho, Ikoma city, Nara, 630-0101, Japan
[2] Graduate School of Infomatics, Kyoto University,
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan

Abstract. In this paper, we propose a node-first causal extraction method for trend analysis. Recently, it has become more important for business leaders, politicians and academics to understand broader and longer environmental trends because of the need to develop better strategies for dealing with current and future issues. Trend analysis is often utilized to identify key factors in political, economical, social and technological trends. We propose a web mining framework that can extract a causal network of key factors underlying macro trends related to a user's interest. The main idea can be described as "node-first" approach, which recursively identifies key factors relevant to a user's query, then verifies causal relations between key factors. As the result of experiment, we demonstrate high precision of key factor identification (P@100 = 0.76) and causality verifications (F-value = 0.74).

1 Introduction

In our complex modern society, it is important for decision-makers such as business leaders, policy makers and academic researchers to develop better strategies for the future. For such future planning, it is essential to identify key factors which can represent important trends or changes and to build a causal network based on the "cause-effect" relationship between them. Thus, various kinds of techniques have been developed to draw causal relations between key factors for future trend analysis [11,9,17,5,7]. However, this process requires a significant amount of expert knowledge and manual effort. Therefore, our research goal is to provide a causal network extraction system for a wide range of key factors related to user's interest.

In the text mining area, there are many studies on extracting causal relations. Chang et al. [8] built a causal network for the term "protein" from 2,000 biological domain paper abstracts. Ishii et al. [16] extracted a causal network from 60 news articles. However, these previous studies have three main drawbacks.

First, in most cases, the target corpus is given by users or experts in advance. However, selecting a target document for each trend analysis is neither feasible

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 63–72, 2011.

nor scalable, and the limitation of a corpus can prohibit finding a wide range of causal relations in different domains.

The second problem is the limitation of target sentences for causal relation extraction. The basic approach in traditional causal network construction can be classified as "link-first," where they collect sentences containing cue phrases of causal linkage. Then cause and effect nodes are extracted from these sentences. This approach can extract only a small set of cause and effect nodes, because the number of sentences containing cue phrases is relatively small. According to empirical study by Inui et al. [15], cue phrases are included in only 30% of sentences describing causal relations.

The third problem is the lack of key factor identification. Since there are so many noisy causal relations, we need to extract only important causal relations between key factors. For example, consider the following two sentences, which both have causal relations: "Traveling by car contributes to global warming." and "I was late because of rain." The former sentence is much more important than the latter for trend analysis because the former sentence describes a fact related to a environmental trend, but the latter just explains a personal experience.

To overcome these drawbacks, we used web search engines as a huge corpus. In addition, we take a "node-first" approach, where we retrieve and extract facts as a combination of noun phrases (NPs) and trend verbal phrases (TVPs), which represent changes, actions, behaviors and movements of NPs. After collecting enough facts, we identify key factors as the most frequent NPs appeared in the collected facts. Finally, we verify the causality between key factors by using retrieved sentences as evidences.

The three main contributions of this paper are the following.

1. We redefine the causal network construction problem as the node-first approach.
2. We design a recursive key factor identification method using Web search engines as a huge corpus.
3. We develop a light weight pattern matching technique for causality verifications.

The remainder of this paper is structured as follows. The next section explains the main components of our proposed system. Section 4 describes experimental settings. Section 5 presents results and discussions. Section 2 introduces some related works. Section 6 is the conclusion.

2 Related Work

Currently, there are several methods available for future planning. PEST analysis is one of the methods commonly used by decision makers in business, [1] government [12] and academics [20] to understand the future opportunities and threats posed by the forces of political, economic, social and technological change. In these future planning, there are several methods to illustrate how the critical

factors of future activity can interact each other, e.g., Futures wheel [11], Cross-impact matrix [9], Causal Loop Diagram [17], Causal Chain Analysis [5] and Pathfinder Networks [7].

In the text mining area, many studies have attempted to extract cause-effect relations from text. Khoo et al. [19] used hand-coded, domain-specific knowledge to identify explicit causation relations in English text. Inui et al. [14] used a causation marker *tame* to detect causal relations in Japanese text. More recently, other researchers used automatic discovery of lexico-syntactic patterns referring to causation. In English, triplet patterns with a causation verb such as $< NP_1$ *causes* $NP_2 >$ [10] and syntactic patterns with preposition and conjunction such as *because, since, after* and *as* [6] have been further investigated. In Japanese, Higashinaka et al. [13] derived causal expression patterns from a causally annotated EDR corpus.

There are several works about building ontologies that include causality relationship. Joshi et al. [18] are aiming to build a causation-based ontology. EcoLexicon [2] is a specialized knowledge base on the environment. SACOT framework [3] exploits several semantic relation markers including causality to build broad domain ontologies from texts. Chang et al. [8] built a causal network for the term "protein" from 2,000 biological domain paper abstracts. Ishii et al. [16] extracted causal networks from 60 news articles. Basically, these previous works relied on domain-specific corpora collected by experts in advance. On the contrary, our proposed method exploits World Wide Web as a huge corpus, and has a key factor identification mechanism to extract only important causal relations related to users' interest.

3 Node-First Causal Network Construction

An overview of our node-first causal network construction system is shown in Fig. 1. The system consists of two parts: (A) key factor crawling and (B) causality verification processes. In the key factor crawling process, the system recursively retrieves *facts* relevant to a user's query q with a web search engine, and selects the most important *facts* as *key factors*. In this paper, we assumed that user's query q and *key factors* can be represented as NPs. As the result of key factor crawling, we get a key factor network whose nodes are NPs and whose edges are co-occurrences between NPs. Next, in the causality verification process, the system verifies the causality of co-occurring edges in the key factor network. Finally, causal network G can be visualized by a graph visualization tool. In this paper, we impremented the system for Japanese texts.

More formally, we defined causal network $G = (N, E)$ as a directed graph with a finite set N of nodes and a set E of edges, which is a subset of $N \times N$. If there is a causality edge $ce = (n_c, n_e)$, then n_c is a cause node and n_e is an effect node, where n_c and n_e are elements of N.

We defined a *fact* f as a combination of NPs and TVPs.

$$f =< np, tvp >, \tag{1}$$

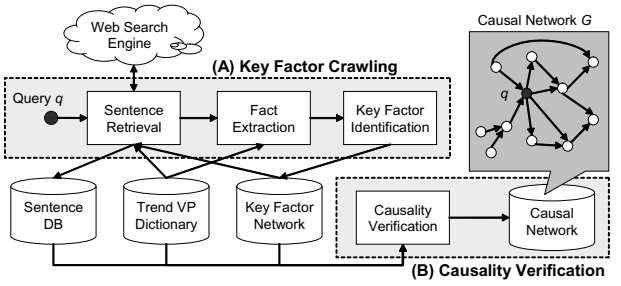


Fig. 1. Overview of node-first causal network construction

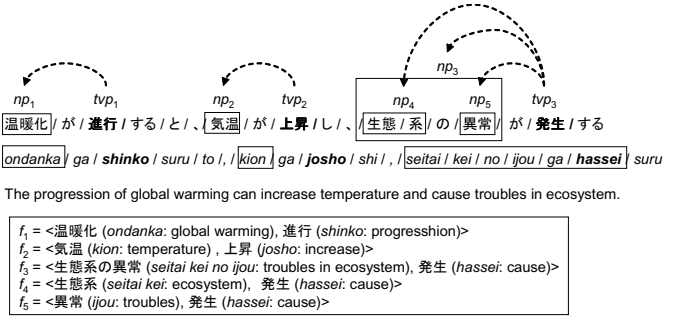


Fig. 2. Example of fact extraction

Table 1. Example of trend VPs

Political	Economical	Scientific	Technologycal	General
規制 (kisei: regulate)	急騰 (kyuuto: escalate)	研究 (kenkyuu: research)	設計 (sekkei: design)	変化 (henka: change)
制定 (seitei: establish)	急落 (kyuuraku: plunge)	発見 (hakken: discover)	開発 (kaihatsu: develop)	増加 (zouka: increase)
公布 (kouhu: publish)	提携 (teikei: tie up)	発明 (hatsumei: invent)	革新 (kakushin: innovate)	減少 (gensho: decrease)
台頭 (taito: rise)	協業 (kyougyou: cooperate)	分析 (bunseki: analyze)	進歩 (shimpo: advance)	改善 (kaizen: improve)
失脚 (shikkyaku: downfall)	独占 (dokusen: monopoly)	実験 (jikken: experiment)	製造 (seizou: manufacture)	発生 (hassei: generate)

Note that we used not only verbs but also verbal nouns as TVPs. An example of fact extraction is shown in Fig. 2. In this process, the module detects all the TVPs in a sentence s, and scan leftward to find NPs. If an NP consists of several sub-NPs joined by particles " (*no*: of)", " (*to*: and)" and " (*ya*: or)", the combination of sub-NPs and TVP is also extracted as facts. In Fig. 2, there are three TVPs and five facts are extracted. We constructed a TVP dictionary consisting of 469 TVPs by hand. Example of TVPs are shown in Table 1.

For a given user's query q, consider *context* C_q as a set of sentences related to the query q. This context can be defined as a set of surrounding sentences, a passage, the document's title or even the entire document. For each query q, all the C_q retrieved by a web search engine are stored in the Sentence DB. For a context C_q, a set of facts F_q relevant to the query q can be expressed as follows:

$$F_q = \{f | f \in s, s \in C_q\}, \tag{2}$$

where $f \in s$ denotes that np and tvp of a fact f appear in a sentence s. Now, we formally define a set of *key factors* KF as a set of NPs appearing in the fact set F_q:

$$KF = \{np_1, np_2, ..., np_m\},\ s.t.\ Freq(np_i|F_q) \geq \theta, \tag{3}$$

where θ is a threshold for *key factor*, and $Freq(np_i|F_q)$ is a frequency of np_i in a fact set F_q.

$$Freq(np_i|F_q) = |\{f | f.np = np_i, f \in s, s \in C_q\}| \tag{4}$$

where $f.np$ denotes an NP of a fact f. In the following sections, we will explain more details of (A) key factor crawling and (B) causality verification processes in Fig. 1.

3.1 Key Factor Crawling

After the key factors are identified, the most important key factor is used as the next query recursively. An overview of the recursive key factor identification is shown in Fig. 3. For a given user's query q_0 (Fig. 3(a)), five key factors $\{np_{01}, np_{02}, ..., np_{05}\}$ are identified (Fig. 3(b)). The system links between q_0 and these key factors and stores them as a key factor network. Note that in this stage, dotted undirected edges in Fig. 3(a)-(d) represent just co-occurrences, not causalities.

Next, suppose that np_{01} is used as query q_1, and three NPs $\{np_{11}, np_{12}, np_{13}\}$ are found as new key factors(Fig. 3(c)). At the same time, it is also found that np_{02} and np_{03} are relevant key factors to np_{01}. By iterating these processes recursively, we can grow the key factor network (Fig. 3(d)). The order of key factors used as queries can be determined by a score function. In this paper, we assumed that the more frequent and closer to a user's query q_0, the more important a key factor is. Thus, we defined the following scoring function:

$$Score(np_i) = \sum_{j=0}^{\tau} \frac{H(np_i, q_j)}{d(q_0, q_j) + 1}, \tag{5}$$

where τ is the number of iterations, and $d(np_i, np_j)$ is a distance between np_i and np_j on the key factor network: for example, $d(q_0, np_{02}) = 1$ and $d(q_0, np_{13}) = 2$ in Fig. 3(d). $H(np_i, q_j)$ is a relevancy function between np_i and query q_j, and we

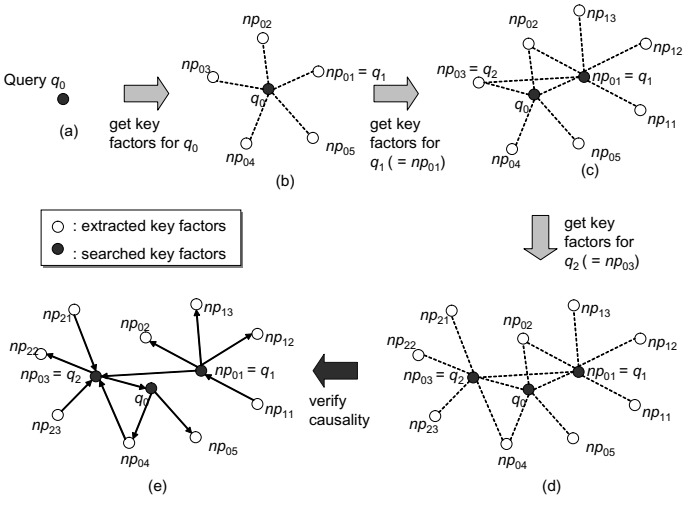


Fig. 3. Overview of recursive key factor identification

Table 2. Top 10 common causality patterns

Pattern	Roman	Freq.
NP_1によるNP_2	(NP_1 *niyoru* NP_2)	62,293
NP_1へのNP_2	(NP_1 *heno* NP_2)	22,792
NP_1に対するNP_2	(NP1 *ni taisuru* NP2)	15,939
NP_1TVPによるNP_2	(NP_1TVP *niyoru* NP_2)	12,265
NP_1TVPでNP_2	(NP_1TVP *de* NP_2)	9,713
NP_1に伴うNP_2	(NP_1 *ni tomonau* NP_2)	9,295
NP_1をTVPし、NP_2	(NP_1 *wo* TVP *shi*、NP_2)	7,964
NP_1がTVPし、NP_2	(NP_1 *ga* TVP *shi*、NP_2)	7,359
NP_1によりNP_2	(NP_1 *niyori* NP_2)	6,325
NP_1のTVPはNP_2	(NP_1 *no* TVP *ha* NP_2)	6,061

used $H(np_i, q_j) = Freq(np_i|F_{q_j})$ for the proposed method. And the np^* whose $Score(np^*)$ is the maximum is selected as the next recursive query $q_{\tau+1}$.

3.2 Causality Verification

After key factor crawling, the causality verification module verifies the causality of each co-occurrence edge on the key factor network. In this process, the system decides whether each edge $e = (np_i, np_j)$ has causality or not.

For determining the causality between np_i and np_j, we used common causality patterns extracted from 4.5 million sentences in 11 domains collected as a preliminary experiment. In the preliminary experiment, we count all the patterns between np_i and np_j in the key factor networks and manually classified most common patterns as causality or not. Finally, we got 54 common causality patterns. Top 10 common causality patterns are shown in Table 2. In the causality verification process, the system decides that a sentence s indicates a causal relation if one of causality patterns appears between np_i and np_j in the sentence.

4 Experimental Settings

The system was implemented in Perl on a single 1.86GHz Intel Xeon CPU Linux server with an 8-GB memory. We utilized Bing API [1] as a search engine. we used cytoscape for the network visualization [2]. The experimental period was from 1 to 31 March 2011. In the experiment, we used the sentence s_q, which contains a query q, and five sentences following s_q as a context C_q. Finally, we collected 32 million sentences for 76 test user queries. For each test user query, we crawled 100 NPs in the key factor crawling process, and got 4,200 sentences per NP on average. For the test user queries, we used company names, product names, country names, social issues, keywords from news headlines, technology terms and generic topic words such as finance, health and entertainment.

We evaluated the accuracy of the key factor crawling and causality verification. For the accuracy evaluation of key factor crawling, we compared our proposed system with two kinds of baselines. One is a link-first method, the other is a TF-IDF method.

In the link-first method, we joined query q_j to typical causal patterns such as "(*niyori*: by)" and extracted NPs just after the causal pattern. We used 21 causal patterns for the link-first approach, and we selected recursive query $q_{\tau+1}$ by using relevancy function $H(np_i, q_j) = Freq(np_i, q_j)$ in equation (5), where $Freq(np_i, q_j)$ is the frequency of np_i extracted by the query q_j.

In the TF-IDF method, we made a virtual document D_{q_j} joining all of the contexts C_{q_j} for a query q_j. Also we selected recursive query $q_{\tau+1}$ by using relevancy function $H(np_i, q_j) = TF(np_i, D_{q_j}) \times \frac{N_d}{logDF(np_i)}$, where $TF(np_i, D_{q_j})$ is a term frequency of np_i in D_{q_j}, $DF(np_i)$ is a document frequency of np_i, and N_d is the number of web pages indexed in the search engine.

We annotated the 500 NPs for 5 test user queries that are searched in the key factor crawling. The annotation labels were "2 (relevant to user's query q_0)", "1 (partially or indirectly relevant to user's query q_0)", and "0 (obviously irrelevant to user's query q_0)". We used precision at k (P@k) and normalized Discounted Cumulative Gain (nDCG) as performance measures widely used in the information retrieval [4]. We treated only label "2" as the correct answer.

For the causality verification, we randomly selected 2,800 sentences for arbitrary edges $e = (np_i, np_j)$ in key factor networks collected in 15 test user queries. We manually annotated causality labels, and compared with the result of system output. In this evaluation, we have used precision, recall and f-value with 5-fold cross validation. The precision PR is defined as the number of correctly labeled causalities divided by the total number of causalities detected by our system. The recall RE is defined as the number of correctly labeled causalities divided by the total number of causalities labeled by humans. F-value is defined as $2PR \times RE/(PR + RE)$. We used SVM-light [3] as an implementation of the SVM algorithm.

[1] http://www.bing.com/developers/
[2] http://www.cytoscape.org/
[3] http://svmlight.joachims.org/

Table 3. Evaluation of key factor crawling

	Node-first	Link-first	TF•IDF
P@10	**0.920**	0.740	0.760
P@20	**0.910**	0.560	0.610
P@50	**0.860**	0.480	0.556
P@100	**0.764**	0.386	0.504
nDCG	**0.978**	0.954	0.886

5 Results and Discussion

The result of key factor crawling is shown in Table 3. For all the performance measures, our proposed approach could extract the most key factors of future trends in various domains. In general, we found that the proposed approach tended to extract macroscopic factors, while a link-first approach tended to extract microscopic factors. Also, the TF-IDF method tended to extract related keywords but not ones causally related to each other. For example, in the proposed approach, key factors such as "labor costs", "China" and "globalization" were successfully extracted, however, key factors such as "Kyoto Protocol", "carbon dioxide" and "forest" were judged as "not directly related" to the query "apparel" in the evaluation process. In the link-first approach, "bipartisan" and "revolution" were successfully extracted for the query "Egypt". However, in the worst case of the link-first approach, the crawling was stopped because there were not enough NPs extracted by the causal patterns. In the TF-IDF approach, many related keywords were extracted such as "men's" and "casual" for the query "apparel business", but there were few causal relations between them. The worst case of TF-IDF happened for the query "Egypt", where all the extracted keywords were names of ancient kings, gods, and locations of pyramids. These keywords may come from travel information pages about Egypt.

As a result of causality verification, Precision, Recall and F-Value were 0.946, 0.603 and 0.736 respectively. Even though light weight pattern matching, precision of causality verification was almost 95%, Higher precision is much more important than recall for two reasons. First, there may be many sentences explaining the causality between NPs, but the user is typically interested in only a few evidences. Furthermore, for the visualization of the causal network, only the most typical relation can be drawn as an edge for a given pair of NPs. However, improving the recall should be done in future work.

Finally, a subset of a causal network extracted for the query "Egypt" is shown in Fig. 4. Note that we limited the number of edges only three in-links and three out-links for each node because it is too noisy to show all edges in the network. We can see the background of the Egyptian revolution in which the oil prices affected the cost of living, including food prices, and higher food prices triggered riots in Egypt. Also, the potential effect on Israel and crude oil can be seen in Fig. 4.

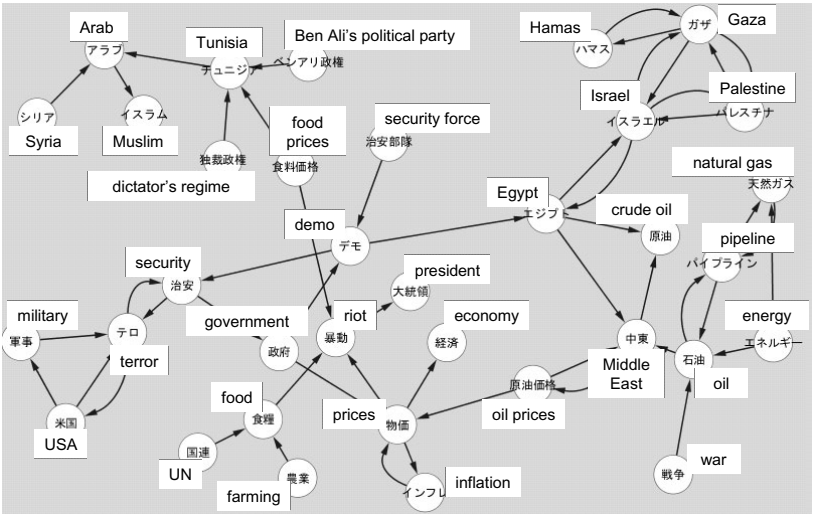


Fig. 4. Example of constructed causal network about "Egypt"

6 Conclusion

In this paper, we proposed a node-first causal network extraction method for trend analysis. We defined facts and key factors relevant to a user's query q, and crawled relevant key factors recursively. As the result of experiment, we demonstrated that high precision of key factor identification (P@100 = 0.76) and causality verifications (F-value = 0.74).

In future work, we would like to treat with a clustering method for summarizing a certain group of key factors. We also want to develop language independent methods to see different views of macro trend in different countries.

References

1. Abe, H., Suzuki, A., Etohc, M., Sibagaki, S., Koike, S.: Towards Systematic Innovation Methods: Innovation Support Technology that Integrates Business Modeling, Roadmapping and Innovation Architecture. In: Proc. PICMET 2008 Conference, Cape Town, South Africa, pp. 2141–2149 (July 2008)
2. Araúz, P.L., Faber, P.: Natural and contextual constraints for domain-specific relations. In: The Workshop Semantic Relations, Theory and Applications, Valletta, Malta, pp. 12–17 (May 2010)
3. Auger, A.: Capturing and Modeling Domain Knowledge Using Natural Language Processing Techniques. In: 10th International Command and Control Research and Technology Symposium (ICCRTS 2005) (June 2005)
4. Bao, S., Wu, X., Fei, B., Xue, G., Su, Z., Yu, Y.: Optimizing Web Search Using Social Annotations. In: Proc. the 16th International World Wide Web Conference (WWW 2007), Alberta, Canada, pp. 501–510 (May 2007)

5. Belausteguigoitia, J.C.: Causal Chain Analysis and Root Causes: The GIWA Approach. Ambio 33(1-2), 7–12 (2004)
6. Blanco, E., Castell, N., Moldovan, D.: Causal Relation Extraction. In: The 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, pp. 310–313 (May 2008)
7. Burandt, S.: Effects of an Educational Scenario Exercise on Participants Competencies of Systemic Thinking. Journal of Social Sciences 7(1), 51–62 (2011)
8. Chang, D.S., Choi, K.S.: Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. Information Processing and Management 42, 662–678 (2006)
9. Chao, K.: A New Look at the Cross-Impact Matrix and its Application in Futures Studies. Journal of Futures Studies 12(4), 45–52 (2008)
10. Girju, R., Moldovan, D.: Mining Answers for Causation Questions. In: The AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases, Palo Alto, CA, USA, pp. 15–25 (March 2002)
11. Glenn, J.C., Gordon, T.J.: Futures Research Methodology Version 3.0. Amer Council for the United Nations, Washington, D.C, USA (2009)
12. Ha, H., Coghill, K.: E-Government in Singapore – A Swot and Pest Analysis. Asia-Pacific Social Science Review 6(2), 103–130 (2006)
13. Higashinaka, R., Isozaki, H.: Automatically Acquiring Causal Expression Patterns from Relation-annotated Corpora to Improve Question Answering for why- Questions. ACM Transactions on Asian Language Information Processing 7(2) (2008)
14. Inui, T., Inui, K., Matsumoto, Y.: Acquiring Causal Knowledge from Text Using the Connective Marker tame. ACM Transactions on Asian Language Information Processing 4(4), 435–474 (2005)
15. Inui, T., Okumura, M.: Investigating the Characteristics of Causal Relations in Japanese Text. In: The Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, Stroudsburg, PA, USA, pp. 37–44 (June 2005)
16. Ishii, H., Ma, Q., Yoshikawa, M.: An incremental method for causal network construction. In: The 11th International Conference on Web-Age Information Management (WAIM 2010), Jiuzhaigou, China, pp. 495–506 (July 2010)
17. Jafari, M., Amiri, R.H., Bourouni, A.: An Interpretive Approach to Drawing Weighted and Most Frequent Causal Loop Diagram using ELECTRE III and SUBDUE Methods. International Journal of Intelligent Information Technology Application 2(3), 116–120 (2009)
18. Joshi, S., Pangaonkar, M., Seethakkagari, S., Mazlack, L.J.: Lexico-Syntactic Causal Pattern Text Mining. In: The 14th WSEAS International Conference on Computers, Corfu Island, Greece, pp. 446–452 (July 2010)
19. Khoo, C.S.G., Kornfilt, J., Oddy, R.N., Myaeng, S.H.: Automatic Extraction of Cause-Effect Information from Newspaper Text Without Knowledge-based Inferencing. Literary and Linguistic Computing 13(4), 177–186 (1998)
20. Peng, G.C.A., Nunes, M.B.: Using PEST Analysis as a Tool for Refining and Focusing Contexts for Information Systems Research. In: Proc. The 6th European Conference on Research Methodology for Business and Management Studies (ECRM 2007), Lisbon, Portugal, pp. 229–237 (July 2007)

Consumer Behavior Analysis from Buzz Marketing Sites over Time Series Concept Graphs*

Tetsuji Kuboyama[1], Takako Hashimoto[2], and Yukari Shirota[3]

[1] Computer Center, Gakushuin University
Mejiro 1-5-1, Toshima, Tokyo 171-8588, Japan
ori-kes11@tk.cc.gakushuin.ac.jp
[2] Commerce and Economics, Chiba University of Commerce
1-3-1 Konodai Ichikawa-shi Chiba, 272-8512, Japan
takako@cuc.ac.jp
[3] Faculty of Economics, Gakushuin University
Mejiro 1-5-1, Toshima, Tokyo 171-8588, Japan
20010570@gakushuin.ac.jp

Abstract. This paper proposes a text mining method for detecting drastic changes of consumer behavior over time from buzz marketing sites, and applies it to finding the effects of the flu pandemic on consumer behavior in various marketing domains. It is expected that more air purifiers are sold due to the pandemic, and it is, actually, observed. By using our method, we reveal an unexpected relationship between the flu pandemic and the reluctance of consumers to buy digital single-lens reflex camera. Our method models and visualizes the relationship between a current topic and products using a graph representation of knowledge generated from the text documents in a buzz marketing site. The change of consumer behavior is detected by quantifying the difference of the graph structures over time.

1 Introduction

Analyzing word-of-mouth in social media such as blogs and buzz marketing sites has recently become an active area of research [3,6,9]. In analyzing product reviews or reputation by word-of-mouth in social media, almost all existing research focuses first on specific products, and extracts typical evaluation expressions such as "favorite," "dislike," "expensive," and "useful." They then calculate positive/negative degrees of extracted expressions. We have also researched data mining techniques on home electrical appliances such as air purifiers and front loading washing machines with automatic drying systems; we have proposed a reputation analysis framework for buzz marketing sites [10]. It may be easy to

* This work is partially supported by Grand-in-Aid for Scientific Research 23500185 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 73–83, 2011.

analyze a specific product's reputation, because the target product's characteristics can be illustrated by the ontology for the product, which is constructed with relatively little effort. On the other hand, it is very difficult to analyze unexpected consumer behavior for "unspecified products." Because the target product is not explicit, it is not possible to prepare a specific ontology in advance. We can say that unexpected consumer behavior is implicit (hidden) information. In this paper, we would like to discover unexpected consumer behavior from word-of-mouth in social media as well as expected consumer behavior.

We conducted experiments to detect anomalous behavior of consumers attributed to the super-flu spawn in 2009, in which we discovered an unexpected consumer behavior. In threads about digital single-lens reflex cameras (digital SLR cameras), we discovered that many persons wrote about the flu. The flu pandemic made consumers hold off buying digital SLR cameras. In this case, we guessed that the flu made people cancel plans for children's PE festivals and trips during Golden Week in Japan because their children were confined at home. And those who had been planning to take photos at those events were reluctant to buy digital cameras. We could easily expect more air purifiers to be sold due to the flu pandemic. However, the reluctance in buying digital single-lens reflex cameras due to the flu was not something we had expected. The relation between the flu and digital SLR cameras can be recognized as an unforeseen and indirect relationship. This paper clarifies this sort of unforeseen and indirect relationships between a current topic and unspecified products. In analysis, we adopt the concept graph due to Hirokawa [1], which makes relevance hypernym relations of keywords appearing in a set of documents. Beyond that, we apply a graph topology-based distance measure [11] to the detection of changes of consumer behavior over time series concept graph structures.

In the following section illustrates the unexpected consumer behavior we discovered concerning the super-flu. Section 3 refers to related work. In section 4, we use concept graph due to Hirokawa to analyze the relationship between the flu pandemic and consumers holding off buying digital SLR cameras, and quantify the degree of differences between two concept graphs to detect changes of consumer behavior. Finally, Section 5 concludes this paper with remarks and future work.

2 Consumer Behavior in a Buzz Marketing Site

This section illustrates expected and unexpected consumer behavior. We focus on the super-flu pandemic in 2009. First, we conducted text mining on the bulletin board system (BBS) of kakaku.com [2], which is the most popular buzz marketing site in Japan. We used it to research the effects of the flu on consumer behavior. On the site, we can read word of mouth episodes about various products and various current affairs and events. One person begins a thread with one topic. Others then post their word of mouth views sequentially until the first person closes the thread. We call the posted individual document a post document.

2.1 Expected Consumer Behaviors

We define "expected consumer behavior" as consumer behavior that has explicit relationships with current topics. We conducted counting the word "flu" for all products on the BBS site. In Japan, the first infected patient of the super-flu was detected in May 2009 and was important news. In September to November 2009, the great epidemic was noticed and precautions were strengthened. The number of post documents with the word "flu." correlates the symptoms in autumn with news reports. We then show the number of post documents with the word "flu" for four products (Digital camera, Digital SLR camera, Lens, and Air purifier). We started by conducting text mining over all products and found four numeral effects. We discovered that the number of post documents on the "digital SLR camera" and the "air purifier" were greater than others in Figure 1. Virus elimination functions-the province of air purifiers-were quite popular, because air purifier manufacturers propagated them through the web. Therefore, the relationship between the flu and the air purifier is expected. Figure 2 illustrates the relationship between real sales of the air purifiers and the number of post documents, where we detect a strong correlation. We recognize this kind of explicit relationships as expected consumer behavior.

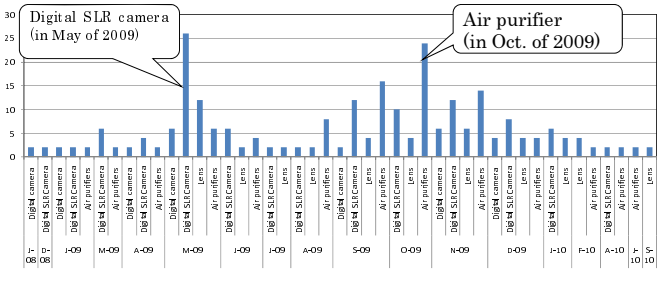


Fig. 1. The number of post documents of main products (Digital camera, Digital SLR camera, Lens, and Air purifier) that included the word "flu" from BBS of kakaku.com (from June 2008 to September 2010)

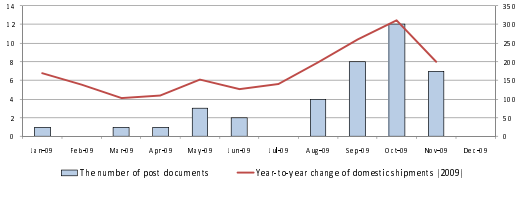


Fig. 2. The number of post documents in the BBS of kakaku.com and the volume of shipments for air purifiers. (Cited: GfK Marketing Services Japan Ltd., http://www.gfkjpn.co.jp/).

2.2 Unexpected Consumer Behaviors

We never intuitively forecast the negative relationship between digital SLR camera sales and the flu pandemic. However, the number of post documents that include the word "flu" increased most in May 2009. Reading word-of-mouth contents, we found many messages like "owing to the flu, we cannot take a vacation (or a business trip)" and "children's PE festivals may be called off owing to the flu" repeatedly. From these comments, we can guess that various events had been cancelled due to the flu pandemic, that consumers could not go out to take photos, and that, consequently, they had held off buying cameras.

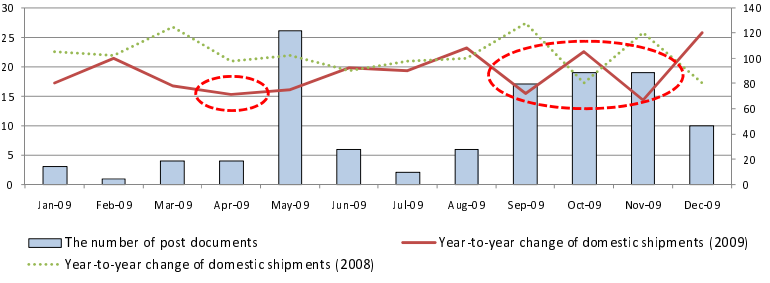


Fig. 3. The number of post documents in the BBS of kakaku.com and the volume of shipments for digital single-lens reflex camera in 2009. (Cited: The Camera Information Center: Camera information Center Report, http://www.camera-info.net).

We provide supportive evidence of our guess in Figure 3, where 2009 sales of digital cameras are shown. In 2008, an ordinary year, the slack sales of 2009 does not exist. In April to May and September to November 2009, however, we found a drastic drop in digital camera sales in our domestic market as shown in Figure 3. In September to November 2009, there is a significant reverse increase-decrease pattern compared to 2008. The drop in sales is related to the number of post documents with the word "flu;" we consequently guess there is a strong correlation between the flu and digital camera sales. The relationship between the flu and digital camera sales cannot be detected without mining technology, because it was unforeseen. We define this kind of an unforeseen and indirect relationship triggered by a current topic as unexpected consumer behavior. This would be a new approach to marketing analysis.

3 Related Work

Various researchers have analyzed product reviews and reputation from social media [3,6]. Nagano et. al [3] propose the word-of-mouth engine to present product reputation on the Web. In their system, users first specify the products by taking pictures using cell-phone cameras. Spangler et. al [6] propose an automated way to monitor social media to analyze the specific corporate brand, reputation, consumer preferences and buying habits. They also offer a mechanism

for developing the ontology, near-real-time gathering of word-of-mouth information and the calculation of positive/negative measures. This related work targets specific products, extracts evaluation expressions from word-of-mouth in social media and calculates sentiment orientations of extracted expressions to analyze product reviews and reputation. They require specific ontology. Our proposed method, however, does not target specific products, and a specific ontology is not needed. We focus on a current topic and visualize the unforeseen relations between a current topic and unspecified products from buzz marketing sites. Through the visualization, we can detect unexpected consumer behavior.

Sekiguchi et. al [7] treat recent blogger posts and analyze word co-occurrence and the rate words are repeated. They visualize the relationship between words and show topics in social media through the visualization results. Wang et. al [8] propose a graphic reputation analysis system for Japanese. It presents information on the relation between the products and users' evaluations using simple graphs (pie charts and line graphs). Both also require that specific products be targeted. Our proposed method has a similar approach, which focuses on word co-occurrence and the visualization of relations. However, we visualize the relation between a current topic and unspecified products.

4 Consumer Behavior Analysis by Concept Graphs

This section explains our analysis method in finding the unexpected consumer behavior concerning the flu pandemic in 2008.

4.1 Concept Graph

We use the concept graph proposed by Hirokawa et al. [1] to show the unforeseen relations between the current topic and unspecified products. They proposed a simple method to construct a hierarchy of words from a set of documents automatically and dynamically. The method first retrieves the set of documents according to given keywords, and extracts related words. Then hypernym relations of these related words are obtained using co-occurrence frequencies. A concept graph is a directed acyclic graph whose nodes are characteristic words of the set of documents and whose edges represent the upper and lower relation of words. It can present meaningful structures. For example, a set of documents is retrieved by the query "wine" from an English-Japanese dictionary. The concept graph of the trivial hyponym "white wine," which indicates the names of areas and brands of wine such as "cuve," "chardonnay," and "blanc," is then constructed.

Hirokawa et. al formalize upper-lower relationships among words in documents as a concept graph. The set of whole target documents is represented as D. Given a subset of D as D', and keywords w_1 and w_2, $df(w_1, D')$ represents the number of documents in D' that contain the keyword w_1,and $df(\{w_1, w_2\}, D')$

represents the number of documents that contain both w_1 and w_2 in D'. The relevance between w_1 and w_2 is defined as follows:

$$r_{D'}(w_1, w_2) = \frac{df(\{w_1, w_2\}, D')}{df(w_1, D')}. \quad (1)$$

If $r_{D'}(w_1, w_2) > \theta$ and $df(w_1, D') > df(w_2, D')$ then w_1 is defined to be greater than w_2 from the standpoint of document frequency according to Eq. (1). We set $\theta = 0.5$ in this paper. The hypernym/hyponym relation then determines an order structure among characteristic words and can be drawn as a directed acyclic graph. Visualization of a concept graph can be obtained by placing words of high frequency on the left and ones with lower frequency on the right. Thus a directed edge looks like an arrow from left to right.

Iino et. al [9] showed that the concept graph can effectively depict the time series variation of organizational structure by researchers, which is a form of implicit information in patent documents.

4.2 Concept Graph Generation from Buzz Marketing Web Sites

We decided to use the bulletin board system(BBS) of kakaku.com [2] because the site is the most popular buzz marketing site in Japan. We would like to find post documents among many products related to a specific topic of current affairs, such as the flu. We therefore think the kakaku.com BBS is suitable for us to conduct a crossover retrieve on many products.

The procedure for constructing a concept graph of a specific current topic is as follows:

1. Select the most impressive topic word of current topic w.
2. Search the kakaku.com Web site to find a set of post documents that includes the topic word w. The result is defined as a set of post documents $D = \{d_1, \ldots, d_{|D|}\}$ where d_i represents one post document that includes word w. Each d_i has four attributes (posted_time, target_id, user_id, content_id). The content_id represents content of the post document and points to the i-th content among a whole set of post document contents. The target_id shows a target product or topic of the post document such as a digital SLR camera, a camera lens, and an air purifier. The user_id specifies the person who posted it.
3. We extract keywords that are nouns, verbs, adjectives, and adverbs from d_i using morphological analysis and then calculate the value of RIDF (residual IDF) for an individual keyword.
4. For each d_i, we select a set of keywords of which RIDF values are greater than a certain threshold level T. The posted date is then delimited monthly at the beginning of the month, and D is clustered by month.
5. For each cluster, we construct concept graphs, in which each edge is labeled by the product name where two co-occurring words are found.

In this paper, we set $w =$"flu" and $T = 2.0$, and retrieve post documents from January 2009 to December 2009 in `kakaku.com`. Table 1 lists the excerpts in the

result of step 4. In May 2009, there were keywords with high RIDF value (> 2.0) such as "lens," and "photograph," on digital SLR cameras. These keywords are not typical opinions for the flu, and digital SLR cameras seem to be irrelevant to the flu, because it is difficult to extract these keywords using existing research approaches that target specific products in advance. In contrast, there are keywords like "ion," "purify," and "virus" regarding the function of air purifiers. These keywords are typical opinions against flu and we can say there is a strong relation between air purifiers and the flu.

Table 1. The excerpt in the result of keyword extraction (with high RIDF values) of step 4 (in May of 2009)

ID	Product Name	Keywords
9599867	Digital SLR Camera	flu, lens
9560041	Digital SLR Camera	flu, photograph
9586600	Digital SLR Camera	flu, lens, photograph
9587481	Digital SLR Camera	flu, photograph
9563058	Digital Camera	flu, photograph, traveling overseas
9566776	Air Purifier	flu, ion, virus, specific, product
9628120	Air Purifier	flu, virus, Daikin(manufacturer)

According to our preliminary survey, however, with regard to digital SLR cameras, because of the flu pandemic, people who had plans for children's PE festivals and trips during Golden Week in Japan were confined to their homes and those who had been planning to take photos of the events were reluctant to buy digital cameras. From this viewpoint, keywords such like "PE festival," "trip," "Golden Week," and "cancel" should have been extracted from post documents. In general, we used tf-idf values to extract keywords. However, when we use tf-idf values, values of keywords extracted from regular post documents tend to become high. On the other hand, values of keywords we would like to extract tend to become low. We have to improve measures to extract useful keywords that more precisely express unexpected user behavior. In our method, edges in concept graphs represent not only hierarchical relationship of words but also the name of the product having a correlation between both nodes.

5 Change Detection of User Behavior over Time Series Concept Graphs

A concept graph is helpful to discover unexpected consumer behavior. By analyzing time series variation of concept graphs, user behavior can be detected more precisely. However, it is required to detect structural differences among time series concept graphs. For large graphs, this task is not tractable for manual handling. In this paper, we use graph topology-based distance to detect changes of time series graphs.

5.1 Graph Topology-Based Distance Measures

A concept graph is denoted by $G = (V, E, \alpha, \beta)$, where V is the finite set of nodes, and $E \subseteq V \times V$ is the set of edges. Each node is labeled by a labeling function $\alpha : V \to L_V$, where L_V is a set of node labels, and Each edge is labeled by a labeling function $\beta : E \to L_E$, where L_E is a set of edge labels.

We employ *graph edit distance* and *maximum common subgraph distance*(MCS distance) [11] for measuring changes in concept graph topology over time. Edit distance is the cost of edit operations to transform one graph to the other. The computation of graph edit distance is, in general, known to be intractable. Fortunately, the concept graphs we consider in this paper are graphs with unique node labels. Therefore, there are efficient algorithms [11] for computing these distances with $O(n^2)$ time, where n is the maximum node size of two graphs.

The graph edit distance D_e and MCS distance D_m between two graphs $G_1 = (V_1, E_1, \alpha_1, \beta_1)$ and $G_2 = (V_2, E_2, \alpha_2, \beta_2)$ are significantly simplified and formulated as follows:

$$D_m(G_1, G_2) = 1 - \frac{|\mathrm{MCS}(G_1, G_2)|}{\max\{|G_1|, |G_2|\}}, \tag{2}$$

$$D_e(G_1, G_2) = |V_1| + |V_2| - 2|\alpha(V_1) \cap \alpha(V_2)| + |E_1| + |E_2| - 2|\beta(E_1) \cap \beta(E_2)|, \tag{3}$$

where $\mathrm{MCS}(G_1, G_2)$ denotes the maximum common subgraph (MCS) of G_1 and G_2, and $|G|$ denotes the size of G. As the size of G, we use the number of edges in G for computing Eq. (2). We also define $\alpha(V)$ as $\{\alpha(v) \in L_V \mid v \in V\}$, and $\beta(E)$ as $\{\beta(e) \in L_E \mid e \in E\}$, respectively, in Eq. (3). The graph edit distance measures the absolute change of graph structures over time. Thus, the graph edit distance is useful for detecting changes of global structures, while the MCS distance is useful for detecting relative changes of preserved substructures over time.

5.2 Quantifying Structural Transition of Concept Graphs

Figure 4 shows the plots of normalized MCS distance(left) and normalized graph edit distance(right) in time series concept graphs, which are monthly concept graphs about the flu extracted from messages of kakaku.com (from January 2009 to December 2009). Before measuring these distance measures, we split these monthly concept graphs into two sets according to the edge labels related to "camera" and "purifier." In the plot, the distance in a month(e.g. Feb) shows the graph distance between the current month(e.g. Feb) and the previous month(e.g. Jan). On the other hand, Figure 5 shows the corresponding monthly concept graphs, and some of the substructures related to the products "air purifier," "camera" are boxed by dotted lines and solid lines, respectively (although the graph structures with labels are not shown in detail due to the space limitation, the structures are roughly recognized). Note that, in the concept graph, the substructures related to the same products form clusters in the lower parts since

intuitively the upper nodes represent more general concept, while the lower are more specific.

Regarding this time series variation, we first recognized that major structure changes happened in May, October and September. The sudden appearance and disappearance in the graphs over time show drastic changes of topics discussed in the BBS. We can confirm that the graph edit distance in Figure 4 shows the change of global graph structures, while the MCS distance shows the occurrences of preserved substructures.

Our hypothesis was that these major changes caused specific consumer behavior. In Japan, the first infected patient of the super-flu was detected in May 2009 and was reported in the mass media. Further, from September to November, the great epidemic was noticed and precautions were strengthened. These major structure changes happened according to the topical problem "flu."

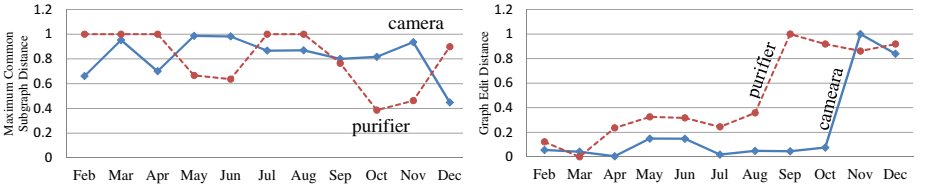


Fig. 4. Plots of MCS distance(left) and graph edit distance(right) in time series concept graphs in 2009

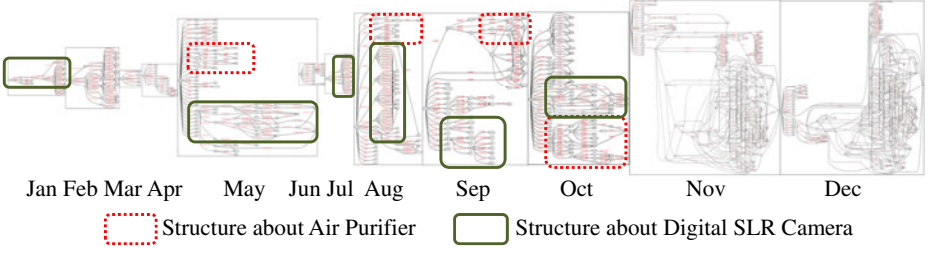


Fig. 5. Monthly concept graphs in 2009, and substructures related to "air purifier," "camera"

We confirm that the drastic structure changes happened in April and September Compared to real sales of air purifiers. We can therefore guess that the structure change in April and September illustrates consumer behavior related to air purifiers.

In Figure 5, graph structures about cameras are recognized in the concept graphs of January, May, July, August, September and October of 2009. With the plots in Figure 4, we recognize that the major structure changes happened in May and July, and a part of substructure emerged in July is preserved until October. We guess that these structure changes cause unexpected consumer behavior.

Compared to real sales of digital SLR cameras (Figure 2), sales increased in June (after May) and October (after September). We can guess, therefore, that the structure change in May and September illustrates consumer behavior. We can say that these structure changes can express changes in consumer behavior.

This type of unexpected consumer behavior detection based on graph structure changes cannot be conducted using existing reputation analysis research. It is not possible because existing research focuses on specific products first, and then extracts typical evaluation expressions such as "favorite," "dislike," "expensive," and "useful." Time series analysis based on concept graph structures is useful to visualize unexpected user behavior.

6 Conclusion and Future Work

This paper visualizes expected and unexpected consumer behavior from messages on buzz marketing sites using concept graphs. Existing data mining research cannot show this kind of consumer behavior. Moreover, we employ two graph distance measures to detect structural changes in time series concept graphs, and show that the MCS distance is useful to recognize the changes and preservation of substructures, while the graph edit distance is useful to recognize global structural changes.

In future work, the results of concept graph analysis should be integrated with appropriate marketing data for developing a system that can extract unexpected consumer behavior semi-automatically. We will obtain other data examples that can express unexpected consumer behavior from buzz marketing sites, and evaluate the effectiveness of our proposed method using the concept graph.

References

1. Shimoji, Y., Wada, T., Hirokawa, S.: Dynamic Thesaurus Construction from English-Japanese Dictionary. In: The Second International Conference on Complex, Intelligent and Software Intensive Systems, pp. 918–923 (2008)
2. kakaku.com, http://kakaku.com
3. Nagano, S., Inaba, M., Mizoguchi, Y., Iida, T., Kawamura, T.: Ontology-Based Topic Extraction Service from Weblogs. In: IEEE International Conference on Semantic Computing, pp. 468–475 (2008)
4. Kobayshi, N., Inui, K., Matusmoto, Y., Tateishi, K., Fukushima, S.: Collecting evaluative expressions by a text mining technique. IPSJ SIG NOTE 154(12), 77–84 (2003)
5. Asano, H., Hirano, T., Kobayasi, N., Matsuno, Y.: Subjective Information Indexing Technology Analyzing Word-of-mouth Content on the Web. NTT Technical Review 6(9), 1–7 (2008)
6. Spangler, W.S., Chen, Y., Proctor, L., Lelescu, A., Behal, A., He, B., Griffin, T.D., Liu, A., Wade, B., Davis, T.: COBRA - mining web for COrporate Brand and Reputation Analysis. Web Intelligence and Agent Systems (WIAS) 7(3), 243–254 (2009)

7. Sekiguchi, Y., Kawashima, H., Uchiyama, T.: Discovery of Related Topics Using Serieses of Blogsites' Entries. In: The 22nd Annual Conference of the Japanese Society for Artificial Intelligence, pp. 2I1-1 (2008)
8. Wang, G., Araki, K.: A Graphic Reputation Analysis System for Mining Japanese Weblog Based on both Unstructured and Structured Information. In: AINA Workshops 2008, pp. 1240–1245 (2008)
9. Iino, Y., Hirokawa, S.: Time Series Analysis of R&D Team Using Patent Information. In: Velásquez, J.D., Ríos, S.A., Howlett, R.J., Jain, L.C. (eds.) KES 2009. LNCS, vol. 5712, pp. 464–471. Springer, Heidelberg (2009)
10. Hashimoto, T., Shirota, Y.: Semantics Extraction from Social Computing: A Framework of Reputation Analysis on Buzz Marketing Sites. In: Kikuchi, S., Sachdeva, S., Bhalla, S. (eds.) DNIS 2010. LNCS, vol. 5999, pp. 244–255. Springer, Heidelberg (2010)
11. Bunke, H., Dickinson, P.J., Kraetzl, M., Wallis, W.D.: A Graph-Theoretic Approach to Enterprise Network Dynamics. Ch.4, pp. 63–92 (2007)

Fuzzy Image Labeling by Partially Supervised Shape Clustering

G. Castellano, A.M. Fanelli, and M.A. Torsello

Computer Science Department, University of Bari "A. Moro"
Via E. Orabona, 4 - 70126 Bari, Italy
{castellano,fanelli,torsello}@di.uniba.it

Abstract. In this paper, a fuzzy shape annotation approach for automatic image labeling is presented. A fuzzy clustering process guided by partial supervision is applied to shapes represented by Fourier descriptors in order to derive a set of shape prototypes representative of a number of semantic categories. Next, prototypes are manually annotated by attaching textual labels related to semantic categories. Based on the labeled prototypes, a new shape is automatically labeled by associating a fuzzy set that provides membership degrees of the shape to all semantic categories. Experimental results are provided in order to show the suitability of the proposed approach.

Keywords: Fuzzy shape annotation, fuzzy shape clustering, image labeling, partial supervised clustering, shape representation.

1 Introduction

The exponential growth in image databases has given rise to an increasing interest for automated tools to efficiently store, organize and retrieve images. Consequently, Content-Based Image Retrieval (CBIR) emerged as a promising technology that involves the development of systems that are able to extract the low-level visual features from the images such as texture, color and shape, in order to characterize the salient information in the image, and to exploit this information in the indexing and retrieval processes [1], [2]. Performances of CBIR systems are strongly affected by the well-known semantic gap problem that refers to the lack of coincidence between the information automatically extracted from the visual data and the semantic meaning, i.e. the interpretation that the same visual data have for a user in a given situation [3]. Of course, it was recognized that, to improve CBIR systems, semantic information has to be incorporated generally consisting in textual keywords assigned to images. The task of attaching labels to images is known as image annotation or image labeling and it has become a core research topic in CBIR [4], [5]. Labels assigned to images are related to their content, providing textual descriptions of the semantic categories to which images belong.

Hence, automatic image labeling can be seen as a problem of image classification where two main issues have to be addressed, i.e. the identification of the

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 84–93, 2011.

semantic categories and the association of a new image to the identified categories. The most of image labeling approaches assume the existence of a number of semantic categories strictly dependent on the considered image domain. However, the number of categories is not always known in advance. In many cases, experts may have some knowledge about the image domain as, for example, the membership of a number of images to semantic categories. Thus, a method able to exploit such kind of knowledge for the automatic discovery of semantic categories and the corresponding visual prototypes could be useful to improve the overall process of image labeling. In addition, the process of attaching labels to images is usually crisp, i.e. an image is exactly classified into one semantic category [6], [7]. However, due to the presence of noise and ambiguity in image features, it is very difficult or even impossible to classify them into precisely one category. Therefore, the uncertainty characterizing the labeling process can be properly captured by annotating images with multiple labels leading to a fuzzy labeling process that assigns multiple keywords to an image together with values representing the membership degree of the same image to each semantic category [9], [10], [11].

To address these problems, we propose a fuzzy shape annotation approach for image labeling. We automatically define semantic categories by means of a fuzzy clustering process and assign to shapes multiple labels together with membership values of shapes to categories. Specifically, the approach first creates a database of unlabeled object shapes represented by Fourier descriptors. Then, objects are grouped together on the basis of shape similarity by applying a process of fuzzy clustering equipped with a partial supervision mechanism that allows to take advantage from the knowledge of few labeled shapes providing an helpful guidance during the activity of cluster discovery. For each derived cluster, a prototypical shape is determined that is manually associated to a textual label corresponding to a particular semantic category. Finally, to label a new shape, its visual descriptors are matched with visual descriptors of all prototypes and the similarity values are used to create a fuzzy set expressing the membership degrees of the shape to each identified semantic category.

The rest of the paper is organized as follows. The next section describes the proposed approach. Section 3 provides some experimental results obtained by testing the proposed approach on a shape data set. Section 4 closes the paper by drawing some conclusions and future directions.

2 The Proposed Approach

We suppose that a collection of shapes expressed in the form of boundary coordinates is available. All the shape boundaries are Fourier descriptors that, among the different approaches proposed for describing shape information, are well-recognized nowadays to provide robustness and invariance, obtaining good effectiveness in shape-based indexing and retrieval [12]. Fourier descriptors represent the outside contour of a shape by means of a limited number of coefficients in the frequency domain. Since such coefficients also carry information about the

size, the orientation, and the position of the shape, they have to be properly normalized in order to achieve invariance properties with respect to transformations. In addition, to obtain a compact description of a shape, a common approach is to retain only a subset of Fourier coefficients, corresponding to those with frequency closer to zero [13], [14].

The choice of an appropriate number M of coefficients to be used has to trade off the accuracy in representing the original boundary with the compactness and simplicity of the representation. We indicate by $\mathbf{s} = (s_1, s_2, ..., s_M)$ the representation of an object shape by means of its M Fourier descriptors.

Once all the available shapes have been represented by Fourier descriptors, the effective process of shape annotation can start. In our proposed approach, two main steps can be distinguished: (i) shape clustering for the derivation of shape prototypes representative of a number of semantic categories and (ii) shape labeling for the fuzzy annotation of shapes. In the following, these steps are detailed.

2.1 Shape Clustering

In the step of shape clustering, a set of shape prototypes is automatically defined by applying a clustering process so as to group similar shapes into a number of clusters and represent each cluster by means of a prototypical shape.

In this work, to group similar shapes into overlapping clusters representing several semantic categories, we employ a fuzzy clustering algorithm augmented by a partially supervised mechanism described in [15] that represents a modified version of the well-known Fuzzy C-Means (FCM) algorithm. In [8], FCM was applied for shape clustering. However, poor results were obtained when grouping very similar shapes belonging to different semantic classes. In this case, it can be helpful having some guidance in the form of a limited number of shapes labeled by the user who imposes his point of view about the semantic of the selected shapes. This is the underlying rationale behind the use of FCM with partial supervision adopted in this work. In the following, we recall the essence of the clustering algorithm. However, the interested reader can find more details about the clustering procedure in [15].

Let $S = \{\mathbf{s}_j\}_{j=1}^{N}$ be a set of N shapes represented by Fourier descriptors and K a given number of clusters, we denote by $S_1 \subset S$ a small set of labeled shapes and by $\mathbf{b} = [b_j]_{j=1}^{N}$ a boolean vector defined as follows:

$$b_j = \begin{cases} 1 & \text{if shape } s_j \text{ is labeled} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Analogously, membership values of the labeled shapes to the clusters are arranged into a matrix $\mathbf{F} = [f_{jk}]_{j=1...N}^{k=1...K}$.

The partially supervised clustering algorithm works in the same manner of FCM. This is an process that iteratively mines K clusters by minimizing an object function consisting in a modified version of the object function of the original

FCM obtained by adding the supervised learning component encapsulated in the form of **b** and **F** as follows:

$$O = \sum_{k=1}^{K} \sum_{j=1}^{N} u_{jk}^{m} d_{jk}^{2} + \alpha \sum_{k=1}^{K} \sum_{j=1}^{N} (u_{jk} - b_j f_{jk})^{m} d_{jk}^{2} \tag{2}$$

where d_{jk} represents the Euclidean distance between the shape s_j and the center of the k-th cluster, m (the fuzzification coefficient) is any real number greater than 1 and α is a parameter that serves as a weight to balance the supervised and unsupervised components of the clustering process. In general, the higher the value of α, the higher is the impact coming from the supervised component. It has been observed that a value of α proportional to the rate between the total number of shapes and the number of labeled shapes ensures that the impact of the labeled shapes is not ignored in the clustering process. The second term of the objective function captures the difference among the true membership of shapes (encapsulated in **F**) and the membership computed by the algorithm. The aim to be reached is that, for the labeled shapes, these values have to coincide.

The object function O is minimized by updating the partition matrix $\mathbf{U} = [u_{jk}]_{j=1\ldots N}^{k=1\ldots K}$ in the following way:

$$u_{jk} = \frac{1}{1+\alpha} \left[\frac{1 + \alpha(1 - b_k \sum_{l=1}^{K} f_{lk})}{\sum_{l=1}^{K} d_{jk}^{2} / d_{lk}^{2}} \right] + \alpha b_k f_{jk} \tag{3}$$

The iterative process ends when the difference between the values of the objective function obtained in two consecutive iterations does not exceed a prefixed threshold or when the established maximum iteration number is reached.

When the clustering process is completed, as a result, the algorithm provides a fuzzy partition matrix **U** containing the membership degrees of each shape to each discovered cluster. These values are exploited to derive a prototypical shape for each cluster. Namely, for each cluster, the shape with maximal membership degree is selected as prototype. We denote by $\mathbf{p}_k$ the prototypical shape of cluster k.

Once shape prototypes have been derived, these are manually annotated by a domain expert according to a set of C semantic categories. Precisely, each derived shape prototype is associated to a unique textual label corresponding to a semantic class represented by the prototype. Of course, different prototypical shapes may convey the same semantic content (i.e., several different shapes may convey the same class of objects), i.e. $K \geq C$. We consider such prototypes to belong to the same semantic class: thus, such prototypes will have attached the same class label. As a result, we may have different shape prototypes with attached the same textual description.

2.2 Fuzzy Shape Labeling

Fuzzy shape labeling is the step devoted to the effective annotation of the shapes. Precisely, every time a new shape is added to the database, its Fourier descriptors

$\mathbf{s}_{j*}$ are matched against Fourier descriptors of all prototypes $\mathbf{p}_k$ by computing the Euclidean distance $\|\mathbf{s}_{j*} - \mathbf{p}_k\|$. Then, membership degrees of the shape to clusters are calculated according to a Gaussian membership function as follows:

$$\mu_{j*k} = e^{\frac{-\|\mathbf{s}_{j*} - \mathbf{p}_k\|}{2\sigma}} \tag{4}$$

where $\sigma = \frac{\|\mathbf{p}_k - \mathbf{p}_h\|}{r}$ is calculated by using the *first-nearest-neighbor* heuristic where $\mathbf{p}_h$ is the cluster prototype nearest to $\mathbf{p}_k$ and r is an overlap parameter ranging in $[1.0, 2.0]$.

Thus, each object shape $\mathbf{s}_{j*}$ is associated with a fuzzy set of labels L_{j*} defined as follows:

$$L_{j*} = \{\mu_{j*1}, \mu_{j*2}, \ldots, \mu_{j*C}\} \tag{5}$$

where μ_{j*i} represents the membership degree of the $j*$-th shape to the i-th semantic category that are obtained by computing the maximum value of membership degrees of the shape with respect to all prototypes $\mathbf{p}_k$ representative of the i-th semantic category, namely:

$$\mu_{j*i} = \max_{\mathbf{p}_k \in C_i} \{\mu_{j*k}\} \tag{6}$$

As a result, an object shape usually belongs to multiple semantic categories with different degrees of membership.

Whenever a new shape is labeled, shape prototypes have to be updated in order to take into account the information coming from the association of the new shape to a given semantic category. To accomplish this, we firstly select the semantic category C_i with the maximal membership degree of the new labeled shape. Hence, the visual prototype corresponding to such category is updated. Since, in our approach, a category may have several shape prototypes we choose to update, among the different prototypes, the visual prototype having the minimum distance with the labeled shape, namely

$$\mathbf{p}^* = \arg \min_{\mathbf{p}_k \in C_i} d(\mathbf{s}_{j*}, \mathbf{p}_k) \tag{7}$$

Finally, we compute the average distances among the determined prototype $\mathbf{p}^*$ and all shapes of its cluster and these last with the labeled shape. The new prototype will be represented by the shape that has the minimum average distance value.

When the matching among a new shape and visual prototypes of all semantic classes provides membership values lower than a given threshold, a new category is added. Precisely, Fourier descriptors of the new shape are considered as the visual prototype of the new semantic category (i.e. $\mathbf{p}_{K+1} = \mathbf{s}_{j*}$). Then, the newly created prototype is manually labeled by the domain expert with a keyword that describes the new added semantic category.

The use of shape prototypes, which represent an intermediate level of visual signatures, facilitates the annotation process, since only a reduced number of shapes (the prototypical ones) need to be manually annotated. Secondly, the use

Table 1. Sample images of the employed data set

 Shark	 U-Eel	 Tonguefish	 Crustacean	 Eel
 Sole	 Ray	 Seamoth	 Seahorse	 Pipefish

of prototypes simplifies the search process in a retrieval system. Indeed, since any single user query is likely to match with high degree only a small number of objects, a large number of unnecessary comparisons is avoided during search by performing matching with shape prototypes rather than with specific shapes. In other words, prototypes act as a filter that reduces the search space quickly while discriminating the objects.

3 Experimental Results

To test its effectiveness, the proposed approach was tested on a dataset consisting of 1,100 text files that contain the coordinates of boundary points of objects representing a marine animal. In our experiments, we considered a portion of the data set composed of 265 images that it has been possible to manually classify into 10 different semantic categories, as follows: "Seamoths" (11), "Sharks" (58), "Soles" (52), "Tonguefishes" (19), "Crustaceans" (11), "Eels" (26), "U-Eels" (20), "Pipefishes" (16), "Seahorses" (11) and "Rays" (41). In table 1 some sample images of the employed data set portion are shown, along with their respective semantic categories.

For each shape of the employed data set, Fourier descriptors were computed by using a number of 32 coefficients. Such number was empirically established by relying on preliminary experiments where we found that 32 coefficients allow to achieve a good trade-off between compactness and accuracy of shape representation.

The collection of the obtained 265 shape descriptors was divided into a training set (90%) and a test set (10%). In the next step, the partially supervised FCM algorithm was applied to the training set. We performed several runs of the algorithm by varying the cluster number from 10 to 20 and the percentage of labeled shapes from 10% to 30%. In all runs, we fixed the fuzzification coefficient $m = 2$, while the parameter α was modified so as to make it proportional to the percentage of labeled shapes. The experiments were repeated 10 times for each scenario in order to achieve more stable results. At the end of each trial, we calculated the Xie-Beni index [12] that is one of the most employed indexes to determine the cluster number corresponding to a good partition in terms of compactness and separation of the identified shape groups. Figure 1 shows that, on the average, the best partitions are found in correspondence of $K = 15$ (where

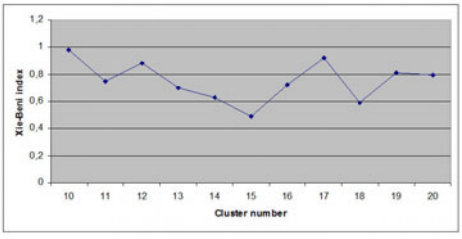


Fig. 1. The obtained Xie-Beni index values

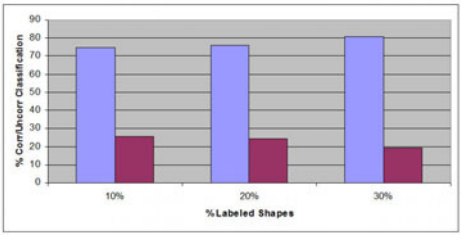


Fig. 2. Percentage of correctly and uncorrectly classified shapes

the validity index has the minimum value). Moreover, we observed (see Fig. 2) that by increasing the number of labeled shapes, the goodness of the obtained partitions improves in terms of percentage of shapes correctly classified.

Hence, to determine shape prototypes, we considered as a final result of shape clustering the partition obtained by running the algorithm with $K = 15$ and the 30% of labeled shapes. Successively, for each discovered cluster, a shape prototype was determined by selecting the shape having the highest membership degree to that cluster. Therefore, 15 different shape prototypes (one for each derived cluster) were determined together with the matrix containing the membership degrees of each shape to each derived cluster. Then, shape prototypes were manually annotated by associating to each of them a label related to the corresponding semantic category. Table 2 reports some information about the obtained clusters. Each row indicates the cardinality (expressed in terms of percentage) of the dominant category (DCC) belonging to that cluster, the associated label (corresponding to the most represented semantic category) and the respective shape prototype. It can be observed that the clustering algorithm for some semantic classes, such as "Soles" and "Sharks", has derived more than one prototype. In effect, such classes include very dissimilar shapes that, therefore, the algorithm has categorized in different clusters. This shows that different shape prototypes can be representative of the same semantic class so that different shapes convey the same semantic concept. Successively, the effective phase of shape annotation was performed by exploiting the derived shape prototypes. Each shape included in the testing set was matched with the derived shape prototypes and, then, it was annotated by assigning a fuzzy set derived by computing

Table 2. The resulting clusters with the respective prototypes and category labels

Cluster	DCC	Label	Prototype
1	77,70%	Sharks	
2	100%	U-eels	
3	87,50%	Rays	
4	32%	Pipefishes	
5	83,33%	Eels	
6	83,33%	Seamoths	
7	47,36%	Tonguefishes	
8	86,90%	Soles	
9	100%	Crustaceans	
10	90%	Seahorses	
11	80%	Soles	
12	50%	Sharks	
13	91,30%	Sharks	
14	81,25%	Rays	
15	80%	Soles	

the membership degrees according to the eq. 4. In our experiments, the value for the overlap parameter r was fixed to 1.5.

To better to assess the suitability of FCM with partial supervision as a method for creating prototypes, we made a comparison with the well-known FCM algorithm with no supevision. The overall annotation process was evaluated using Precision and Recall measures on the testing set. Comparative results are shown in Table 3. As it can be observed, a better annotation accuracy can be achieved by using shape prototypes derived by FCM with partial supervision. It can be seen that the use of standard FCM leads to null values of recall and precision in correspondence of some semantic classes, namely "Pipefishes", "Seahorses", "Seamoths" and "Crustaceans". This is because FCM recognizes shapes belonging to these classes as very similar to shapes of other classes, thus failing in determining a prototype for each of these semantic classes. Obviously, this strongly affects the annotation result for images including shapes of the unidentified categories.

Finally, it should be noted that the proposed annotation approach is quite accurate compared to manual annotation. Moreover, it turns out to be more efficient than a fully manual annotation since only a small number of shapes (i.e. the number of the identified shape prototypes) have to be manually annotated by the domain expert.

Table 3. Precision and Recall values for the annotation process

Class	FCM with partial supervision		FCM with no supervision	
	Precision	Recall	Precision	Recall
U-Eels	0.67	1.00	0.50	1.00
Sharks	0.50	0.70	0.60	0.60
Tonguefishes	1.00	0.50	0.20	0.50
Rays	1.00	0.75	1.00	0.50
Soles	0.80	0.80	1.00	0.60
Eels	1.00	1.00	1.00	0.40
Pipefishes	1.00	1.00	0.00	0.00
Crustaceans	1.00	1.00	0.00	0.00
Seamoths	0.50	1.00	0.00	0.00
Seahorses	0.50	1.00	0.00	0.00

4 Conclusions

In this paper, we presented an approach to fuzzy annotation of object shapes for automatic image labeling. A clustering process equipped with a partial supervision mechanism is applied to discover a number of clusters grouping similar shapes of several semantic classes. Shapes are annotated by assigning them a fuzzy set including membership degrees of the considered shape to each of the discovered semantic classes. Experimental results show the suitability of the proposed approach and they encourage its application to wider contexts. Future work in the direction outlined in this paper will be addressed to the definition of further mechanisms for refining the discovery process of shape prototypes. Moreover, research will investigate appropriate methods for the dynamical update of the derived prototypes that, for example, could take advantage from mechanisms of relevance feedback expressed by the users when they interact with the system.

References

1. Lew, M., Sebe, N., Djeraba, C., Ramesh, J.: Content-based Multimedia Information Retrieval: State of the Art and Challenges. ACM Transactions on Multimedia Computing, Communications, and Applications, 1–19 (2006)
2. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys 40(2), 1–60 (2008)
3. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transaction on Pattern Analysis and Machine Intelligence 22, 1349–1380 (2000)
4. Inoue, M.: On the need for annotation-based image retrieval. In: Proc. of the Workshop on Information Retrieval in Context, pp. 44–46 (2004)
5. Akbas, E., Vural, F.Y.: Automatic Image Annotation by Ensemble of Visual Descriptors. In: Proc. of Conf. on Computer Vision (CVPR) 2007, Workshop on Semantic Learning Applications in Multimedia, pp. 1–8 (2007)
6. Platt, J., Cristianini, N., Shawe-Taylor, J.: Large margin DAGs for multiclass classification. Advances in Neural Information Processing Systems 12, 547–553 (2000)

7. Wu, G., Chang, E., Li, C.: SVM binary classifier ensembles for image classification. In: Proc. of ACM Conf. on Information and Knowledge Management, pp. 395–402 (2001)
8. Castellano, G., Fanelli, A.M., Torsello, M.A.: A fuzzy set approach for shape-based image annotation. In: Proc. of International Workshop on Fuzzy Logic and Applications (WILF 2011), Trani, Italy (in Press 2011)
9. Chen, Y., Wang, J.Z.: A region-based fuzzy feature matching approach to content-based image retrieval. IEEE Transaction on Pattern Analysis and Machine Intelligence 40(9), 1252–1267 (2002)
10. Chander, V., Tapaswi, S.: Shape Based Automatic Annotation and Fuzzy Indexing of Video Sequences. In: Proc. of the 2010 IEEE/ACIS 9th International Conference on Computer and Information Science (ICIS 2010), Washington, DC, USA, pp. 222–227 (2010)
11. Athanasiadis, T., Mylonas, P., Avrithis, Y., Kollias, S.: Semantic Image segmentation and object labeling. IEEE Transaction on Circuits and Systems for Video Technology 17(3), 298–312 (2007)
12. Bartolini, I., Ciaccia, P., Patella, M.: WARP: Accurate retrieval of shapes using phase of Fourier descriptors and Time warping distance. IEEE Trans. on Pattern Analysis and machine Intelligence 27(1), 142–147 (2005)
13. Rafiei, D., Mendelzon, A.O.: Efficient Retrieval of Similar Shapes. The Very Large Data Bases Journal 11(1), 17–27 (2002)
14. Zhang, D., Lu, G.: A Comparative Study of Fourier Descriptors for Shape Representation and Retrieval. In: Proc. Fifth Asian Conf. Computer Vision, pp. 646–651 (2002)
15. Pedrycz, W., Waletzky, J.: Fuzzy clustering with partial supervision. IEEE Transaction System Man Cybernetics 27(5), 787–795 (1997)

Intelligent E-Learning System for Training Power Systems Operators

Liliana Argotte, Yasmin Hernandez, and G. Arroyo-Figueroa

Instituto de Investigaciones Electricas
Cuernavaca, Morelos, 62490, Mexico
{largotte,garroyo}@iie.org.mx

Abstract. Training of operators has become an important problem to be faced by power systems: updating knowledge and skills. An operator must comprehend the physical operation of the process and must be skilled in handling a number of normal and abnormal operating problems and emergencies. We are developing an intelligent environment for training of power system operators. This paper presents the architecture of the intelligent environment composed by reusable learning objects, concept structure maps, operator cognitive and affective model, tutor and adaptive sequence, and learning interface. The operator model and adaptive sequence are represented by probabilistic networks that select the best pedagogical and affective action for each specific operator. The model was evaluated using scholar environments with good results. The general aim of our work is to provide operators of complex industrial environments with a suitable training from a pedagogical and affective viewpoint to certify operators in knowledge.

Keywords: adaptive learning, intelligent environment, learning objects, SCORM, sequencing model.

1 Introduction

Learning and training are becoming strategic task for the electricity utilities. CFE, the mexican national utility, like other electricity utilities in the world is having a generational change, many employees are in the process of retirement. In this context, training of operators is an important problem faced by power systems. A power system can be described by a great variety of processes with multiples state variables, events and disturbances. The processes are generally nonlinear and the operating conditions may vary over a wide range subjected to various disturbances and noise. The process of learning how to control, maintain and diagnose power systems take years of practice and training. An operator must comprehend the physical operation of the process and must be skilled in handling a number of abnormal operating problems and emergencies. The problem increases when the complexity of the system obstructs the efficiency, reliability and safe operation.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 94–103, 2011.

The training requirements for utilities ask for advanced training systems based on new information technologies and artificial intelligent algorithms. Some tools used for advance training systems are: 3D simulation systems, adaptive interfaces, multimedia interfaces, virtual reality systems, learning objects reusable, intelligent systems, virtual laboratories and so on.

In the field of intelligent systems, most of the learning proposals in education are based on Intelligent Tutoring Systems (ITS). Intelligent Tutoring Systems are interactive learning environments that have the ability to adapt to a specific student during the teaching process. In general, the adaptation process can be described in three phases: (i) getting the information about the student, (ii) processing the information to initialize and update a student model, and (iii) using the student model to provide the adaptation.

This paper describes the development of an intelligent e-learning system (IES) for training of power systems operators. The IES takes elements of advanced learning systems such as reusable learning objects (RLO) based on 3D simulation systems and virtual reality systems, intelligent sequence, adaptive operator models (cognitive and affective). In contrast with a traditional training system, the main goal of the intelligent environment is to certify operators in knowledge, skills, expertise, abilities and attitudes for operation of power systems.

2 Design of the Intelligent E-Learning System

The architecture of the intelligent environment is based on the generation of dynamic courses generating systems proposed by Brusilovsky [1]. The intelligent environment is composed of four main modules (see Figure 1): domain knowledge, tutor, operator model, and the learning management system (LMS).

The first module contains the knowledge of the domain as learning objects and concept structure maps. The tutor module is the component that generates the intelligent sequence of learning objects to be presented to the operator as a course. Taking as a basis the concept map, the pedagogical component and the operator model, the sequencing model generates a course for each operator. The operator model is used to adapt the IES to each operator. The operator model integrate cognitive, operator and affective components. Finally, the LMS controls the interactions with the operator, including the dialogue and the screen layout.

3 Domain Knowledge

The domain knowledge contains the expert's operation abilities in procedure and malfunction operations and its representation considers both theoretical and practical concepts. The knowledge of the domain is represented by concept structure maps and reusable learning objects (RLO). The RLO as well named Shareable Content Objects (SCO) complies with the SCORM (Shareable Content Object Reference Model) standard [2].

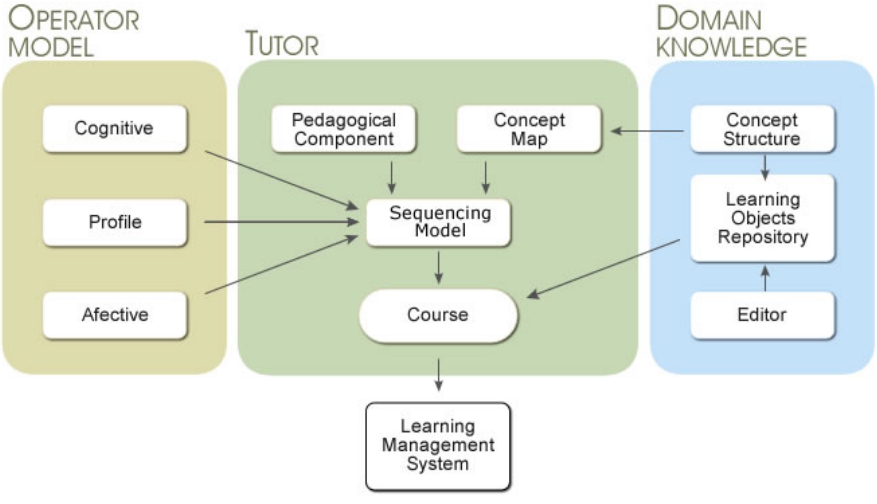


Fig. 1. Architecture of the IES

The concept structure contains the concept/topic structure of the subject knowledge to be taught (see figure 2). It is possible to organize the domain concepts/topics into a set of smaller, possibly interrelated AND/OR graphs, representing relatively independent sub-areas of the knowledge, different views, or different levels of granularity. It is represented as an AND/OR graph, where nodes represent the concepts domain or elements of knowledge, such as electrical topics, components of control board, rules, procedures and so on; and arcs represent relationships between concepts, such as a prerequisite for learning a concept or a sequence. Every node is associated with a set of teaching and testing materials labeled as (RLO), which instantiate different ways to teach the concept/topic (e.g. introduce, explain, give an example, and give a simulation, exercise, or test).

For the training of power system operators, the concept structure map is made based on the structural decomposition of the main processes of power systems. The concept structure has a higher expressive power because it allows representing not only prerequisites, but also many different types of relationships between concepts; and it enables the use of AI planning techniques for the generation of alternative courses. Therefore, it guarantees a wide variety of different teaching goals and several courses for achieving these goals.

RLO are self-contained learning components that are stored and accessed independently. The learning materials consider both theoretical and practical concepts contained in: electronic books, simulation, multimedia, virtual reality, and others digital applications to present to the operator pedagogical actions such as explanations, exercises, tests, and so on. RLO are any digital resource that can be reused to support Web-based learning using learning management systems (LMS). The learning content authors can create, store, reuse, manage and deliver digital learning content. The editor contains tools for edition of teaching and testing materials based on learning objects. The RLOs are stored in a learning object repository (LOR) [3].

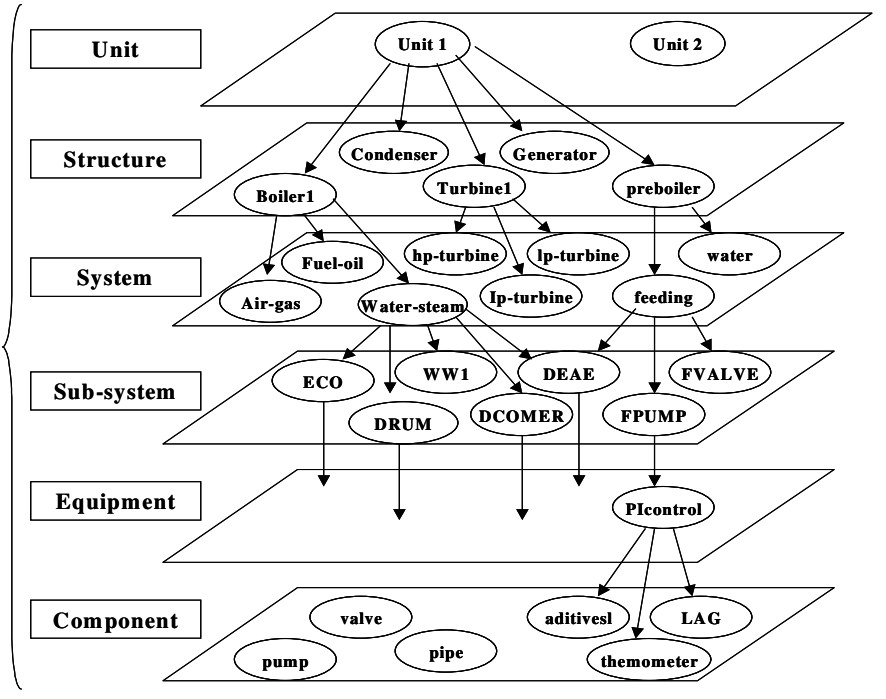


Fig. 2. Concept structure map for power plant process

By the end of the year 2009, CFE (National company for power generation and distribution in Mexico) has a collection of more than 400 instructional courses and 130 power plant courses wrapped as SCORM complaint learning objects in the RLO. Courses have also been developed for maintenance of energized lines and for the maintenance and operation of electrical substation [4]. The LOR manages, spreads and promotes the knowledge of power system, by mean of the search and recovery of RLOs.

4 Tutor

The heart of the IES is the tutor module. The module is composed by four components: concept map, pedagogical component, sequencing model and course. The concept map is an abstraction of the concept structure map.

The pedagogical component contains the learning goals and several pedagogical strategies to teach operators and to select the pedagogical strategy to be used, based on the Operator Model. This strategy represents the specific necessities of the operator from both, affective and pedagogical point of view. The pedagogical actions (explanations, exercises, tests, and so on); exist as learning object (LO) in the repository of the Domain Knowledge Module.

The main component of the tutor module is the sequencing model that contains the sequence of learning objects to be presented to an operator as a self instructional non-traditional course. The adaptive sequence is represented as a decision network that selects the best pedagogical action for each specific operator. The goal is to present to the operator the learning materials that better fits his learning needs.

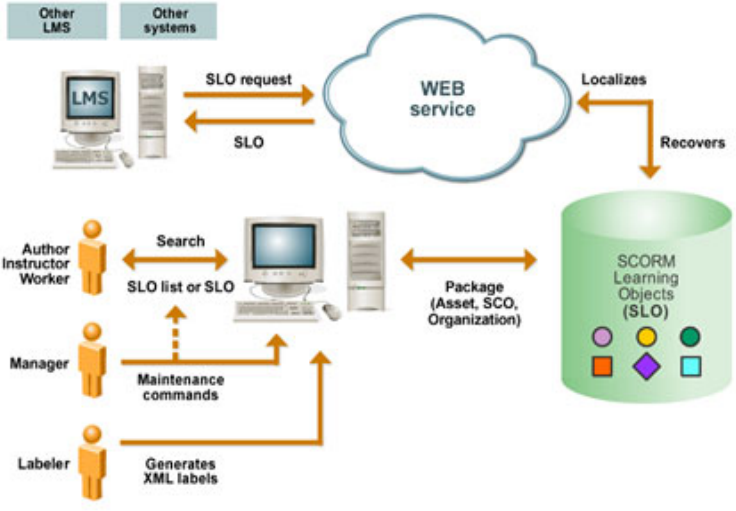


Fig. 3. Learning object repository

A decision network or influence diagram represents information about the current state of the tutor, their possible actions, the state resulting from the action of the tutor and the usefulness of the resulting state. It can be seen as an extension to Bayesian Networks with random nodes, incorporating decision nodes and utility nodes. The sequencing model consists of the following random variables (see Fig. 4): knowledge of LO, satisfaction of the objective, progress of the objective, progress of the activity, score of LO, quiz, project, task, practice [5].

As shown in the figure above relations between the nodes are causal. KwLO random node has four values: Very Good, Good, Enough and Not Enough. The decision node considers the pedagogical actions that will be evaluated according to the utility function to select the best one, and in this work four possible pedagogical actions are identified: LO challenge, Next LO, Repeat refined and Repetition of LO. The calibration of the decision network is given by the experts in the domain. The utility table entries are set based on the teacher's experience about the best over all possible pedagogical actions according to the given knowledge state and the student's interaction.

An empirical evaluation of the sequencing model was applied to undergraduate courses: Mathematics II, Electricity and Magnetism, and Introduction to Physics. The instructors designed four LOs of a specific topic for each course, which were built under the standard SCORM. Each course was divided by two groups: focus and control. The focus group consisted of students who used the system during a specific time

period, while the control group had no access to the system. The total population was of 58 students (N = 58).

The total population of each group was divided by the instructors at random to form two groups: Focus and Control. The focus group consisted of students who used the E-Learning Intelligent System during a specific time period, while the control group had no access to the system and used the remedies provided by the instructor from the Blackboard platform (with the same amount of learning resources in the E-Learning IS). The three focus groups in each course were heterogeneous in the level of subject knowledge, skills or even interest, since they were formed by students of various disciplines or careers. In Figure 5 one can compare the learning gains for each group obtained from the average of the group obtained in the Pre-Test, focus groups being those with higher learning gain.

E-Learning Intelligent System is being validated in CFE (Comisión Federal de Electricidad – the National Electric Utility in Mexico). A course is generated in a dynamic way based on particular characteristics of the operator, and it can be transformed dynamically to reflect changes in the Operator Model. Each course contains the sequence of learning objects which will be presented to the operators as self instructional and nontraditional courses. Each course supports different learning paths for each operator and provides an intelligent sequence of the instructional material towards to the training goal, starting from the current state of student knowledge as recorded in the Operator Model.

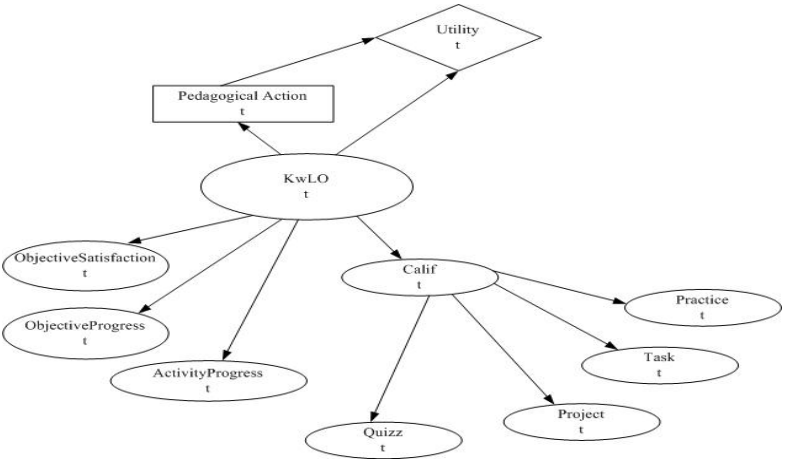


Fig. 4. Decision Network for sequencing model

5 Operator Model

The operator model is used to adapt the intelligent system to each specific operator. The operator model is built from observations that the IE makes about the operator. These can come in the form of responses to questions, answers to problems, behavior, etc. The operator model is divided into three subcomponents: pedagogical component,

operator profile and affective component. The pedagogical model represents the state of knowledge of the operator.

The pedagogical model is constantly updated during the training session as the pedagogical state of operator changes. The pedagogical model is an overlay model, where the operator's knowledge is a subset of the knowledge represented by the concept structure. The pedagogical model is active during the training session and collects all the information related to trainee performance in the current session: i.e. instructional plan, errors, objectives requested, actions performed, etc.

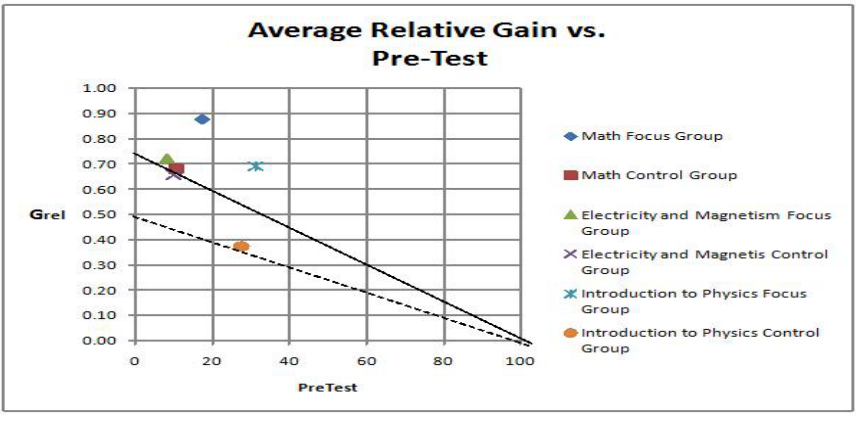


Fig. 5. Average Relative Gain vs Pre-Test

The affective model of the operator is a representation of the affective state of the trainee. The operator model must contain knowledge about the affective state of the student, in addition to knowledge about her pedagogical state, in order to give her an affectively adequate response and at the pedagogically appropriate time. The Figure 6 shows the pedagogical and affective model [6].

The operator profile represents the relevant operator's features which are important for the training process: capacity of pedagogical development, learning style, expertise, and so on. The profile component contains information related to: operator curriculum, expertise, skills, learning characteristics, operator's errors history, didactical material used with the trainee, and a history of the whole instructional process.

The affective component of the operator is modeled using a OCC cognitive model of emotion. To determine the operator affective state we use the following factors: 1) operator personality traits, 2) operator knowledge state, 3) goals and 4) tutorial situation. The goals for our domain are: 1) to learn the topics related to the operation, 2) to perform tests and simulations successfully, and 3) to complete the experiment as fast as possible. Once the affective student model has been obtained, the tutor has to respond accordingly, and in order to do that, the tutor needs a model of affective behavior (ABM) which establishes parameters that enable a mapping from affective and cognitive student model to responses of the tutor. Figure 6 shows a block diagram for the operator model.

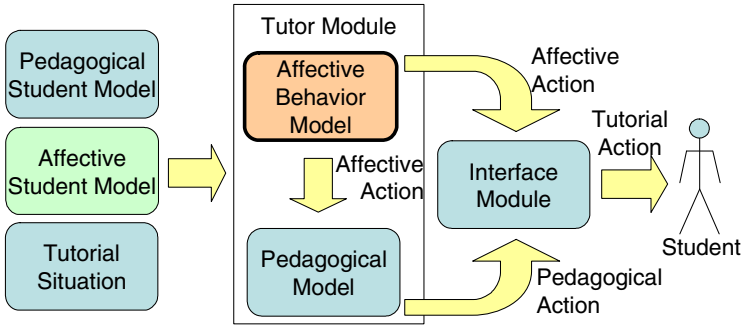


Fig. 6. Operator model

The general structure of our model is based on the proposed in [6]. The affective student model consists of a dynamic Bayesian network which is presented in Figure 7. This network is a high level representation of the model, thus the nodes in the figure are actually a set of nodes in the detailed model.

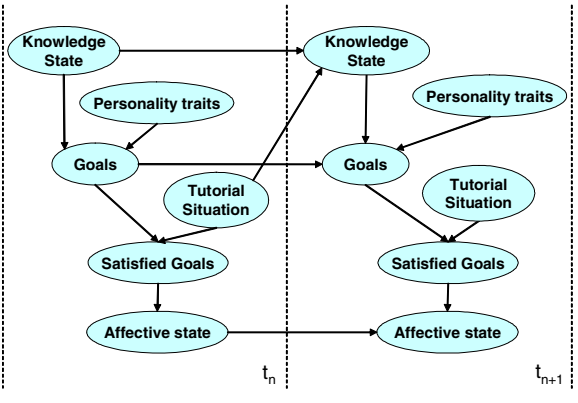


Fig. 7. Affective Operator model

6 Learning Management System

The learning management system (LMS) manages user learning interventions. The LMS supervises the training activities, such as the access to the intelligent environment, manages the user accounts, and keeps track of the training course of every user and generates reports. The importance of the LMS is the functionalities they offer in contrast to traditional training systems, such as self-registration, training workflow, on-line learning, on-line assessment, management of continuous professional education, collaborative learning, and training resource management. The one of the main activities of the LMS is provide communication services between users, authors and trainers such as videoconferencing, discussion threads.

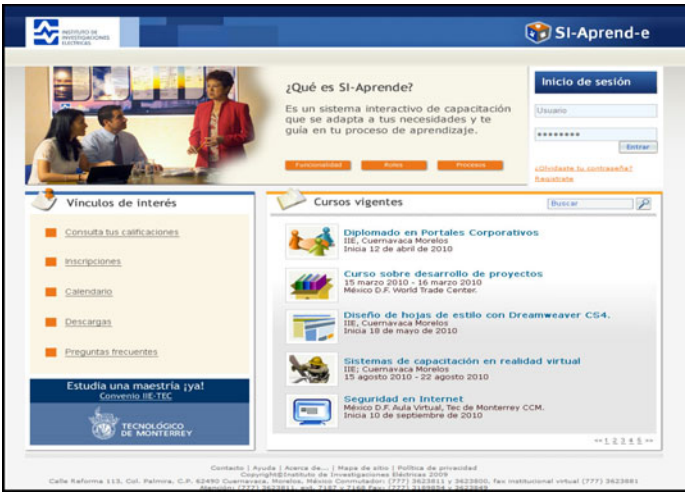


Fig. 8. Home page of ILS

7 Conclusions and Future Work

This paper presents the design and development of an Intelligent E-Learning System for training of Power Systems Operators. The aim was build an advance training system which includes: Computer based Training components labeled as Reusable Learning Objects; a Learning Object Repository; concept structure map of the power plant domain where the nodes represents concept of domain and the links relation between concepts; a tutor module based on adaptive sequence and operator model based on cognitive and affective components.

The adaptive sequence is represented as a decision network that selects the best pedagogical action for each specific operator, providing feedback to the operator and generating the sequence of LO that best match the operator learning process.

Operator model proposed is based on cognitive and affective components that have several advantages: flexibility, it allows to consider different models for each operator in a common framework; adaptability, by obtaining an initial model of a new learner from similar operator models, and modularity, it can be easily extended to include more trainees, and more experiments and other domains.

The results are encouraging, next step is to complete the integration of the IES for learning power systems processes. The implementation of the intelligent learning system in the National Electrical Sector provides a more dynamic and interactive training to the Operators of Power Systems, where the employees really experience the acquisition and transfer of skills and knowledge. Currently the system has been loaded with 130 learning objects for electric power operators training.

References

1. Brusilovsky, P., Vassileva, J.: Course sequencing techniques for large-scale web based education. Int. Journal Cont. Engineering Education and Lifelong Learning 13(1/2), 75–94 (2003)

2. ADL, Sharable Content Object Reference Model version 1.2: The SCORM Overview, Advanced Distributed Learning (2001), http://www.adlnet.org
3. Rodríguez-Ortiz, G., Paredes-Rivera, J., Argotte-Ramos, L., Arroyo-Figueroa, G.: Learning Objects Planning for the Training of the Power Generation Operation and Maintenance Personnel. In: IEEE Electronics, Robotics and Automotive Mechanics Conference, vol. II, pp. 349–354 (2006)
4. Galvan, I., Ayala, A., Muñoz, J.: Virtual Reality System for Power System Training. In: International Conference on Education and Information Technologies (ICEIT 2010), Proceedings of the World Congress on Engineering and Computer Science, WCECS 2010, San Francisco, USA, October 20-22, vol. I (2010)
5. Argotte, L.: Intelligent E- Learning model for adaptive sequence of learning objects (In Spanish), Msc Thesis, Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Ciudad de México (2010)
6. Hernández, Y., Noguez, J., Sucar, E., Arroyo-Figueroa, G.: A probabilistic model of affective behavior for Intelligent Tutoring Systems. In: Gelbukh, A., de Albornoz, Á., Terashima-Marín, H. (eds.) MICAI 2005. LNCS (LNAI), vol. 3789, pp. 1175–1184. Springer, Heidelberg (2005)

Interaction Based on Contribution Awareness in Collaborative Learning

Yuki Hayashi[1], Tomoko Kojiri[2], and Toyohide Watanabe[1]

[1] Graduate School of Information Science, Nagoya University
[2] Faculty of Engineering Science, Kansai University
{yhayashi,watanabe}@watanabe.ss.is.nagoya-u.ac.jp,
kojiri@kansai-u.ac.jp

Abstract. In the collaborative learning, each participant progresses their discussion by observing his/her contribution to other participants. By grasping these contributive atmosphere through networks, participants will be promoted to discuss more actively. As contribution information in the discussion, there are two types of contributions. *Short-term contribution* corresponds to the immediate contribution for other participants. *Long-term contribution* indicates the contribution of each participant through the whole discussion. In this paper, we propose a visualization method of contribution in the collaborative learning environment. For short-term contribution, the reference action to the useful utterance is represented by moving lighting balls from the utterer locations to the referrers. In addition, referred utterances are appeared in the visualization window as long-term contribution. Experimental result showed that our method could appropriately give the awareness for thinking/making useful utterances.

Keywords: Round-table interface, CSCL, contribution awareness.

1 Introduction

In the collaborative learning, participants actively progress their discussion by observing the learning situation. In the face-to-face environment, participants can feel the contribution of their utterances by observing other participants' actions of taking a note when useful utterances are occurred. By grasping this nonverbal information timely, participants are promoted to make further contributive utterances for their discussion.

In order to support the learning in the distributed environment, the issue in computer-supported collaborative learning (CSCL) is one of the interesting research fields [1]. Participants can easily study with others through networks regardless of times and locations. However, participants cannot keep successfully the motivation for their discussion because of the restricted communication means. By reflecting the contributive atmosphere of the discussion, it will enhance the motivation of the utterer for making useful utterances. In order to activate the conversation, Viegas, et al. proposed the chat system in which participants are represented by colored circles with their utterance texts [2]. The circles become bigger when participants post a message

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 104–113, 2011.

so as to reflect the active conversations. Kotani, et al. proposed the discussion support system in which roles of participants are timely represented as indexes for activating their discussion [3]. For calculating the index, the system uses the utterance information such as targets, types, and so on. However, in these systems, participants cannot feel how much they contribute in the discussion. For attaining the effective discussion, two types of contributions need to be considered. One is the *short-term contribution* that indicates immediate reaction of other participants toward the utterance. It does not reflect the exact contribution to the whole discussion, but it shows the situation that participants at least paid attention to each utterance. The other is the *long-term contribution* that represents effect of the utterance for the whole discussion. By providing these contributions, spontaneous motivation for making useful utterances for not only current topics but also whole discussion may be promoted.

For reflecting the contributive information through networks, the concept of *awareness* needs to be considered. Awareness provides the information: who is around, what activities are occurring, who is talking with whom and so on [4]. In order to support the real-time communication among participants, we have proposed a collaborative learning support system [5, 6]. In the interface, participants can be aware of the existence of other participants by changing view and grasp the flow of utterances by observing the moving utterance texts. The purpose of this paper is to improve our collaborative learning interface so as to grasp participants' contributions in the discussion. In order to reflect short-term contribution, the reference action of the useful utterance is represented by moving lighting balls from the utterer locations to the referrers. By observing this action intuitively, participants may be promoted to discuss the current topic. In addition, referred utterances are visualized as contribution utterances for representing long-term contribution. In the visualization window, contribution utterances are represented as colored circles whose sizes change according to the contribution degrees. The circles of the similar topic are connected and arranged nearer to each other. From observing the distribution of circles, participants can understand the contribution for the discussed topic. They may be lead to think an utterance considering the effect of whole discussion.

2 Approach

2.1 Round-Table Interface

Currently, we focus on the participants who study the learning subject in which many terms and their relations should be remembered such as history. In this kind of learning, the understanding knowledge of participants varies from person to person. Through the learning, they try to acquire the new knowledge.

In order to support the real-time communication through networks, we have proposed a collaborative learning support system in which focusing intentions of participants are reflected [5, 6]. Figure 1 shows the interface of our system. Here, participant *X* himself/herself, and other participants *A, B* and *C* are in the learning environment. Round-table window corresponds to each participant's view. In the window, other participants are represented as their real-time camera images through

the web camera. Participants can discuss the learning subject by using text-based chat in the window. According to the action such as making utterance, focusing target of each participant is automatically estimated, and its size changes based on the target. In addition, utterance texts are moved in the interface based on the input utterance target information to represent who is talking with whom. For providing the important utterances for participants, focusing utterances are estimated and emphatically represented in the interface. Figures 2(a) and (b) show the example of utterance transfer and their displaying images in the round-table window. If *C* sends a chat message as her utterance to *A*, the text moves from the location *C* to *A*. Figure 2(c) is the case that *B*'s utterance is detected as the *X*'s focusing utterance. Its text is represented with a different color (red). In addition, its font size becomes bigger to distinguish them from other utterances.

In this research, we try to reflect the participants' contribution in discussion into this interface.

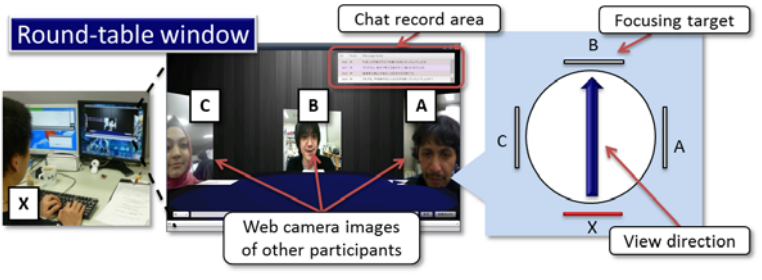


Fig. 1. Interface of our collaborative learning support system

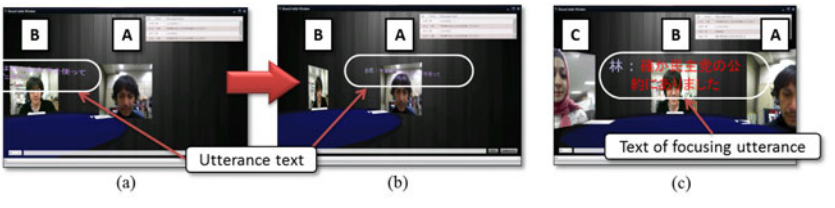


Fig. 2. Example of moving utterance among participants and displaying focusing utterance

2.2 Contribution Awareness in Collaborative Learning

In order to attain the fruitful discussion in the collaborative learning, it is important for participants to exchange many useful utterances relating to the learning. In the real world, participants timely take a note when useful utterances are occurred. This action brings the feeling of his/her contribution for other participants. It is considered that participants tend to refer to the utterances which include much new/interesting knowledge. By grasping the reference action, the motivation of the utterer for making useful utterances may be enhanced. However, this kind of nonverbal information cannot be grasped by existing chat system.

In this research, we regard the process of making useful utterances as three steps: *(i) being aware of useful utterance*, *(ii) thinking of what is useful utterance for the learning*, and *(iii) trying to make useful utterances*. Step *(i)* is the basis of thinking and making useful utterances in steps *(ii)* and *(iii)*. We try to support step *(i)* by reflecting two aspects of contribution awareness: *short-term contribution* and *long-term contribution*. Short-term contribution corresponds to the immediate contribution for other participants. This contribution occurs when utterance is referred by other participants. If participants can observe this reference action of others, they may be promoted to make an utterance for the current topic. On the other hand, long-term contribution indicates the contribution of each participant through the whole discussion. When many participants refer to the utterance, the utterance is considered to be contributive for their learning. Through the discussion, topics vary according to the knowledge acquisition activity. We reflect the long-term contribution for the scaffolding of step *(ii)*. If participants notice who have contributed for each discussed topic, they may think of what is the useful utterance for their whole discussion.

For reflecting short-term contribution, our interface shows the utterance reference action. It is desirable to intuitively visualize the information without disturbing their real-time learning. Therefore, when the reference action occurred, the lighting ball moves from the utterer to the referrer in the round-table window. In addition, the referred utterances are displayed in the visualization window for representing long-term contribution. In order for participants to grasp the contribution of whole discussion, the referred utterances are appeared in the visualization window. In this paper, we define *contribution utterances* as the referred utterances which appear in the visualization window. The contribution utterance is visually represented as a colored circle whose size becomes larger when many participants acquire knowledge from the utterance. The circles of related topics are connected. Participants can understand whose utterances and what kind of utterances are contributive through the learning from the distribution of circles. Table 1 summarizes our approach for each contribution type and their expected effects in our target learning environment.

Table 1. Contribution awareness types and approach of this research

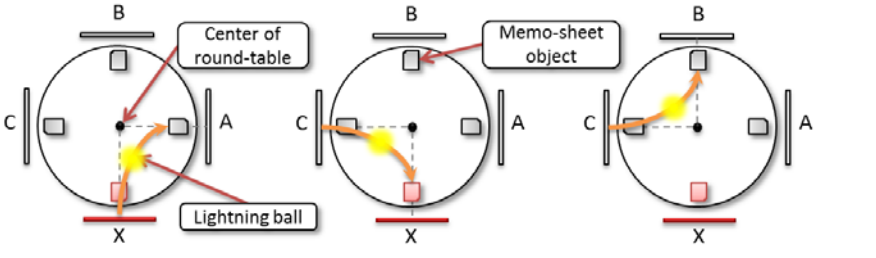
Contribution Type	Approach	Expected effects
Short-term contribution	Reflecting utterance reference action	Making utterances for current topic
Long-term contribution	Visualizing contribution utterances	Making utterances considering the effect of whole discussion

3 Visualization of Utterance Reference Action

In order to represent the utterance reference action as short-term contribution of participants, lighting balls move in the round-table window. When participants refer to the useful utterances, lighting balls move from the utterer locations to the referrer locations. Figure 3 shows trajectories of lighting ball. *X* represents the participant himself/herself, and *A, B* and *C* represent other participants. Figure 3(a) represents the case where *A* refers to *X*'s utterance as useful utterance. In this case, the lighting ball

moves from *X*'s location to the *A*'s location. *X* can observe this action from his/her view. The trajectory of the moving ball is determined by using Bézier Curve which interpolates the center of the round-table. Figure 3(b) is the situation that *X* refers to *C*'s utterance. The ball moves from the location of *C* to the *X*'s location as the case of Figure 3(a). In the interface, participants can grasp not only the reference action which relates to him/her but also other participants. Figure 3(c) shows the case that *B* refers to *C*'s utterance.

In order for the system to grasp the reference action, selecting function of useful utterances for the chat record is provided. In this function, participants can click the useful utterance on the round-table window to copy the utterance text to their memo windows. Based on the click, the lighting ball moves from the utterer to the participant who clicks the utterance.



(a) *X*'s utterance is referred by *A*. (b) *X* refers utterance of *C*. (c) Reference action occurred among other participants *C* to *B*.

Fig. 3. Trajectories of lighting ball in round-table window

4 Visualization Method of Contribution Utterances

For visualizing long-term contribution of participants, our system calculates contribution degrees of the referred utterances based on quantity of the knowledge that is acquired from the utterance. The learning goal of history is to understand the relations between items, such as parson, year and event. So, the acquired knowledge is represented as a set of relations among terms that are written on the participants' memo-sheets. When participants click and copy the utterance to the memo window and edit it, the contribution degree is calculated by counting the number of different pair of the terms in the edited texts. In order to detect the pair of terms, the terms included in the edited text are extracted as a successive noun or undefined words by using morphological analyzer for Japanese language [7]. The terms in participant *x*'s edited text derived by utterance *u* are expressed as $\boldsymbol{t}(u, x)=(t_a, t_b, \ldots t_n)$. Acquired knowledge $\boldsymbol{k}(u, x)$ is represented as a set of available combinations (t_i, t_j) where $t_i \neq t_j$. Here, the number of acquired knowledge $|\boldsymbol{k}(u, x)|$ equals ${}_{|\boldsymbol{t}(u, x)|}C_2$. The contribution degree of referred utterance *u*, such as $cd(u)$, is calculated as the following expression 1.

$$cd(u) = \sum_{\forall p \in P} \left| k(u, p) \right| \tag{1}$$

In this expression, ***P*** corresponds to the set of participants except the utterer of *u*. *cd(u)* becomes large if many participants refer to utterance *u* and edit the texts that include many terms.

In the visualization window, contribution utterances which indicate the referred utterances are appeared according to the calculated contribution degrees. Contribution utterances are represented as colored circles. The color distinguishes each participant who made the contribution utterance. In order to represent the contribution degree, the radius of the circle *u* is determined as the following expression 2 using *cd(u)*.

$$r(u) = (1 + \log(cd(u) + 1)) \times \alpha \tag{2}$$

Here, α indicates the predefined minimum value of the circle radius. *r(u)* becomes large in proportion to *cd(u)*.

For grasping the relation among contribution utterances, terms included in the edited text appear on each circle. If plural participants edited using the same terms by the same referred utterance, the terms are represented by bold type. In addition, when the same term is included among different contribution utterances, the circles of the utterances are connected by the edge. Figure 4 represents the visualization of contribution utterances. In circle *a*, terms k_2 and k_3 are emphasized because these terms are used in plural editing text. Circles *b* and *c* are connected by edge since the term k_7 is included in both circles. In the visualization window, each circle is automatically assigned based on spring model.

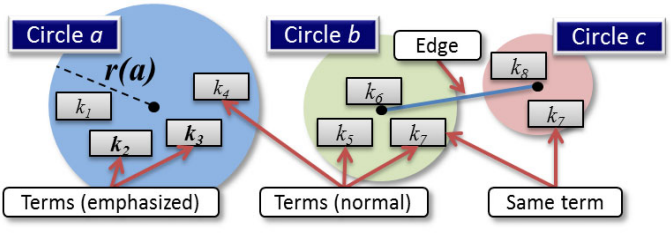


Fig. 4. Visualization of contribution utterances

5 Prototype System

We have implemented the prototype system that embeds mechanisms for displaying the utterance reference action and the contribution utterance into our collaborative learning system. Figure 5 shows the example of utterance reference action in the round-table interface. When participant *A* double-clicks the *X*'s utterance in the round-table window, lighting balls move from the location of *X* to the location of *A*. Then, the text of the referred utterance is appeared in the memo window (Figure 6). Participants can edit the text freely. By pushing the input button in the window, the text is added to the acquired knowledge area. Participants can revise/delete these texts from the right-click menu.

Figure 7 is the example of displaying contribution utterances in utterance visualization window. When participants edit the referred utterance, the utterances

appear in the visualization window as contribution utterances. On the circle, terms in the edited text that are extracted by morphological analyzer are appeared. When participants edit the utterance texts which are already appeared in the visualization window, the terms that are not appeared on the circle are added. Terms by which more than one participant edits the text are emphatically represented. Circles are connected by edge if the circles contain the same term. Participants can view the original utterance text by clicking the circle.

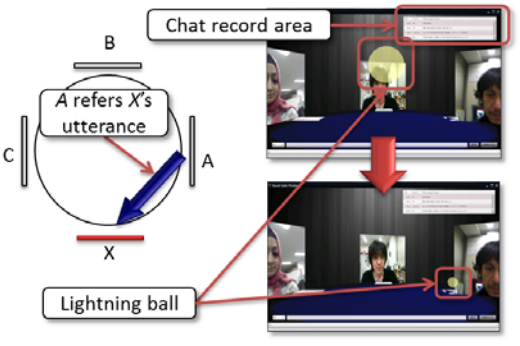


Fig. 5. Example of utterance reference action in round-table window

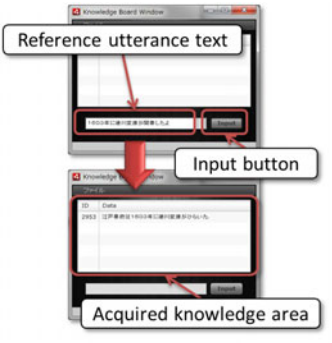


Fig. 6. Example of edited reference texts in memo window

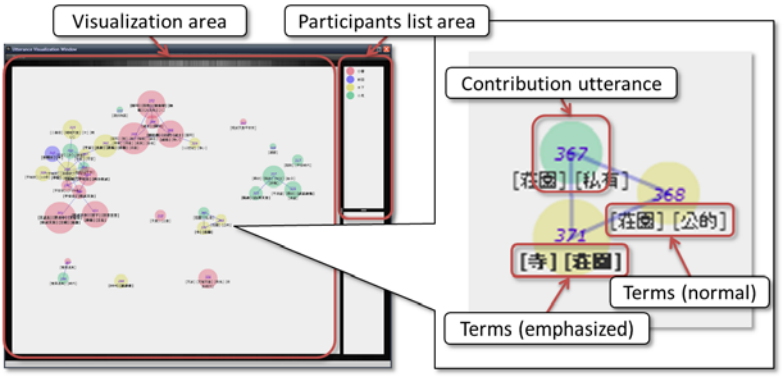


Fig. 7. Example of displaying contribution utterances in utterance visualization window

6 Experiment

6.1 Experimental Setting

We have conducted an experiment to evaluate the reflection method for contribution awareness. For this experiment, groups 1 and 2 of four participants in our laboratory were participated. Each group was asked to discuss two topics of Japanese history: Heian Period and Muromachi Period. Each participant has already studied about these topics in junior high school. Before the experiment, they were given the printed paper

of Wikipedia corresponding to these topics. In order to become accustomed to the interface, participants were asked to manipulate it. Then, they were asked to discuss using our system. Through the discussion, participants could use the web browser if they wanted to search the topic.

In the first round of the learning, participants of both groups used the system without visualization window to evaluate the effect of utterance reference action. After the discussion, they were asked to answer questionnaires of the utterance reference action as shown in Table 2. For each question, participants chose from 1 (worst) to 5 (best). After the first discussion, they discussed the second topic using visualization window. Then, they were asked to answer questionnaires about the appropriateness of the display method of contribution awareness as listed in Table 3. For answering the questionnaires, they were allowed to operate the visualization window. In addition to these questions, they were asked to answer the questions about their attitudes in discussion; how to refer the information and what kind of utterances you tried to make by observing the information.

Table 2. Questionnaire about display method of reference actions

No.	Questions
1-1 (a)	Could you recognize who refer to whose utterance?
1-1 (b)	Could you recognize which utterance is referred?
1-2 (a)	Did you feel that you could contribute to other participants by observing reference actions?
1-2 (b)	Did you think about what was the useful utterance by observing reference actions?
1-2 (c)	Did you try to make useful utterances by observing the reference actions?

Table 3. Questionnaire about display method of contribution utterances

No.	Questions
2 (a)	Could you grasp contribution degrees of participants?
2 (b)	Could you grasp contribution degrees of contribution utterances?
2 (c)	Could you grasp differences of topics?

6.2 Experimental Results

Figure 8 shows the average score of the questionnaires in Tables 2 and 3. Questions 1-1(a) and (b) inquired the validity of visualizing the reference action as moving lighting ball. From the result, participants could intuitively observe the reference action without their discussion disturbed. Questions 1-2(a) to (c) asked about the effectiveness of the reference action. A number of participants gave high scores for each question. They commented that the reference action for his/her utterances impressed the feeling of helping other participants timely, and they were promoted to make the utterances which relate to the current topic. As a result, our utterance reflection method could enhance the motivation for making useful utterances for the current topic.

Questions 2(a) to (c) asked the validity of visualizing contribution utterances. According to the results of 2(a) and 2(b), the majority of participants could grasp the participants who had contributed to their discussion and the contribution degree of their utterances. Therefore, our system could calculate the contribution degrees of each participant correctly. On the other hand, some participants selected low scores in question 2(c). They commented it became difficult to understand the topic of circles which were connected by many edges because of many terms. Since current method directly display the terms included in contribution utterance, it is necessary to improve the representative terms that should summarize method not only one node but also the connected nodes: e.g., using network analysis method of measuring centrality in data mining [8].

Table 4 summarizes the result of referred information in the visualization window during the discussion. From the comments, participants often referred to the circle colors and the sizes information at the same time. They also referred to the cluster of circles and terms information together. Some participants used the information to make useful utterances considering not only the current topic but also whole accumulated contribution utterances in the visualization window.

Through the discussion, participants in both groups frequently concentrated on the round-table window as the discussion became active. Thus, the rate of participants using visualization window is not so large as shown in Table 4. For our future, we have to modify the usability of our interface so as to access the information easily in visualization window.

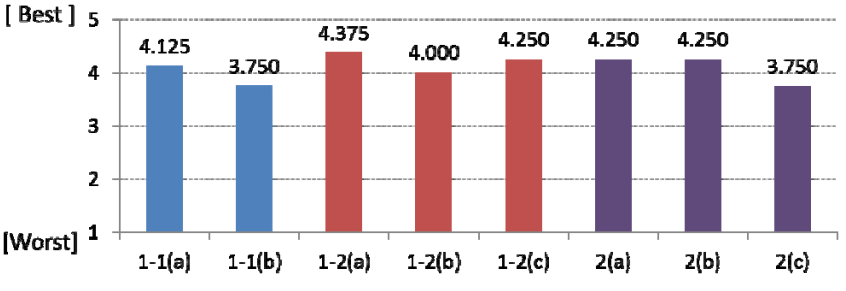


Fig. 8. Average score of questionnaires

Table 4. Referred information in visualization window

Referred Information	Rate of referred participants	The kind of utterances participants tried to make by observing the referred information
Colors of circles	75.0% (6/8)	- Utterance to appear as a large circle - Utterance to urge the participant whose utterance did not appear as circle - Utterance including knowledge that other participants might not know
Sizes of circles	62.5% (5/8)	
Clusters of circles	50.0% (4/8)	- Utterance related to the past discussed topics - Utterance that trigger new topics - Utterance including new terms which are not discussed so far
Terms on circles	37.5% (3/8)	

7 Conclusion

In this research, we proposed the reflection method for short-/long-term contribution in the collaborative learning. In the interface, lighting balls intuitively move from the utterer locations to the referred participants as short-term contribution. In visualization window, the referred utterances are appeared as colored circles so as to reflect the long-term contribution. Based on the experimental results, participants were enhanced to make the utterances which relate to the current topic. In addition, they could grasp whose utterances and what kind of utterances were useful through their discussion by observing the visualization window, and some of them tried to make useful utterances considering the effect of the whole discussion.

In this evaluation, we conducted experiments on only 2 groups and evaluated the effectiveness of the reflection method from the subjective answers of participants. For our future work, further experiments are required to confirm how short-/long-term contribution influences the quality of discussion in more detail, and to clarify the effective situations for this interface.

Acknowledgement. This research was supported by Grant-in-Aid for Research Fellow of the Japan Society for the Promotion of Science (No. 21-1764).

References

1. Andriessen, J.H.E.: Working with Groupware. Springer, Heidelberg (2003)
2. Viegas, F.B., Donath, J.S.: Chat circles. In: Proc. of SIGCHI 1999, pp. 9–16 (1999)
3. Kotani, T., Seki, K., Matsui, T., Okamoto, T.: Development of Discussion Supporting System Based on the Value of Favorable Words' Influence. Transactions of the Japanese Society for Artificial Intelligence: AI 19, 95–104 (2004) (in Japanese)
4. Gutwin, C., Stark, G., Greenberg, S.: Support for Workspace Awareness in Educational Groupware. In: Proc. of CSCL 1995, pp. 147–156 (1995)
5. Hayashi, Y., Kojiri, T., Watanabe, T.: Focus Support Interface Based on Actions for Collaborative Learning. International Journal of Neurocomputing 73, 669–675 (2010)
6. Hayashi, Y., Kojiri, T., Watanabe, T.: A Method for Detecting Focusing Utterance through Discussion in Collaborative Learning. In: Proc. of ICCE 2010, pp. 186–190 (2010)
7. Morphological Analyzer: Sen, `http://ultimania.org/sen/`
8. Maimon, O., Rokach, L.: The Data Mining and Knowledge Discovery Handbook, pp. 417–432. Springer, Heidelberg (2005)

The MATHESIS Semantic Authoring Framework: Ontology-Driven Knowledge Engineering for ITS Authoring

Dimitrios Sklavakis and Ioannis Refanidis

University of Macedonia, Department of Applied Informatics,
Egnatia 156, P.O. Box 1591, 540 06 Thessaloniki, Greece
{dsklavakis,yrefanid}@uom.gr

Abstract. This paper describes the MATHESIS semantic authoring framework being developed within the MATHESIS project. The project aims at an intelligent authoring environment for reusable model-tracing tutors. The framework has three components: an intelligent web-based model-tracing algebra tutor, an ontology and a set of authoring tools. The tutor serves as a prototype for the development of the ontology. The purpose of the ontology is to provide a semantic and therefore inspectable and re-usable representation of the declarative and procedural *authoring knowledge* necessary for the development of any model-tracing tutor, as well as of the declarative and procedural knowledge of the specific tutor under development. The procedural knowledge is represented via the *process model* of the OWL-S web services description ontology. Based on such an ontological representation, a suite of authoring tools is being developed at the final stage of the project.

Keywords: authoring systems, ontologies, semantic web, intelligent tutoring systems, model-tracing, web based tutors, OWL-S.

1 Introduction

Intelligent tutoring systems (ITSs) and especially model-tracing tutors have been proven quite successful in the area of mathematics. Despite their efficiency, these tutors are expensive to build both in time and human resources [1]. The main reason for this is the classic *knowledge acquisition bottleneck*: the extraction of knowledge from domain experts, the representation of this knowledge and the implementation of it in effective ITSs.

As a solution, authoring programs have been built having as their main purpose the reduction of development time as well as the lowering of the expertise level required to build a tutor. An extensive overview of these authoring tools can be found in [2]. Still, these tools suffer from a number of problems such as isolation, fragmentation, lack of communication, interoperability and re-usability of the tutors that they build.

The main goal of the ongoing MATHESIS project is to develop authoring tools for model-tracing tutors in mathematics, with knowledge re-use as the primary characteristic for the authored tutors as well as for the authoring knowledge used by

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 114–123, 2011.

the tools. For this reason, in the first stage of the MATHESIS project the MATHESIS Algebra tutor was developed to be used as a prototype target tutor [3]. In the second stage, based on the knowledge used to develop the MATHESIS algebra tutor, an initial version of the MATHESIS ontology has been developed [4]. The MATHESIS ontology is an OWL ontology developed with the Protégé OWL ontology editor. As this initial version of the ontology was developed in a bottom-up direction, it emphasized on the representation of the tutor, namely the interface, tutoring and domain expertise models. Of course, the ontology contained also a representation of the authoring knowledge but in a rather conceptual level. The project is now in its third stage were the authoring tools are being developed. These include tools for developing: a) the declarative tutoring knowledge (interface, cognitive tasks), b) the procedural tutoring knowledge (domain expertise and tutoring models) and c) the authoring knowledge.

This paper describes these authoring tools, how they are integrated with the MATHESIS ontology and how they operate on it to build both a model of the tutor under development which can be parsed and produce the tutor's code (HTML and JavaScript) as well as an *executable* model of the authoring processes that produced the tutor's model. It must be stressed out that because of the complex intertwining of the MATHESIS ontology with the MATHESIS authoring tools, the development of the authoring tools led to the following modifications of the initial ontology: a) the Document Object Model (DOM) representation of the tutor's HTML interface was simplified, b) the representation of the procedural tutoring knowledge (domain and tutoring model) was simplified and refined by introducing a taxonomy of atomic processes representing the various JavaScript statements, c) the authoring knowledge was developed as a full *executable model authoring language*, OntoMath, with composite authoring processes and atomic authoring statements.

The rest of the paper is structured as follows: Section 2 describes the newer, refined and extended version of the MATHESIS ontology as it is integrated with the authoring tools. Section 3 describes the authoring tools developed on top of this ontological representation. Section 4 presents some related work and, finally, Section 5 concludes with a discussion and further work for the development of the MATHESIS framework.

2 Semantic Knowledge Representation in the MATHESIS Framework: The MATHESIS Ontology

The main component of the MATHESIS authoring framework is the MATHESIS Ontology. The ontology contains three kinds of knowledge: a) the declarative knowledge of the tutor, such as the interface structure and student models, b) the procedural knowledge of the tutor, such as the teaching and math domain expertise models and c) the authoring knowledge, i.e. the declarative and procedural knowledge that is needed to develop the other two kinds. While the declarative knowledge is represented with the basic OWL components, i.e. classes, individuals and properties, the procedural knowledge, both tutoring and authoring, is represented via the *process model* of the OWL-S web services description ontology. By using OWL-S, every authoring or tutoring task is represented as a *composite process*.

2.1 Procedural Knowledge Representation: The OWL-S Process Model

OWL-S is a web service description ontology. Every service is an instance of class Process. There are three subclasses of Process, namely the AtomicProcess, CompositeProcess and SimpleProcess (not actually used). Atomic processes correspond to the actions a service can perform by engaging it in a single interaction; composite processes correspond to actions that require multi-step protocols.

Composite processes are decomposable into other composite or atomic processes. Their decomposition is specified by using control constructs such as Sequence and If-Then-Else. Therefore, any composite process can be considered as a tree whose non-terminal nodes are labeled with control constructs, each of which has children specified using the components property. The leaves of the tree are invocations of other processes, composite or atomic. These invocations are indicated as instances of the Perform control construct. The process property of a Perform indicates the process to be performed. In Figures 3 and 4 (Section 3) the tree structures of three different composite processes can be seen, as they are displayed to the user by the MATHESIS authoring tools. This tree-like representation of composite processes is the key characteristic of the OWL-S process model used in the MATHESIS ontology to represent authoring, tutoring and domain *procedural* knowledge *declaratively*, as it will be described in the following sections.

2.2 Tutor Representation in the MATHESIS Ontology

The MATHESIS project has as its ultimate goal the development of authoring tools that will be able to guide the authoring of real-world, fully functional model-tracing math tutors. The MATHESIS Algebra tutor is a Web-based, model-tracing tutor that teaches expanding and factoring of algebraic expressions. It is implemented as a simple HTML page with JavaScript controlling the interface interaction with the user and implementing the tutoring, domain and student models. The user interface consists of common HTML objects as well as Design Science's WebEQ Input Control applet, a scriptable editor for displaying and editing mathematical expressions.

On the top level of the MATHESIS ontology, every tutor is represented as an instance of class ITS_Implemented, having two properties: a) hasDomainTask which keeps a list of Domain_Task instances, the math tasks that the tutor teaches, e.g. monomial_multiplication, and b) hasTopInterfaceElement which keeps the root of the HTML DOM, a Document instance. Every instance of Domain_Task, like monomial_multiplication, has three properties: a) hasTutoringModel which keeps an instance of ITS_Teaching_Model, the algorithm that the tutor uses to teach this task, e.g. Model_Tracing_Algorithm, and b) hasInputKnowledgeComponents and hasOutputKnowledge-Components which keep lists of Domain_Knowledge_Component instances, the math concepts given and asked for the task. For the monomial_multiplication task, the input knowledge components are two monomials and the output knowledge component is their product, a monomial too. Each Domain_Knowledge_Component like monomial, has two properties: a) hasInterfaceElement which keeps a list of HTMLObject instances, the HTML interface object(s) used to display the structure of the concept, e.g.

WebEQ_Input_Control and b) hasDataStructure which keeps a list of the programming data structures used to represent the concept in the tutor's JavaScript code.

The top level representation of the tutor's procedural knowledge in the ontology is the model-tracing algorithm represented as a composite process, named ModelTracing-Algorithm. The tree structure of the process, adapted for the monomial_multiplication task, is shown in Figure 3a (Element 9) as displayed by the authoring tools. Each step of the algorithm is a top level tutorial action: present the current problem solving state, get the correct solution(s) from the domain expertise model, provide help/hint, get the student solution, provide feedback, assess, discuss. Each of these steps is also a composite process analyzing the tutorial steps further down to more simple ones like giving a hint, showing an example, recalling math formulas or rules and so on. In programming terms, this composite process when translated to JavaScript code will be the main function that controls the whole tutoring process by calling other functions.

This recursive analysis ends when a composite process contains only atomic processes corresponding to JavaScript statements. For the monomial_multiplication task, the execute_monomial_multiplication-Execution process is analyzed in two other composite processes: multiplyCoefficients and multiplyMainParts. These two processes form the tutor's *domain expertise model*, which calculates the correct answer(s) in each step in order to be compared against the student's answer. Figure 1 shows the structure of process multiplyMainParts.

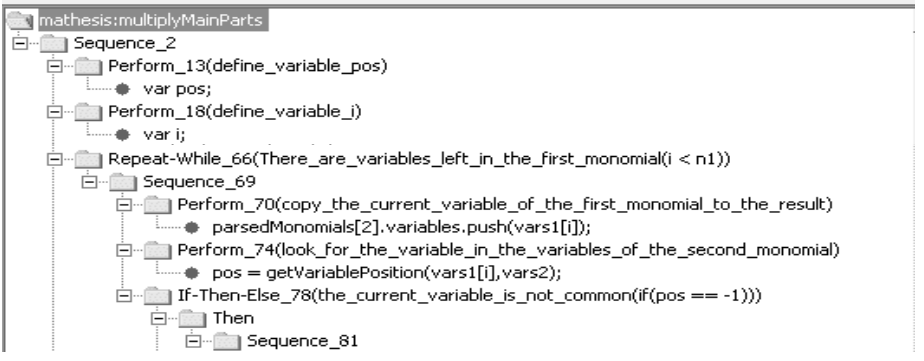


Fig. 1. Representation of the JavaScript function multiplyMainParts

Each JavaScript statement is represented by an instance of the JavaScriptStatement class, a subclass of AtomicProcess. Following the OWL-S representational scheme, these instances are parameters to Perform constructs. The JavaScriptStatement class has subclasses which classify the JavaScript statements in various classes such as DefineVariable, InitializeVariable, AssignValueToVariable, InvokeFunction, InvokeMethod, SetProperty. Each subclass has properties that represent the various parts of the corresponding JavaScript statements. For example, the InvokeFunction class has three

properties, hasInvokedFunction, hasArgumentsList and hasAssignedVariable. From the values of these properties the authoring tools create and display the actual JavaScript code as shown in Figure 1 (the dots under the Perform constructs). Such a detailed model of the JavaScript language allows the authoring processes to guide the non-expert author in building the tutor's code by selecting the appropriate JavaScriptStatement subclass and the values of the related properties. Therefore, a non-expert author does not need to know the JavaScript syntax but only have some general programming knowledge. As far as it concerns the semantic validation of the produced JavaScript code, the tools do not provide any special assistance to the authors. Therefore, the produced JavaScript code is syntactically correct but whether this code produces the intended behavior is a matter of correct design and analysis of the tutoring processes. In turn, this is a matter of correct design and analysis from the authors' part as is the case with any other system implementation.

At last, the representation of the HTML code and the corresponding Document Object Model (DOM) of the user interface are shown in Figure 2. Each object defined in the HTML code is represented as an instance of the corresponding HTMLObject subclass (Head, Body, Div, Input). Each instance has the corresponding HTML properties (body-onload, input-type). The DOM tree is represented via the two properties hasFirstChild and hasNextSibling. This representation allows for bi-directional creation of the HTML part of the user interface: a) The author creates in the ontology the representation of the DOM and then by traversing the DOM tree, the authoring tools generate the corresponding HTML code (top-down) or b) The user interface is created using any Web-page authoring program, the HTML file is parsed by Java's XML parser creating a DOM structure which in turn is transformed into its corresponding ontological representation for further authoring (bottom-up).

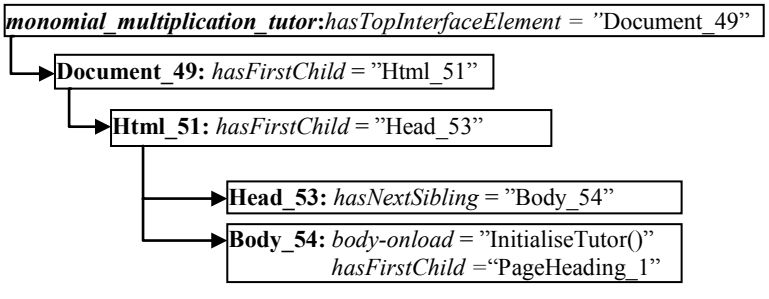


Fig. 2. The HTML User Interface Document Object Model Representation

2.3 Tutor Authoring Knowledge Representation

Within the MATHESIS framework, any authored tutor is represented in the MATHESIS ontology as described in the previous section. Expert authors have to create this ontological representation of the tutor. This is done by using the Protégé OWL interface to create classes, individuals, properties and values. The MATHESIS framework allows expert authors to capture the whole authoring effort by providing an *executable authoring model building* language, OntoMath. In OntoMath, each

authoring step is represented as an authoring process, composite or atomic. The authoring steps of a composite authoring process can be further decomposed in other authoring processes in exactly the same way that was described for tutoring processes in the previous section. Composite authoring processes correspond to functions of a programming language that can be called, getting and returning values. This is achieved by two properties: a) each composite process has a property, hasFormalParameters, that keeps a list of the process formal parameters, b) each Perform - the construct used to call a composite process – has a property, hasRealParameters, that keeps the list of the parameters at call time. During execution of a Perform construct, the interpreter matches the values of the real parameters to these of formal parameters. The values of the two properties are defined by the author in the ontology with the help of the authoring tools.

The recursive analysis of composite authoring processes ends to atomic authoring processes which are instances of the OntoMathStatement atomic process subclass. Each OntoMathStatement instance corresponds to an operation that must be performed to the MATHESIS Ontology such as create class, create instance, create property, get/set property value. OntoMath statements are *grounded* to actual Java program code. When the MATHESIS authoring tools interpret a Perform construct that calls an OntoMath statement, they execute its corresponding Java code, which performs various operations on the ontology that represents the tutor under development. It must be noted that these statements are not fixed. Expert authors can define their own atomic authoring statements by a) using the tools to define in the ontology the values of property hasFormalParameters for the new statement and b) writing the Java code that during execution gets the values of property hasRealParameters of the calling Perform construct and performs the statement's intended operation(s). The interpretation and execution of the OntoMath code by the MATHESIS authoring tools, as it will be described in section 3, leads to the creation of the ontological representation described in section 2.2, and therefore to the implementation of the authored tutor.

Therefore, the OntoMath authoring processes form a *meta-program* that handles the ontological representation of the tutor as its data. They capture the authoring expertise of expert authors and make it available to non-expert authors. They are the expertise model of an ITS authoring shell that, when executed, it guides non-expert authors to develop their own tutors.

3 The MATHESIS Authoring Tools

The MATHESIS authoring tools are currently implemented as a tab widget in Protégé. The tools are grouped in three windows according to their functionality: the Tutor authoring window, the Authoring processes authoring window (Author) and the MATHESIS ontology tab (Figure 3). The general philosophy of the tools is "point and click": the author selects a class or an instance in the ontology and clicks on a tool to perfrom an authoring action with it.

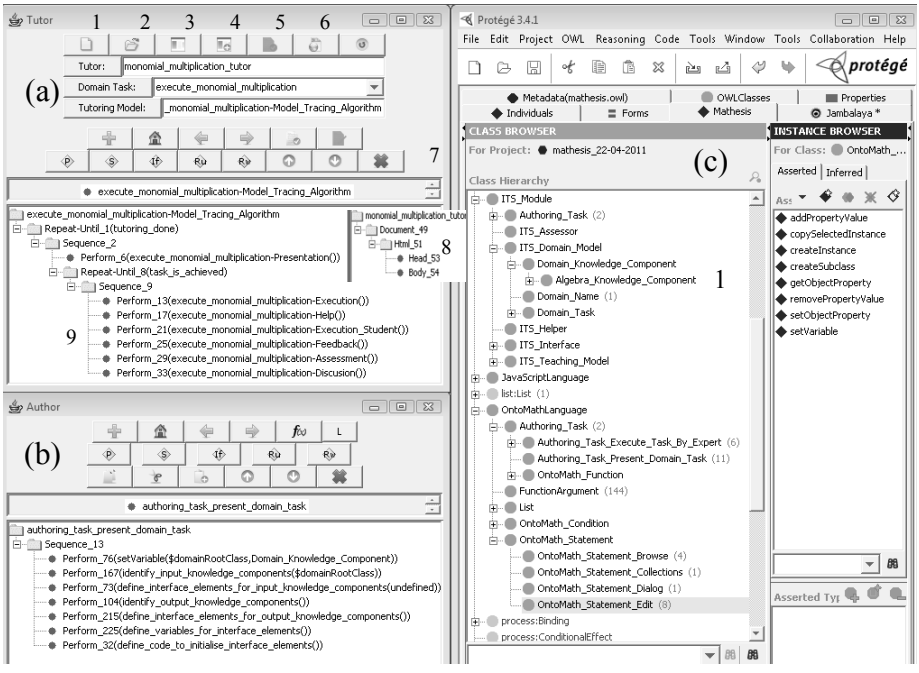


Fig. 3. The MATHESIS Authoring Tools: (a) Tutor window, (b) Authoring Processes Window, (c) The MATHESIS Ontology Tab

The Tutor authoring tools lie in the Tutor window (Figure 3a). With these tools an author can:

- Create a new instance of class ITS_Implemented for a new tutor (button a1) or select an existing one (button a2). In Figure 3a, the monomial_multiplication_tutor has been selected.
- Create a new or select an existing domain task that the tutor will teach (buttons a3, a4). New domain tasks are added as values to the hasDomainTask property of the tutor. The assigned tasks are displayed in a drop-down list. In Figure 3a the execute_monomial_multiplication task has been chosen to be authored.
- Assign to the selected domain task a generic tutoring model (button a5). Currently, only the Model_Tracing_Algorithm is defined in the ontology. The authoring tools create a new tutoring process by copying the structure of the generic tutoring process. In Figure 3a, the execute_monomial_multiplication-Model_Tracing_Algorithm has been assigned. Its structure is displayed by the tools (a9). As explained in section 2.2, the tutoring model is the main tutoring process from which all other processes are called. It is from this process that the authoring of the tutor starts.
- Switch views between the tutoring model (a9) and DOM representation of the tutor's HTML interface (a8). In reality the two structures are displayed separately.

The Tutor window also contains the *Tutoring Processes Authoring Tools,* arranged in two rows (a7). The tools of the first row allow the author to browse through the

tutoring processes that represent in the ontology the tutor's program code: when a Perform construct is selected, the author can move to the Performed (called) composite process. In the same time, a call stack is maintained so that the author can return back to the calling process. These tools are to be used by expert and non-expert users alike.

The second row of the *Tutoring Processes Authoring Tools* contains advanced tools, used mainly by expert authors. These tools allow the author to create and edit directly the tutor's processes (code). The reason for this is that some tutoring processes demand a very high level of authoring skills to be created. Expert authors can create these processes and provide them to non-expert authors as libraries. Examples of such tutoring processes are the model-tracing algorithm or an algorithm for parsing the MathML presentation code of an algebraic expression from the WebEQ Input Control applet.

Non-expert authors create tutoring processes by executing the corresponding authoring processes. This binding is done by the expert authors who set the hasAuthoringProcess property of a tutoring process through the advanced tools. When the hasAuthoringProcess property of a tutoring process points to an authoring process, with a click of a button the authoring process is displayed in the Author window. In Figure 3a, the tutoring process execute_monomial_multiplication-Presentation() has as a corresponding authoring process the authoring_task_present_domain_task, displayed in the Author window (Figure 3b). From this point, the next step is to execute this authoring process that will implement the tutoring process execute-monomial-multiplication-Presentation().

The Authoring Processes Tools lie in the Author window, shown in Figure 3b. They provide the same creating and editing operations for the authoring processes that form the *executable authoring model* of the implemented tutor. Their significant difference lies in the tracing (execution) of the Authoring Processes. When a composite authoring process is Performed (called) its tree structure appears in the Author window and its code is executed. When an atomic authoring process is Performed its corresponding Java code is executed.

As an example, the execution of the identify_input_knowledge_components authoring process (Fig. 4) will be traced. This authoring process is called by authoring process authoring_task_present_domain_task (Fig. 3b) and guides the non-expert author to select the domain concepts considered to be known for the currently selected domain task, execute_monomial_multiplication, of the authored tutor, monomial_multiplication tutor (Fig. 3a). For the monomial multiplication, the given domain concepts are two instances of class monomial. The OntoMath code is executed as follows:

1. Statement Perform_45(setSelectedClass($domainRootClass)) is performed. The interpreter gets the value of $domainRootClass, which is Domain_Knowledge_Component, and sets this class as selected in the Class Browser of the MATHESIS ontology tab (Fig. 3c). The author must browse into the subclasses of this class to find the monomial instance. Let's suppose that the instance exists (created by an expert author) and it has been found by the author.
2. Construct Repeat-While_2 starts an iteration executed for each knowledge component the author wants to identify.
3. Construct If-Then-Else_1 is executed having as its condition the authorConfirmation OntoMath predicate. To evaluate this predicate the interpreter displays a Yes/No question to the author ("Does the knowledge component already exist?").

According to step 1, the answer is "Yes", the predicate is true and execution continues with the "Then" part, i.e. Sequence_20.

4. Statement Perform_47(showMessageDialog(…)) prompts the author to select the existing domain concept in the Instance Browser of the MATHESIS ontology tab.
5. In construct If-Then-Else_22 a new instance of monomial is created, marked as selected and stored in variable $selectedInstance.
6. Statements Perform_55(getObjectProperty(…)) and Perform_58(setObjectProperty(…)) add the newly created instance of monomial to the list of values of property hasInputKnowledgeComponents of the execute-monomial-multiplication domain task.
7. Steps 3-7 are repeated until loop Repeat-While_2 is interrupted by the author, and the execution of the identify-input-knowledge-components authoring process is completed.

Fig. 4. The identify_input_knowledge_components authoring process

4 Related Work

The use of ontologies and semantic web services in the field of ITSs is relatively new. Ontological engineering is used to represent learning content, organize learning repositories, enable sharable learning objects and learner models, facilitate the reuse of content and tools [5]. The most relevant work to the MATHESIS framework is the OMNIBUS/SMARTIES project [6]. The OMNIBUS ontology is a heavy-weight ontology of learning, instructional and instructional design theories. Based on the OMNIBUS ontology, SMARTIES (SMART Instructional Engineering System) is a theory-aware system that provides a modeling environment and guidelines for authoring learning/instructional scenarios.

While the OMNIBUS/SMARTIES system provides support mainly for the design phase of ITS building, the MATHESIS framework aims at the analysis and development phases. It provides a semantic description of both tutoring and authoring

knowledge of any kind in the form of composite processes and the way to combine them as building blocks of intelligent tutoring systems. Thus, it provides the ground for achieving reusability, shareability and interoperability.

5 Discussion and Further Work

It is well known that the development of ITSs demands a considerable effort. The ontological representation of both the developed tutor and the authoring knowledge to develop it, definitely puts extra effort to the authoring endeavor. We believe that this extra effort pays off for the following reasons: a) the semantic representation of the tutor makes all parts of it open to inspection by other authors and therefore reusable, shareable and interoperable, b) the same holds for the authoring knowledge used to build the tutor and c) the OntoMath language is a proper programming language, completely open and configurable, therefore authoring processes can in principle build open, shareable and reusable authoring models for any kind of tutor.

Before this extra effort starts paying off, the MATHESIS semantic authoring framework must represent in its ontology a considerable amount of authoring and tutoring knowledge that now lies hidden and fragmented inside the various authoring tools and the authored tutors. As a first step, this will be done by representing a considerable part of the MATHESIS Algebra tutor into the ontology. For this purpose new authoring tools are needed: parsers for transforming between HTML and MATHESIS DOM representation; parsers for transforming between JavaScript and MATHESIS tutoring processes; extension of the OntoMath language and elaboration of its interpreter. These tools constitute our current research line.

References

1. Koedinger, K.R., Anderson, J.R., Hadley, W.H., Mark, M.A.: Intelligent Tutoring Goes to School in the Big City. International Journal of Artificial Intelligence in Education 8, 30–43 (1997)
2. Murray, T.: An overview of intelligent tutoring system authoring tools: Updated Analysis of the State of the Art. In: Murray, Ainsworth, Blessing (eds.) Authoring Tools for Advanced Technology Learning Environments, pp. 491–544. Kluwer Academic Publishers, Netherlands (2003)
3. Sklavakis, D., Refanidis, I.: An Individualized Web-Based Algebra Tutor Based on Dynamic Deep Model-Tracing. In: Darzentas, J., Vouros, G.A., Vosinakis, S., Arnellos, A. (eds.) SETN 2008. LNCS (LNAI), vol. 5138, pp. 389–394. Springer, Heidelberg (2008)
4. Sklavakis, D., Refanidis, I.: Ontology-Based Authoring of Intelligent Model-Tracing Math Tutors. In: Dicheva, D., Dochev, D. (eds.) AIMSA 2010. LNCS, vol. 6304, pp. 201–210. Springer, Heidelberg (2010)
5. Dicheva, D., Mizoguchi, R., Greer, J. (eds.): Semantic Web Technologies for e-learning, The Future of Learning, vol. 4. IOS Press, Amsterdam (2009)
6. Mizoguchi, R., Hayashi, Y., Bourdeau, J.: Inside a Theory-Aware Authoring System. In: Dicheva, D., Mizoguchi, R., Greer, J. (eds.) Semantic Web Technologies for e-learning, The Future of Learning, vol. 4, pp. 59–76. IOS Press, Amsterdam (2009)

An Intelligent Tutoring System Architecture for Competency-Based Learning

Miguel Badaracco[1] and Luis Martínez[2]

[1] National University of Formosa,
Faculty of Economics and Business Administration, Formosa, Rep. Argentina
[2] University of Jaén Spain,
Department of Computer Science, Jaén, Spain

Abstract. An Intelligent Tutoring System (ITS) aims to customize teaching processes dynamically according to student's profile and activities by means of artificial intelligence techniques. The architecture of an ITS defines its components where the pedagogical model is crucial, because the ITS complexity will depend on its scope (specific or generic). Our interest is focused on generic ITS that are very complex due to the fact that could be applied to different educational domains. This contribution proposes an architecture for ITS that uses a Competency-based learning pedagogical model, in order to manage the complexity and make them easier to understand, together a diagnosis process for such a type of systems.

Keywords*:* Intelligent Tutoring Systems, Competency-based education, knowledge representation.

1 Introduction

An ITS provides direct customized instruction or feedback to students in their learning processes by means of Artificial Intelligence (AI) techniques, mainly for knowledge representation, managing a teaching strategy like an expert both in the teaching and pedagogical domains in order to diagnose the student learning status. Hence the system should offer solutions, actions, processes that support the student learning process [1].

Due to the fact that, the ITS deals with the teaching purpose about an educational domain such as a course, a module, a subject, etc., together pedagogical criteria [2], they can be classified as *generic* or *specific*. The latter uses pedagogical criteria suitable for a specific educational domain, being such systems limited to just one domain. However the former deals with a multi-domain view and should be able to adapt its criteria to cover different teaching domains and make easier the learning process to the students.

Even though a generic ITS is more flexible, it is also more complex and such complexity implies problems as:

Model comprehension: It is difficult to build a system that joins different theoretical frameworks, evaluation criteria, knowledge representations, heuristics, etc.

Implementation: the model configuration and the interpretation of the different parameters might sometimes become confuse.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 124–133, 2011.

To solve the previous problems this contribution introduces a novel generic ITS architecture and its diagnosis process whose pedagogical model is the Competency-Based Education (CBE) [3], [4], [5]. The joint of ITS and CBE involves innovations in the components, processes and relationships by introducing new features regarding the components of an ITS to fit the CBE model in with it. We will present the knowledge representation of the domain, the student model and the diagnosis process according to the CBE. This architecture provides the teacher a platform that facilitates the implementation of customizing training proposals and tools that improve the performance of generic ITS.

The contribution is organized as follows, Section 2 reviews the main features of an ITS architecture. Section 3 presents the concepts and principles of competency-based education (CBE). In section 4 is introduced the proposal of an ITS model based on competences (ITS-C) and its diagnosis process. Finally we point out some conclusions.

2 Intelligent Tutoring Systems

An ITS is a dynamic and adaptive system for personalized instruction based on the students' characteristics and behavior that involves a combination of various fields such as: AI (expert systems, Bayesian networks, fuzzy logic, etc.), cognitive psychology and educational research. To achieve its objectives the design of an ITS must include an architecture that supports the associated processes. This section reviews a general architecture of an ITS, its knowledge representation and eventually it is presented a classification of ITS according to their scope and pedagogical criteria.

2.1 Architecture

In [6], [7] is defined the general architecture of an ITS by four components: i) domain model, ii) student model, iii) pedagogical model and iv) interface model. These components interact each other to accomplish different functions.

These components are described in further detail below according to [8] and [9]:

Domain model or what is taught? It contains knowledge about the subjects that must be learned. Anderson in [10] asserts that the more knowledge in a domain model the more powerful. The domain models can be classified in:

Black box model: It does not require an explicit coding of underlying knowledge. The domain model has been previously encrypted, being of interest their behavior.

Model based on the methodology of expert systems: It follows the same steps of an expert system. It involves extracting knowledge from experts and decides the way in which it will be codified and implemented.

Cognitive model: The domain model is obtained by abstracting the way in which humans make use of knowledge. This type of model is most effective from a pedagogical point of view, though implementing effort is higher.

Student Model or who is taught? It also implies what the student knows or does not know about the domain. The use of student models in ITS arises because these systems must work with uncertain and incomplete information about the students [11]. Many

ITS infer this model from the student's knowledge about the domain model. The process of instruction is adapted to the students' needs. The structure that stores the student's knowledge status is *his/her own model*, while the update process is called *diagnosis of the student*. Holt et al. [12] extend the classification presented in [13] about the student model proposing the following approaches:

Overlay Model: It represents the student's knowledge about the domain. The student's behavior is compared with an expert. The differences are assumed as gaps in knowledge of the learner.

Differential model: It defines the student's knowledge in two categories, expected knowledge that student should know and expected knowledge that should not know. This approach modifies the previous one representing explicitly the differences between expert's and student's knowledge.

Disturbance model: In this model, student's knowledge is not considered a subset of the expert's knowledge, but it is possible that the student knows some knowledge in different quantity and quality regarding the expert.

Instructional Model or how is it taught? It defines the teaching and tutorial strategies. We highlight three tutorial features that should have an ITS [14]:

- Control on the representation of knowledge to select and sequence the parts that should be supplied to the student.
- Ability to answer students questions about his/her instructional objectives and contents.
- Strategies to determine when the student needs support and how to select the appropriate help.

Interface: It supports the man-machine interaction. Additional efforts are needed to develop this element of the architecture, to make it intuitive and transparent to the student. Burton identifies key issues in the interface design [15]: wishing representation aspects of the domain, level of abstraction and accuracy of representation, order in the presentation of contents, proofing and support tools and assistance, level of control performed by the tool.

Different Intelligent Tutoring Systems are using the advantages offered by Internet, such as Haskell-Tutor [16], WHAT [17] (Web-Based Haskell Adaptive Tutor), ActiveMatch [18], etc.

2.2 Knowledge Representation

An ITS organizes an educational proposal according to pedagogical criteria. These criteria determine the knowledge representation in the domain model, student model and its update process. A common knowledge representation adopted by an ITS is a hierarchical network where nodes are concepts and related concepts are connected by arches or arrows. Generally, this structure may take the form of semantic networks, conceptual maps or Bayesian networks [8], [9]. The relationship among nodes can be of different types: aggregation, part-of, etc.

2.3 Classification of ITS

The customization of the teaching process for each student according to his/her necessities by an ITS is based on the use of training proposals and domains of knowledge (subject, capabilities, professional roles, etc.). Every training proposal is based on the design of a curriculum which might use different approaches that are driven by their teaching and learning views. Consequently the designers must assume pedagogical criteria that underlie the ITS.

The ITS can take different views to implement the pedagogical criteria according to its educational scope that generates a classification into *specific* and *generic* ITS:

1. *ITS for specific domain:* It uses pedagogical criteria suitable for just one specific educational domain, such systems are limited to that educational domain. Therefore, a specific ITS assumes pedagogical criteria adapted to a specific domain, whenever the designer is an expert or worked closely with an expert in the domain model, and the associated processes. This type of ITS does not present problems of implementation, because they suit the specific domain, achieving generally satisfactory results. Some examples are: ELM-ART (Episodic Learner Model Adaptive Remote Tutor) [19], ActiveMatch [18], WHAT (Web-Based Haskell Adaptive Tutor) [17], etc. The main problem of this type of ITS is that its implementation is limited to a single domain.
2. *ITS for generic domain:* An ITS for generic domains is designed to provide a framework to design and implement training proposals for multiple educational domains. Because of this, it could arise a problem when the teacher must adapt various components of the curriculum design to the specifications of the domain model and student model (which will depend on how knowledge is represented in these models). Another issue is the difficulty of producing correct interpretations of the parameters provided by the ITS, especially those ones that use heuristics for student's diagnosis. Some generic ITS are: TANGOW (Task-based Adaptive Learner Guidance on the Web) [20], ALICE (Adaptive Link Insertion in Concept-based Educational System) [21], etc.

3 Competency-Based Education (CBE)

This section reviews concepts about CBE that is the educational model used in our proposal.

The CBE is an emerging curriculum model that tries to satisfy the demands of learning contexts, by the developing competencies, enabling students to act in a complex world in constant transformation, [3]. A competence is the ability to perform effectively in a given situation, it is based on knowledge but it is not limited to it [22]. Competences are complex knowledge and represent the know-how to integrate conceptual, procedural and attitudinal knowledge.

The current importance and relevance of the CBE is showed in Tuning Educational Structures in Europe [5], The contribution of universities to the Bologna process and the Tuning Latin America project. Tuning serves as a platform for developing reference points at subject area level. These are relevant for making programmes of studies (bachelor, master, etc.) comparable, compatible and transparent. Following it

is described some key processes and concepts of the CBE such as the curriculum design based on competences and its descriptors.

3.1 Curriculum Design Based on Competences (CDBC)

The Curriculum Design Based on Competences (CDBC) is the process that performs a training proposal based on CBE. The training proposal is organized according to competency norms that are benchmarks to evaluate the performance achieved by students. Such norms contain a set of descriptors that reflect good professional practices that guide the development of competences. The validity of competency norms should have agreed among social actors as government, industry, education system, etc. [4].

3.2 Descriptors of CDBC

The basis of representation of the domain model and student model are the descriptors of the CDBC. Here we revise the set of descriptors that reflect good professional practices and guide the development of competences that integrate competency norms.

Competency unit (cu): It is a main function that describes and groups the different activities concerning the role or profile chosen.

Competency element (ce): It is the disaggregation of a main function (cu) that aims to specify some critical activities. A function (cu) can be specified by one or more competency elements (ce), according to its complexity or variety.

Evidence of performance (evd): It checks if a process is performed according to best practices.

Evidence of product (evp): It is a descriptor of tangible evidence in the results level, when the best practices have been used.

Evidence of knowledge (evk): It is a descriptor about scientific-technologic knowledge that allows the user understands, reflects and justifies competent performance.

4 An Architecture for an Intelligent Tutoring Systems Based on Competences (ITS-C)

In this section is introduced the proposal for the architecture of a generic ITS based on competences (ITS-C) that facilitates the overcoming of the implementation and comprehension problems presented by generic ITS.

It is necessary to establish a link between the use of ITS and the pedagogical model based on competences. To do so, we propose an architecture for ITS-C (Fig. 1 shows graphically) that extends the components of the general architecture presented in Section 2.

In the following subsections the domain and student models together the diagnosis process are further detailed to show the novelties and elements of the new architecture to be consistent with the CDBC approach.

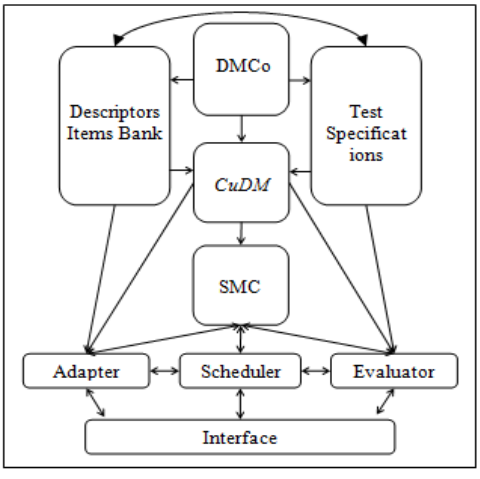


Fig. 1. An ITS-C Architecture

4.1 Domain Model

It contains the expert's competency profile in a knowledge domain. In an ITS-C it consists of four components:

1. *A domain model of competency (DMCo)*: It is represented by a semantic network whose nodes are competence units (*cu*), competence elements (*ce*), descriptors (*evd, evp, evk*) and their relations. In an ITS-C the DMCo does not represent expert's knowledge but the competency profile that is obtained from the functional map and the competency norms.
2. *A curriculum domain model (CuDM):* It rearranges the DMCo according to a teaching strategy that defines the competences associated to a professional profile to perform in different situations. The CuDM is a module based structure, developed from the CDBC of a educational purpose, being each module (M_i) the unit that defines the contents, activities and objectives to learn the necessary skills for solving problems in a professional area. Such skills are inferred from a competence elements whose descriptors are associated to a bank of items used to specify tests that allow the evaluation of a module.
3. *A set of descriptors*: The descriptors associated with the *ce* of the didactic modules are *evd, evp,* and *evk*, that belong to a bank of items.
4. *A bank of problems*: It is associated to the modules of the CuDM and the test specifications provided by the teachers.

4.2 Student Model Based on Competences (SMC)

In an ITS-C the student model of competence (SMC) stores the student's information and its representation in the diagnosis process, by using an overlay model in the semantic network of the CuDM, to evaluate the competences associated to each competence elements (*ce*) belonging to a module M_i (Fig. 2 shows graphically).

The nodes *evp, evd* and *evk* store a probability distribution $P(\theta_{evp} = |\vec{u}_i)$, $P(\theta_{evd} = |\vec{u}_i)$, and $P(\theta_{evk} = |\vec{u}_i)$ corresponding to the level of competency of the student in the respective node. Being θ the student's level of technical-scientific knowledge about the descriptor for a response pattern $\vec{u}_i$ obtained from the responses provided by the student in the test T_s.

Once the components of the CuDM are determined the elements involved in the diagnosis are defined:

1. The items bank associated to the *evk, evp* and *evd* of each *ce;*
2. The tests, T_s, associated to the *ce;*
3. Test specifications that determine the methods for selecting items, end criteria, etc.

Each node ce_i will compute the student's technical-scientific knowledge level about it, θ_{ce_i} (see equation (3)) that aggregates the probabilities for evk_i, evd_i and evp_i that reflects the student technical-scientific knowledge.

Once the distributions of the nodes, ce_i, have been obtained the level of competency for a module M_i is estimated by averaging the values of the respective elements of competence (ce_i).

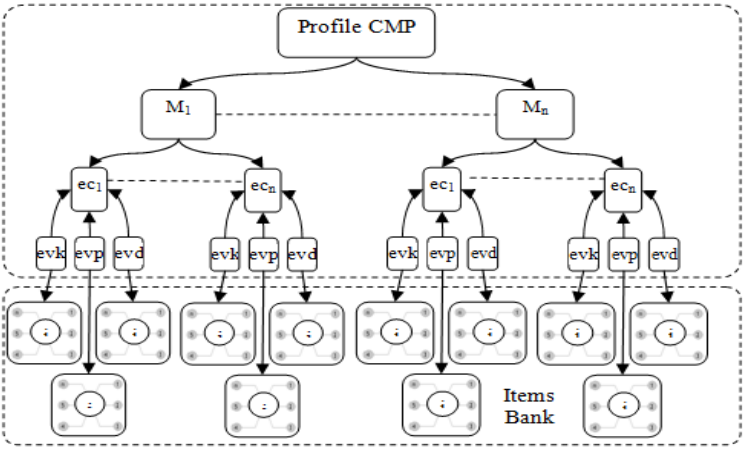


Fig. 2. Modules of a CuDM

4.3 Diagnosis Process Based on Competences

The diagnosis process estimates the level of competence achieved by the student in the didactic modules, M_i, to update the SMC. To do so, our model uses Computerized Adaptive Tests (CAT) [9] with a discrete and non parametric response model that is able to evaluate items with multiple answers. Its main components are:

1. *A response model associated to the items*: It describes the student expected performance according to his/her estimated knowledge. In our case we use the response model based on the Item Response Theory (IRT) [9].

2. *Bank of Items*: The better is its quality the more accurate will be the CAT to achieve the learning objectives. Each item Q_i is associated to the descriptors (evd, evp or evk) that solve, Q_i. And every choice to solve Q_i has assigned a characteristic curve of option (CCO) obtained by a calibration process based on the Ramsay algorithm [23]. Also the model defines the characteristic curve of response (CCR) and of Item (CCI). The former represents the probability that the student answers a given choice according to his/her knowledge and the latter the probability that he/she answers Q_i correctly. Each CCO is represented by a probability distribution noted as, $(\vec{u}_i|\theta_t)$, where each component of distribution represents the probability that the student selects the response pattern $\vec{u}_i$, given their level of competence θ_t.
3. *Initial level of Knowledge:* The initial knowledge estimation is crucial because it determines the CAT for each student.
4. *Criterion for selecting items:* the adaptive mechanism of CAT selects the items that should be solved by the students.
5. *Stop criterion*: The test should end when the student achieves a level of knowledge fixed a priori, though there are other criteria such as a maximum number of items is achieved, etc.

During the administration of a test, the student's knowledge is estimated each time that responds a item. The updating of the distribution of knowledge of the student is carried out using an adaptation of the Bayesian method proposed by [24] that updates the knowledge distribution as follows:

$$P\left(\theta_{ev_t}|\vec{u}_1,..,\vec{u}_i\right) = \begin{cases} \left|P\left(\theta_{ev_t}|\vec{u}_1,..,\vec{u}_{i-1}\right)Po(\vec{u}_i|\theta_t)\right| & if\ Q_i\ asseses\ evd_t, \\ & evk_t\ or\ evp_t. \\ P\left(\theta_{ev_t}|\vec{u}_1,..,\vec{u}_{i-1}\right)\ in\ other\ case. \end{cases} \quad (1)$$

Where $P\left(\theta_{ev_t}|\vec{u}_1,..,\vec{u}_i\right)$ is the a priori student knowledge estimation on evd_i, evp_i or evk_i ; and $Po(\vec{u}_i|\theta_t)$ the CCO for the option of the response pattern.

After the updating of the distribution of the nodes *evk*, *evd* and *evp* of the student, the system can estimate the level corresponding to the distribution by using one of the two choices introduced in the CAT [9]:

Expectation a posteriori (EAP), the value corresponding to the level of knowledge is the average of the distribution of probabilities. Formally:

$$EAP\left(P(\theta_{ev}|\vec{u}_n)\right) = \sum_{k=0}^{k-1} kP(\theta_{ev} = k|\vec{u}_n), \quad (2)$$

where *k* represents the knowledge level.

Maximum a posteriori (MAP), the level of knowledge value corresponds to that with the biggest probability assigned, i.e., the mode of distribution. Formally:

$$MAP\left(P(\theta_{ev}|\vec{u}_n)\right) = max\ P(\theta_{ev} = k|\vec{u}_n) \quad (3)$$

The competency level θ_{ce_i} in ce_i is computed by using the values of the nodes evd_i, evp_i and evk_i.

$$\theta_{ce} = k_d P(\theta_{evd} = k_d|\vec{u}_n) + k_p P\left(\theta_{evp} = k_p\middle|\vec{u}_n\right) + k_k P(\theta_{evk} = k_k|\vec{u}_n) \quad (4)$$

where $k_d P(\theta_{evd} = k_d|\vec{u}_n), k_p P\left(\theta_{evp} = k_p\middle|\vec{u}_n\right)$ *and* $k_k P(\theta_{evk} = k_k|\vec{u}_n)$ are the probability regarding the descriptors *evp, evd* and *evk* and *k* the competency level.

The diagnostic algorithm used is an adaptation of the proposed for CAT [9].

For each answered item the algorithm updates the $P\left(\theta_{ev_t}\middle|\vec{u}_i\right)$ of *evk*, *evd* or *evp*, and checks the stop criterion if it is fulfilled then the value θ_{ce_i} of the ce_i is calculated and updated (by using equation (4)) otherwise it is selected the following item (based on the selection criterion of items established in the test specifications) of the test, repeating the process until fulfills the stop criterion. In [2] we carried out an evaluative study of the algorithm successfully.

5 Conclusions

We have introduced innovations in the architecture of an ITS by presenting a new representation of the domain model, student model, and the diagnostic process based on the pedagogical model Competency Based Teaching (CBT) that is an emerging educational model with a promising the future in global education. Therefore the proposal of an ITS-C architecture based on the principles CBE is a quite interesting because the teachers incorporate into their practice the principles of CBE.

The use of CBE provides a better understanding of ITS-C architecture at the time of its implementation, thus overcoming the problems of ITS architectures based on other pedagogical models.

Acknowledgment. This paper has been partially supported by the research projects TIN2009-08286, P08-TIC-3548 and Feder Fonds.

References

1. Wenger, E.: Artificial intelligence and tutoring systems. Morgan Kaufmann Publishers, Inc., San Francisco (1987)
2. Badaracco, M., Martínez, L.: Design of Architecture for Intelligent Tutor System Based on Competencies (ITS-C). XVI Argentine Congress on Computer Science. Buenos Aires (2010)
3. Zalba, E.a.G., N.: An approach to education based on competencies, http://www.me.gov.ar/spu/guia_tematica/CPRES/cpres-comision.html
4. Catalano, A., Avolio de Cols, S., Sladogna, M.: Curriculum design based on standards of occupational competence. In: Oit, C. (ed.) Concepts and methodological guidelines. CINTERFOR, Buenos Aires (2004)
5. Europe TUNING, E.S., http://tuning.unideusto.org/tuningeu/images/stories/Temp.late/General_Brochure_Spanish_version.pdf
6. Sleeman, D., Brown, J.S.: Intelligent tutoring systems. Academic Press, Inc., London (1982)

7. Polson, M.C., Richardson, J.J.: Foundations of Intelligent Tutoring Systems. Lawrence Erlbaum Associates Publishers, Hillsdale (1988)
8. Millán, E.: Sistema Bayesiano para Modelado del alumno. University of Malaga Spain, Málaga (2000)
9. Guzmán de los Riscos, E.: An evaluation model based on cognitive for STI diagnosis. University of Malaga Spain, Malaga (2005)
10. Anderson, J.R.: The Expert Module. In: Polson, M.C., Richardson, J.J. (eds.) Foundations of Intelligent Tutoring Systems. Lawrence Erlbaum Associates Publishers, Hillsdale (1988)
11. Mayo, M., Mitrovic, A.: Optimising its behaviour with bayesian networks and decision theory. International Journal of Artificial Intelligence in Education 12 (2001)
12. Holt, P., Dubs, S., Jones, M., Greer, J.: The state of student modelling. In: Greer, J.E., McCalla, G. (eds.) Student modelling: The Key to Individualized Knowledge- Based Instruction. Springer, Berlin (1994)
13. Verdejo, M.F.: Building a student model for an intelligent tutoring system. In: Greer, J.E., McCalla, G. (eds.) Student Modelling: The Key to Individualized Knowledge-Based Instruction. Springer, Berlin (1994)
14. Burns, H.L., Capps, C.G.: Foundations of intelligent tutoring systems: An introduction. In: Polson, M.C., Richardson, J.J. (eds.) Foundations of intelligent tutoring system. Lawrence Erlbaum Associates, Hillsdale (1988)
15. Burton, R.: The Environmental Module of Intelligent Tutoring Systems. In: Polson, M.C., Richardson, J.J. (eds.) Foundations of Intelligent Tutoring System. Lawrence Erlbaum Associates, Hillsdale (1988)
16. Xu, L., Sarrafzadeh, A.: Haskell-Tutor: An Intelligent Tutoring System for Haskell Programming. Research Letters in the Information and Mathematical Sciences 6 (2004)
17. López, N., Núñez, M., Rodríguez, I., Rubio, F.: WHAT: Web-Based Haskell Adaptive Tutor. In: Scott, D. (ed.) AIMSA 2002. LNCS (LNAI), vol. 2443, pp. 71–80. Springer, Heidelberg (2002)
18. Melis, E., Andres, E., BÄundenbender, J., Frischauf, A., Goguadze, G., Libbrecht, P., Pollet, M., Ullrich, C.: Activemath: A generic and adaptive web-based learning environment. International Journal of Artificial Intelligence in Education 12 (2001)
19. Weber, G., Brusilovsky, P.: Elm-art: An adaptive versatile system for web-based instruction. International Journal of Artificial Intelligence in Education 12 (2001)
20. Carro, R.M., Pulido, E., Rodríguez, P.: Tangow: A model for internet based learning. International Journal of Continuing Engineering Education and Life-Long Learning, IJCEELL 11 (2001)
21. Kavcic, A., Privosnik, M., Marolt, M., Divjak, S.: Educational hypermedia system alice: an evaluation of adaptive features. Advances in Multimedia, Video and Signal Processing Systems, wseas 2002 (2002)
22. Perrenoud, P.: Building Skills from School. Dolmen Ediciones, Santiago de Chile (2005)
23. Ramsay, J.O.: Kernel smoothing approaches to nonparametric item characteristic curve estimation. In: Psychometrika, vol. 56, Springer, New York (1991)
24. Owen, R.J.: A bayesian approach to tailored testing (Research Report No. 69). Educational Testing Service, Princeton, NJ (1969)

A Majority Density Approach for Developing Testing and Diagnostic Systems

Dechawut Wanichsan, Patcharin Panjaburee,
Parames Laosinchai, and Sasithorn Chookaew

Institute for Innovative Learning, Mahidol University
999, Phuttamonthon 4 Road, Salaya, Nakorn Pathom 73170, Thailand
{kook260g,panjaburee_p,pl_one,nangfa1_edt}@hotmail.com

Abstract. Recently, testing and diagnostic learning systems have been considered as a useful tool for analyzing students' learning problems and giving helpful learning suggestions to them to improve their learning performance. Among the existing methods, a multi-expert approach has introduced a set of rules to integrate test item-concept relationship opinions given by multiple experts. However, when integrating the opinions from multiple experts, there are some problems that might effect on the quality of learning suggestions for students. Furthermore, it is time consuming to reconsidering their opinion when the conflict opinion exists. Therefore, a novel majority density approach is proposed to solve the mentioned problems. The experimental results show that this method can yield more reasonable integrated opinions than the previous approach and also reduce the number of reconsidering opinions.

Keywords: Concept-effect relationships model, Computer-based testing, Computer-assisted learning, Testing and diagnostic learning system.

1 Introduction

In usual testing systems, a student's learning achievement is reflected by his/her grade or a test score after doing a test. However, it is inadequate to know learning problems for improving the learning ability if some guidance is not provided by the system [1]. It implies that learning problem diagnosis after testing can give further suggestions is a significant issue for developing adaptive learning system. Bai and Chen proposed using fuzzy membership and fuzzy rules to discriminate an ability of students who get equal score [2], and using fuzzy rules in education grading system to reduce teachers' subjectivity for students' evaluation [3]. After testing, for separating students who might share similar misconceptions, a hierarchical clustering algorithm is used [4]. For testing and diagnostic learning systems, concept maps, presenting relationship among concepts which students have to learn, is considered as an important tool. In general, most testing and diagnostic systems have been designed and proposed to work in either one of the two environments: a system that automatically generated concept-relationship from students' testing results [5], and a system that a concept map is constructed in dedicated order before it is used for testing students [6-9].

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 134–143, 2011.

This paper focuses on the latter environment, Günel and Aşlıyan proposed statistical language model (SLM) to automatically extract concepts from Latex document [9], however, the output is only concepts, and there is no relationship among concepts to be shown. Hwang proposed a concept effect relationship model for developing intelligent tutoring system [6], however, there system supports only one expert for knowledge acquisition, and human errors is easily appeared. Therefore, based on the idea of the collaboration teachers, a multiple expert approach is proposed to determine the weightings for each test item to the specified concepts by integrating the opinions of diverse experts [7]. It helps experts to easily address all of the relationship between test items and the concepts because they could help each other to verify their knowledge; moreover, they could discuss with others to interchange their opinion when disagreement was happened.

However, there are some problems using their method for determining test item-concept relationships resulting in diagnosing learning problems because the approach considers only some opinions from experts, while the rest expert is omitted. Therefore, the unreliable and low quality integrated opinion might be occurred resulting in the cause of unreasonable results of personalized learning suggestions. Furthermore, since there are many cases that conflict opinion exists, it is time consuming to reconsidering their opinion.

To cope with these problems, in this study, a novel majority density method is presented to determine the relationship between test items and the concepts for diagnosing students' learning problems in testing and diagnostic systems. The proposed method can integrate the relationship values between test items and the concepts given by multiple experts by performing reasoning based on majority and density technique. It provides a useful and reasonable way for presetting the values for test items with respect to concepts for learning problem detection to provide more accurate personalized learning suggestions in testing and diagnostic learning systems.

The rest of this paper is organized as follows. Section 2 reviews some background and related work. Section 3 gives details of the proposed method, followed by an example of using our proposed work in Section 4. Section 5 concludes the work and provides future work.

2 Background

Panjaburee et al. [7] presented the first testing and diagnostic system for education that could work with multiple experts. To briefly describe the characteristic of a multiple-expert approach for testing and diagnostic learning system, an overview of the system is illustrated in Figure 1. In first step, the important thing that this system needs to construct is a predefined concept map, showing step of learning in proper sequence. Concept A, B, C and D are represented as four concepts in well-defined sequence, and the connecting arrows among concepts represent relationship showing learning order, i.e. Concept A, which students should firstly learn, precedes Concept B and C; in other words, Concept A is more fundamental than others.

In next step, the diagnostic test, comprising many test items, with multiple choices was designed. The role of experts is to give their opinion as a weight to represent level of relation between a test item and related concepts. The value ranging from "0"

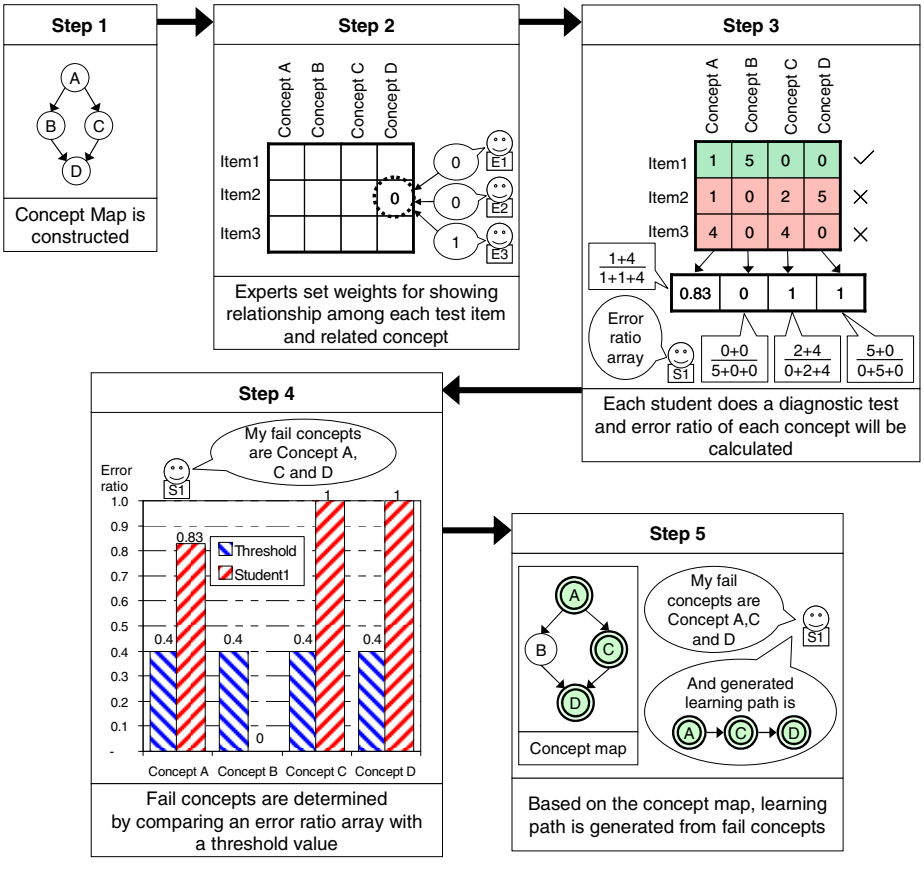


Fig. 1. An example of concept-effect relationship model

to “5” is used to show “No relationship”, “Weak relationship”, “Mildly weak relationship”, “Medium relationship”, “Mildly strong relationship” ,and “Strong relationship”. Moreover, the values ranging from “1” to “3” is defined as “Weak side opinion”, on the other hand, the values ranging from “3” to “5” is defined as “Strong side opinion”. In this case, expert #1 and #2 give “0” as a weight, and expert#3 gives “1” as a weight. Because each expert give different opinions as their weight, it is necessary to have a function used to integrate their opinions to only one value, and the integrated result is “0”. Next step, each student does a diagnostic test sheet, and the system will calculate error ratio, used to consider fail concept. In this case, student A fails to answer the question in item #2 and #3. For each concept, an error ration is calculated from summation of values of fail items divided by summation all values in that column. For example, concept A, summation of values of fail items is 5, calculated from 1+ 4, which are a value of item#2 and #3, and summation of all values of concept A is 6, calculated from 1+1+4. After obtaining error ratio array, fail concepts are determined by comparing an error ratio array with a defined threshold

value, assuming 0.4 in this case, in step 4. If a value in error ratio array exceeds a threshold value, a student will fail in that concept; for example, Student A fails to learn concept A, C and D. Finally, in step 5, the fail concepts will be analyzed by comparing to predefined concept map in order to generate learning path to student. In this case, the learning path is concept A → concept C → concept D.

The main function of existing work [7] is to propose an approach using a set of rules as a function to integrate weights, shown in Step 2 of Figure 1. Apart from a weight given by several experts, level of certainty, comprising "Sure" (S), and "Not sure" (N), is used to help experts to easily decide when they feel uncertain about their opinion. Fourteen rules were proposed using mathematical symbol, including ∀ (for all) and ∃ (for some). A vague point comes from using ∃ in almost all rules, consequently, some values of input weights from all dominate others for rule selection as shown in Table 1.

Table 1. An integrated weight generated from using fourteen rules of previous work

Cases	A weight from expert					An integrated weight (chosen rule)(output)
	E1	E2	E3	E4	E5	
1	**1(S)**	**4(N)**	4(S)	4(N)	4(S)	(R14)(1)
2	**1(S)**	4(S)	**4(S)**	4(S)	4(S)	(R11)(Reconsidering weight)
3	**2(N)**	3(S)	3(S)	**4(N)**	3(N)	(R12)(Reconsidering weight)
4	**0(S)**	3(S)	3(S)	**3(N)**	3(S)	(R9a)(0)
5	**1(S)**	1(N)	4(N)	4(N)	**5(S)**	(R11)(Reconsidering weight)
6	**0(S)**	4(S)	**5(S)**	5(S)	4(S)	(R7)(Reconsidering weight)
7	**0(N)**	1(S)	2(S)	2(S)	**5(S)**	(R8b)(5)

According to Table 1, it is a situation that there are five experts giving their weights in seven cases; the bold characters represent the case that has an impact to a chosen rule and an integrated weight. For instance, in case#1, there are four experts give "4" that is a strong-side weight, and only one expert gives "1" as a weak-side weight. Although almost all weights are strong side values, the integrated weight of rule#14 is "1" that is an unreasonable integrated weight because it lacks of majority consideration and only uses some values to create an output. Conversely, the integrated value should approximately be "4". For another example, in case #2, there are four experts give "4" as their weights with "S", and only one expert gives "1" with "S" as a weight. Although almost all weights are on strong side, all experts have to discuss to reconsider their value to correct disagreement. This case also lacks majority consideration because only opinion of expert1, "1" with "S", can dominate others to generate the output. Other cases also show an integrated that is not similar to all weights from majority except in case#5 that opinions are sparse in two sides.

For a testing and diagnostic system, an algorithm used to manage weights from experts is very important because different weight values directly affect to students' learning paths and suggestions in the future. A good quality of integrated weights produces accurate students' learning paths comparing with those are originated from a bad quality of integrated weight values.

3 Proposed Method

Due to several limitations of the existing method, this work aims to enhance the integrating method, and at the same time reduce the number of reconsidering weight values. This section will first describe scales we use to adjust a weight value with "Not Sure" to another, and then describe our technique for integrating weight values from multiple experts.

3.1 Preparing Data

In developing a testing and diagnostic system, it is difficult for an expert to make a decision to all of the relationships among the test items and the concepts. Therefore, to make them easily work, level of confidence that could be "S" or "N", where "S" represents "Sure" for giving a weight, and "N" represents "Not Sure", will be used together with weight relationship. However, a case that needs to be focused is that an expert gives an input weight with "N" because a weight "1" with "S" should not be the same value as a weight "1" with "N". Therefore, the weight value should be adjusted to another one. A basic idea to change the value is using constant value to multiply old weight value, but it is not good enough because some weights that are on strong or weak value side can be switched to opposite side after being multiplied; for instance, "4" (the strong-side value) multiply by "0.5" (constant value) is "2" (the new on weak side). Therefore, our idea to adjust all weights to a suitable one comprising two basic ideas as shown below:

1. After weight value is adjusted, the value will not be less or more than its adjacent value, e.g. "5" will not be decreased to be less than "4".
2. All weight values should be increased or decreased in direction of the center, average of all possible input weights. In this case, the value of center is "2.5", calculated from average[0, 1, 2 ,3 ,4 ,5].

Because of both ideas, all weights are adjusted to centralize in two aspects, that is center of all weights, and center of own and adjacent value as depicted in Figure 2 representing six possible input values and its adjusted value. There are two directions to change values, i.e. increase or decrease it; "2.5" is the value considered as the center of all adjusted weights. If an expert gives a strong-side weight with "N", the higher possibility that s/he has low confidence in the strong value, so such value should be decreased. For example, "5", a strong-side value, is given as a weight with "N"; such that the value "5" is decreased in direction of the center and the adjusted value is calculated from average of the value and adjacent weight value, ("5"+"4")/2 = "4.5". Clearly, if some experts give "5" as weight with "N", implying that the value should be decreased and the decreased value will not be less than "4". On the contrary, if an expert gives a weak-side weight with "N", it is a higher possibility that an expert think the value should be increased; e.g., "0" is given as a weight with "N", it implies that expert thinks it may have some relationship between a concept and an item. Therefore, the value is increased in direction of the center and calculated in the same way, ("0"+"1")/2 = "0.5" which is still less than "1". All weights with "Not Sure" are adjusted to others before being integrated in next step.

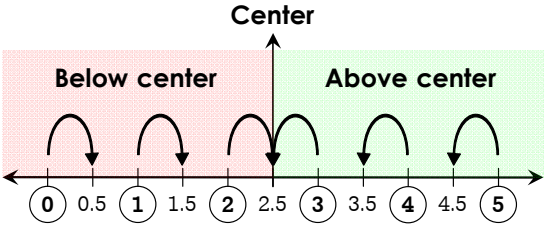


Fig. 2. A Scale for adjusting weights (only case "Not Sure")

3.2 Majority Density Algorithm

Because every member of data is used, the first idea for developing our algorithm to integrate weights to one value is to determine an average value by using an arithmetic mean. In mathematics and statistics, it is a basic and useful measure used to observe central tendency of data. It is calculated from the summation of the values divided by the number of values via formula.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

where $\bar{x}$ is arithmetic mean, x is an observation in i^{th}, n is the number of data. However, it is not enough robust measure because it is considerably influenced by outliers, observations that are distant from other values.

Consequently, for every test item and concept, it is necessary to remove an outlier before calculating an integrated weight from several weights of experts. Our approach

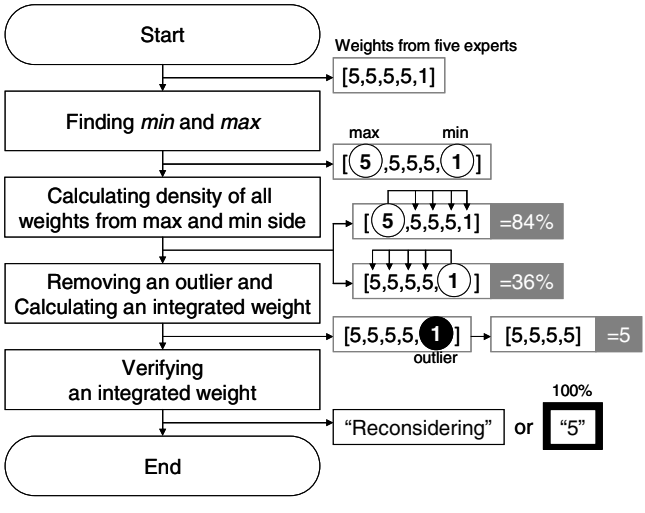


Fig. 3. A diagram describing four process steps of a majority density algorithm

consists of four steps: finding min and max, calculating density of all input weights from max and min side, calculating an integrated weight not including an outlier, and verifying whether the value should be used or all weights will be reconsidered. Figure 3 shows a diagram showing four steps of majority density algorithm and its output.

Finding Min and Max. After an original input of multi-expert testing and diagnostic system is obtained. A set of opinions will be used to find maximum and minimum value. In this case, there are five experts giving "5", "5", "5", "5", and "1" as their weight respectively. *maxWeight* is a variable showing the maximum weight value, i.e. "5", and *minWeight* is variable showing the minimum weight value, i.e. "1". They will be assumed as a center of weights.

Calculating Density from Max and Min Side. In this step, level of density from max side, *dnstMXS*, and min side, *dnstMNS*, will be calculated to observe distribution of opinions in order to define which one is an outlier.

$$dnstMXS = (1 - \frac{\left|\sum_{i=1}^{n} w_i - (maxWeight \times n)\right|}{n \times k}) \times 100\% \tag{2}$$

$$dnstMNS = (1 - \frac{\left|\sum_{i=1}^{n} w_i - (minWeight \times n)\right|}{n \times k}) \times 100\% \tag{3}$$

where *maxWeight* and *minWeight* are maximum and minimum weight obtained from previous step, w is a weight value, showing relationship of a test item and a concept, of i^{th} expert; n is the number of experts, where $1 <= i <= n$, k is the maximum rating scale (in this case, $k = 5$) From Figure 3, the density of weights from *maxWeight* and *minWeight* are 84% and 36% respectively, i.e. majority opinions are closer to *maxWeight* than to *minWeight*. The more value of both shows more level of density of weights.

Removing an Outlier and Calculating an Integrated Weight. The goal of this step is to remove an outlier by comparing values of *dnstMXS* and *dnstMNS*, and to calculate integrated value. There are two possible cases as:

a) *dnstMXS* = *dnstMNS*. There is no outlier because distribution of weights calculated from both sides are same, and temporary integrated weight, *tWeight*, is defined using arithmetic mean:

$$tWeight = \frac{\sum_{i=1}^{n} w_i}{n} \tag{4}$$

b) *dnstMXS* ≠ *dnstMNS*. An outlier is *maxWeight* if *dnstMXS* is less than *dnstMNS*, on the other hand, an outlier is *minWeight* if *dnstMXS* is more than *dnstMNS*. Temporary integrated weight, *tWeight*, is an average of all weights not including an outlier that is minority, and is defined as:

$$tWeight = \frac{\sum_{i=1}^{n} w_i - outlier}{n-1} \tag{5}$$

From Figure 3, "1" is considered as an outlier, and is removed. Consequently, an average of the rests is "5".

Verifying an Integrated Weight. *tWeight*, calculated in previous step, is only a temporary integrated weight and could not be used until the dense level of weight from majority is calculated to consider whether *tWeight* will be properly used as an integrated weight or all experts will be asked to discuss, check, and reconsider their weighing value. If there is no an outlier, density level of majority will be defined as Equation 6; otherwise, it will be defined as Equation 7:

$$dnstMJR = (1 - \frac{\sum_{i=1}^{n} |w_i - tWeight|}{n \times k}) \times 100\% \tag{6}$$

$$dnstMJR = (1 - \frac{\sum_{i=1}^{n} |w_i - tWeight| - |outlier - tWeight|}{(n-1) \times k}) \times 100\% \tag{7}$$

All experts will reconsider their weights if *dnstMJR* is less than 80%, in other words, all weights from experts are so sparse because majority opinion is in different way. On the other hand, *tWeight* is properly used as an integrated weight if *dnstMJR* is more than 80% because majority opinion is in similar way. From Figure 3, because all weights are same value, density level of majority, *dnstMJR,* is 100%". Consequently, "5" is proper to be used as an integrated weight.

4 An Example

To observe effectiveness of our proposed system, we use several input cases that existing method cannot properly handle as shown in Table 1. Table 2 shows the integrated weights, comparing the previous approach with our proposed method.

According to Table 2, there are seven cases, and there are five experts giving their individual weights. First step, to prepare a weight value, a weight value with "Not sure" will be increased or decreased value based on the scale in Figure 2; for example, in case#1, "4" with "Not sure" of expert 2 will be adjusted to "3.5" as same as that of expert 4.

Then, an outlier will be detected and an integrated weight will be calculated; the bold characters represent an outlier detected by our method. In case#1, assuming that "1" is an opinion from expert#1 owing to misconception; our proposed algorithm can detect "1" as outlier, after that it was removed from others; such that the integrated weight would result from the average of the rest weight values, (i.e., 3.5+4+3.5+4)/4 = "3.75" that are closer to opinions of majority comparing with the result of previous work. It implies that our approach could handle the case that an opinion is distant from others.

Table 2. The comparison of results of integrating weight between our proposed method and previous work

Cases	A weight from an expert					An integrated weight	
	E1	**E2**	**E3**	**E4**	**E5**	**Previous work**	**Proposed method**
1	**1(S)**	4(N) =3.5	4(S)	4(N) =3.5	4(S)	1	3.75
2	**1(S)**	4(S)	4(S)	4(S)	4(S)	Reconsidering weight	4
3	2(N) =2.5	3(S)	4(N) =3.5	4(N) =3.5	3(N) =2.5	Reconsidering weight	3
4	**0(S)**	3(S)	3(S)	3(N) =2.5	3(S)	0	2.88
5	**1(S)**	1(N) =1.5	4(N) =3.5	4(N) =3.5	5(S)	Reconsidering weight	Reconsidering weight
6	**0(S)**	4(S)	5(S)	5(S)	4(S)	Reconsidering weight	4.5
7	0(N) =0.5	1(S)	2(S)	2(S)	**5(S)**	5	1.38

Another interesting case is case#3 that distribution of all weights is on medium level ranging from 2 to 4. There is no an outlier and our proposed provides "3" as an integrated weight, comparing with the result of previous approach, i.e. "Reconsidering weight" generated from Rule11. That rule needs only one value on weak side, and one value from strong side to generate "reconsidering weight", however, it is unreasonable because all opinions are almost on same area. This case shows usefulness of our approach using density as one main part to generate reasonable integrated weights. It implies that our approach could handle the case that all weights are in middle-value dense area.

In case#5, an integrated result of our method is "Reconsidering weight" although an outlier was removed. However, it is acceptable because all weights are not in dense area. In other words, all weights are sparse in wide scale. Moreover, dense level, used to verify an integrated weight, is 77.50% that is less than 80%. It implies that our approach could reasonably decide to set "Reconsidering weight".

From an integrated weight of seven cases in Table 2, Results of the proposed method are different to those generated from using previous approach in almost all cases except case#5. Interestingly, in our method, the dense level of weights is calculated, an outlier is also removed, and good-quality integrated weights are calculated from the average of remaining weights; such that, it might generate more accurate learning paths to students.

5 Conclusion

This work proposes a majority density approach for integrating experts' opinions by considering a dense area of majority opinions, and using all weighting values before integrating them. Comparing with the existing works, the results validate the utility of our approach, as it could reasonably integrate opinions of multiple experts in order

that the system might generate more accurate learning paths and suggestions to students, and could reduce several cases that all experts have to replace their opinion with the other values. However, although our new approach reveals good results, it has some limitations, i.e. it can properly work when the number of experts is more than three experts. As a future work, to make a practical system for students, this approach is going to be conducted to demonstrate the effectiveness on a Mathematic and Biology course.

References

1. Gerber, M., Grund, S., Grote, G.: Distributed collaboration activities in a blended learning scenario and the effects on learning performance. Journal of Computer Assisted Learning 24, 232–244 (2008)
2. Bai, S.-M., Chen, S.-M.: Evaluating students' learning achievement using fuzzy membership functions and fuzzy rules. Expert Systems with Applications 34, 399–410 (2008)
3. Bai, S.-M., Chen, S.-M.: Automatically constructing grade membership functions of fuzzy rules for students' evaluation. Expert Systems with Applications 35, 399–410 (2008)
4. Fernández, A., Gómez, S.: Solving Non-Uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms. Journal of Classification 25, 43–65 (2008)
5. Bai, S.-M., Chen, S.-M.: Automatically constructing concept maps based on fuzzy rules for adapting learning systems. Expert Systems with Applications 35, 41–49 (2008)
6. Hwang, G.-J.: A conceptual map model for developing intelligent tutoring systems. Computers & Education 40, 217–235 (2003)
7. Panjaburee, P., Hwang, G.-J., Shih, B.-Y.: A multi-expert approach for developing testing and diagnostic systems based on the concept-effect model. Computers & Education 55, 527–540 (2010)
8. Chen, S.-M., Bai, S.-M.: Learning barriers diagnosis based on fuzzy rules for adaptive learning systems. Expert Systems with Applications 36, 11211–11220 (2009)
9. Günel, K., Aşlıyan, R.: Extracting learning concepts from educational texts in intelligent tutoring systems automatically. Expert Systems with Applications 37, 5017–5022 (2010)

Exploiting Learners' Tendencies for Detecting English Determiner Errors

Ryo Nagata and Atsuo Kawai

Konan University and Mie University

Abstract. This paper proposes a method for detecting determiner errors, which are highly frequent in learner English. To augment conventional methods, the proposed method exploits a strong tendency displayed by learners in determiner usage, i.e., mistakenly omitting determiners most of the time. Its basic idea is simple and applicable to almost any conventional method. This paper combines this idea with countability prediction, which outperforms the conventional methods, achieving an F-measure of 0.613.

Keywords: Determiner errors, essay writing, learners of English, error detection, language learning.

1 Introduction

Determiner usage is one of the major difficulties that non-native speakers of English are faced with in English writing. This is especially true for those whose mother tongue does not have a determiner system similar to that of English (e.g., Chinese and Japanese). It can easily be observed, among other errors, in an essay written by a Japanese learner of English:

> *I became univercity student, I get up early every morning. I go to the school when I listening to music in train. Stady is very different. Especiary I think that programing and math doesn't know.*

The underlines indicate the noun phrases (NPs) that have a determiner error. Because of the difficulty inherent in using determiners, errors in determiners, including article errors, are one of the most frequent grammatical error types in learner English [5,7].

Determiner errors are so frequent that they become problematic in several circumstances. For example, teachers have to identify and correct determiner errors in writing classrooms, which is time-consuming and costly. Similarly, raters have to identify a great number of determiner errors to evaluate writing skills in grammar in writing tests.

Given these circumstances, researchers [5,8,10,12] have done a great deal of work on determiner-error detection, mostly focusing on article errors. Most of this work, including [5,12], has attempted to solve article-error detection as a classification problem by using machine learning algorithms or statistical methods — three-way classification into the indefinite article *a*, the definite article

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 144–153, 2011.

the, and no article ϕ[1]. For instance, Han et al. [5] propose a maximum entropy (ME) classifier-based method for predicting correct articles; if the prediction disagrees with the one actually used, then it is detected as an error. The features are based on lexical and syntactic information around the article in question. They report that their method achieves a recall of 0.40 with a precision of 0.90. It should be noted that these article-error detection methods can naturally be extended to determiner-error detection. One can build an n-way classifier where n corresponds to the number of target determiners. The classifier selects the correct one out of the target determiners in determiner-error detection.

As an alternative approach, Nagata et al. [18] propose using countability prediction[2]. Their method first predicts the countability of the head noun from its surrounding context and then applies some rules to the prediction to examine whether the determiner that modifies the head noun is correct or not.

Although performance has improved, one important factor has been missing in the previous methods. They do not take into consideration that the target texts are written by learners of English. Learners of English have a strong tendency to mistakenly omit determiner most of the time, and this tendency has crucial implications for determiner-error detection. Methods designed to detect determiner errors in learner English should thus exploit this tendency in order to achieve better performance.

Accordingly, this paper takes the first step in exploiting this learner tendency to achieve better performance. The basic idea is very simple and applicable to almost any classification-based method. This paper attempts to combine this idea with the countability prediction-based method, which outperforms the conventional methods.

The rest of this paper is structured as follows. Section 2 introduces the basic idea of the proposed method. Section 3 discusses the proposed methods in detail. Section 4 describes the experiments conducted to evaluate the proposed methods. Section 5 discusses the experimental results.

2 Basic Idea

Let us start by analyzing determiner usage in learner English, which leads us to the basic idea. Izumi et al. [7] report that articles are frequently missing at all proficiency levels (novice, intermediate, and advanced) in the NICT JLE corpus. The frequency approximately ranges from 60% to 80% depending on the proficiency levels. Han et al. [5] expand this into the writing of Chinese, Japanese, and Russian learners of English, focusing on articles errors out of determiner errors. They show that missing and extraneous articles account for around 80% of all article errors. These analyses suggest that learners (at least

[1] Hereafter, *a* refers to both *a* and *an* unless otherwise noted. Also, ϕ refers to no article.

[2] Countability is highly related to determiner usage [1,6]. For example, non-count nouns do not take the indefinite article whereas singular count nouns do not appear without a determiner.

Chinese, Japanese, and Russian) have a strong tendency to often mistakenly omit and extraneously use determiners. A possible reason is that these languages do not have a determiner system similar to that of English. Considering this, other learners whose mother tongue is similar are likely to have the same tendency.

Surprisingly, no one has exploited this tendency in efforts to detect determiner errors in learner English. This paper takes the first step in exploiting it to achieve better performance.

In consideration of this tendency, predicting whether the NP in question requires a determiner or not is sufficient for detecting the vast majority of determiner errors. This leads to the proposed method's basic idea of solving determiner-error detection as a binary classification problem instead of the conventional n-way classification. In other words, it is classification into two categories: (i) YES — the NP in question requires a determiner and (ii) NO — the opposite. If the NP in question is classified into YES and it has no determiner, then an error is detected; likewise if NO and it has a determiner, then an error.

One could argue that the basic idea does not determine which determiner to use. However, this should not be so problematic in terms of feedback in language teaching and learning. Robb et al. [19] show that only indicating where errors exist has good effects in second language learning and that it is a better strategy than providing the correct forms. Several researchers, including Ferris et al. [4] and Lee [11], report similar results. Also, Chodorow et al. [3] and Nagata et al. [17] recently have shown that error detection systems that only indicate where errors exist improve writing skills. Considering these findings, the binary classification setting is expected to be sufficiently meaningful in language learning and teaching.

3 Proposed Method

3.1 Generating Training Data

The source of training data is a raw corpus. As preprocessing, the corpus is tagged with part-of-speech (POS) and chunked using existing tools.

With this preparation, training data are automatically generated as follows. Each NP headed by a noun in the tagged corpus corresponds to a training instance. It is labeled as either YES or NO relying on the POS and chunk tags; it is labeled as YES if it has a determiner, otherwise NO. For instance, the NPs in the sentence "[NP The/DT young student/NN] was on [NP time/NN] ./." would be labeled as "[NP:YES The/DT young student/NN] was on [NP:NO time/NN] ./." These labeled NPs are training instances in the proposed method.

3.2 Detecting Determiner Errors

The proposed method detects determiner errors in three steps. First, a classifier is trained on the training instances. Then, the classifier is applied to each NP in the target text to predict whether the NP requires a determiner or not. Third, errors are detected based on the prediction. If the prediction is YES and the NP

has no determiner, then it is detected as an error; likewise if NO and the NP has a determiner, then an error.

Almost any classification algorithm can be used as the classifier given the training data. In the proposed method, a decision list (DL)-based classifier [18,20] is selected for this purpose. One of the reasons for its use is that it is robust against noise in learner English [16]. Here, *noise* refers to errors other than determiner errors. A wide variety of errors appear in learner English as shown in the example essay in Sect. 1. The noise affects the performance of error detection methods. Thus, it is preferable that error detection methods be robust against noise. Another reason is that there is a natural way of combining the proposed method (classification based-method) with the countability prediction-based method [18] that also uses the DL-based classifier to predict countability.

A DL consists of a set of rules that have the template: *If a condition is true, then make a decision*. To formalize the template in the proposed method, we use a variable L that takes either YES or NO to denote whether the NP in question requires a determiner or not, respectively. We also use w and C to denote a word and a certain context around the head noun, respectively. We define three types of C: (i) np: the NP in question, (ii) $-k$: k words to the left of the NP, and (iii) $+k$: k words to its right. Then the template is formalized by: If w appears in C, then L. Hereafter, to keep the notation simple, it will be abbreviated to $w_C \rightarrow L$. The value of k is set to 3 following Nagata et al. [18]'s work.

Now rules for a DL can be obtained from the training data. All that is needed is to collect words in C from the training data; the following words are excluded as stop words: determiners, personal pronouns, auxiliary verbs, and the head noun itself[3]. All words are reduced to their morphological stem and converted entirely to lower case when collected. In addition, a default rule is defined. It is based on the head noun itself and used when no other applicable rules are found in the DL. It is defined by $h \rightarrow L_{\text{major}}$, where h and L_{major} denote the head noun and the majority of L for the head noun in the training data, respectively. It reads, "If the head noun appears, then predict the label by the majority."

The log-likelihood ratio (LLR) [20] decides in which order rules are applied in prediction. It is defined by $log\frac{p(L|w_C)}{p(\overline{L}|w_C)}$ where $\overline{L}$ is the exclusive event of L and $p(L|w_C)$ is the probability that the label is L when w appears in the context C. The probability $p(L|w_C)$ is estimated by $p(L|w_C) = \frac{f(w_C,L)+\alpha}{f(w_C)+2\alpha}$ where $f(w_C)$ and $f(w_C, L)$ are occurrences of w appearing in C and those in C when the label is L, respectively. In this paper, α is set to 1.0 following Nagata et al. [18]'s work. Rules in a DL are sorted in descending order by LLR. They are tested on the target NP in this order.

3.3 Combining with Countability Prediction

Combining a classification-based method with the countability prediction-based method is an attractive idea because they have different properties in

[3] Note that the information on the head noun is considered in that a DL is learned for each target noun.

determiner-error detection. Countability is highly related to determiner usage and thus should be considered in classification-based methods. At the same time, the countability-based method is not capable of detecting a certain class of determiner errors which classification-based methods are. For example, whether or not the definite article is required cannot be determined by countability.

Before further discussion, let us define two names to avoid confusion between the classification-based method proposed in Subsect.3.2 and the one combined with countability prediction. Hereafter, the former and latter will be referred to as the binary classification-based method and the mixture method, respectively.

Now let us describe the mixture method. Fortunately, there is a natural way of combining the binary classification-based method with the countability prediction-based method. Both use the same type of classifier (i.e., DL-based classifier). Thus, rules in both DLs for a noun are simply merged into one and are sorted in descending order by LLR. Then, rules are applied to the head noun in this order. If the applied rule is from the binary classification-based method, the detection procedure is the same as in the binary classification-based method. If it is from the countability prediction-based method, the detection is done based on the predicted countability. For example, the indefinite article modifying non-count noun can be detected as an extraneous article (see Nagata et al. [18] for the details).

4 Experiments

4.1 Experimental Conditions and Procedures

We collected 119 essays (10,261 words) written by Japanese junior high and high school students for the experiments. The topic was either *My family*, *Memories of junior high school*, or *Future dreams*. Two native speakers of English, who were teachers of English, separately annotated grammatical errors in the target essays, including determiner errors. Another native speaker of English, who was also a teacher of English, double-checked the annotations; if one disagreed with the other, the third determined which to use. As a result, 240 determiner errors were identified in the essays (missing: 75.0%; extraneous: 15.4%; selection: 9.6%).

Recall and precision were used to evaluate the performance. Also, F-measure was used which evaluates the performance considering both recall and precision.

For comparison, we implemented the countability prediction-based method [18] and the ME classifier-based method [5]. In addition, we extended the ME classifier-based method to n-way classification to investigate how well n-way classification performs on determiner-error detection. We selected 18 target determiners (i.e., 18-way classifier)[4]. Also, we implemented a baseline where the prediction was made by the majority of YES/NO for each head noun. Each method was trained on the following corpora that were analyzed by the OAK system[5]; (i) written part

[4] The target determiners are: *a*, *all*, *each*, *every*, *her*, *his*, *its*, *my*, *our*, ϕ, *that*, *the*, *their*, *these*, *those*, *this*, *what*, *your*.

[5] OAK System Homepage: http://nlp.cs.nyu.edu/oak/

Table 1. Experimental results

Method	R	P	F
Mixture	0.717	0.536	**0.613**
Binary classification	0.771	0.440	0.561
Countability Prediction	0.646	**0.550**	0.594
ME	0.583	0.471	0.521
ME (n-way)	0.638	0.274	0.383
Baseline	**0.779**	0.399	0.526

of the BNC corpus [2] (80 million words), (ii) the English concept explication in the EDR English-Japanese Bilingual dictionary and the EDR corpus [13] (3 million words), and (iii) educational materials for Japanese learners of English (180 thousand words).

In the ME classifier-based methods, it was impossible to use all training instances obtained from the corpora because of a computational problem[6]. To solve this problem, training instances were sampled from the corpora. All training instances generated from the corpus consisting of the educational materials were included in the training data because they were designed for Japanese learners of English. As for the remaining two corpora, all training instances whose head noun appeared in the target essays were also included in the training data, considering the fact that head nouns are the most crucial in the ME classifier-based methods [5]. From the rest, 20% of training instances were randomly sampled out. In total, the number of training instances was approximately six million and seven million in the ME classifier-based method and the n-way ME classifier-based method, respectively. Note that the use of all NP instances would improve their performance very little according to the experimental results reported by Han et al. [5][7].

4.2 Experimental Results

Table 1 shows the experimental results. "Mixture" and "Binary classification" refer to the two proposed methods. "Countability Prediction," "ME," and "ME n-way" refer to the countability prediction-based method, the ME classifier-based method, and the n-way ME classifier-based method," respectively.

Table 1 reveals that the binary classification-based method is recall-oriented while the countability prediction-based method is precision-oriented. The binary classification-based method significantly improves, achieving the best F-measure

[6] All ME classifiers were generated by the opennlp.maxent package (http://maxent.sourceforge.net/) on a Linux server (Quad core AMD Opteron 2.3 GHz, 4GMB memory). It was impossible to generate an ME classifier using all training instances, which amounted to approximately 21 million instances, due to the lack of memory.

[7] They report that performance (accuracy) improves from 87.92% to 87.99% when the number of training instances increases from 4.8 million to 6 million.

of 0.613, when it is combined with countability prediction; F-measure of the mixture method is significantly better than the others at $p < 0.01$ (approximate randomization test) except that of the countability prediction-based method where the difference is significant at $p < 0.15$.

5 Discussion

5.1 Comparison with Previous Methods

The experimental results show that the binary setting is effective in determiner-error detection. It seems too difficult to predict correct determiners or even correct articles from the local context according to the performances of the ME classifier-based methods. By contrast, the proposed methods focus on predicting whether the NP in question requires a determiner or not, which is useful in language learning and teaching as explained in Sect. 2.

The fact that the target essays contain noise (i.e, spelling and grammatical errors) makes it more difficult to predict correct determiners. The ME classifier-based methods use all contextual features available surrounding the determiner in question. It is highly possible that noise in the context affect their performances. Especially, spelling errors are problematic for them. Actually, Han et al. [5] excluded, from their experiments, instances whose feature contained a spelling error, whereas the experiments in this work include such instances. By contrast, the proposed methods tend to ignore noise because the proposed methods (DLs) make a prediction solely relying on the most relevant one out of the surrounding features. Simply, they ignore features containing noise because it normally do not appear in the training data.

Compared to the countability prediction-based method, it turns out that the binary classification based-method can detect determiner errors that the countability prediction-based method often fails to detect. The countability prediction-based method tends to overlook missing definite articles. This is often the case when the head noun is one of the nouns that almost always take the definite article such as *the earth* and *the sun*. In the countability prediction-based method, instances modified by the definite article are discarded from the training data because their countability is not determined by the rules used for generating the training data. Because of this, the countability prediction-based method fails to detect errors in such cases. By contrast, the binary classification-based method successfully tells that the NPs require a determiner. This explains why the recall of the binary classification-based method is much higher than that of the countability prediction-based method.

The binary classification-based method significantly improves when it is combined with the countability prediction-based method. Namely, the mixture method achieves the best performance in terms of F-measure. Intuitively, the mixture method adaptively selects a better rule from the binary classification-based method and the countability prediction-based method according to LLR. Considering this, a more sophisticated method for selecting rules may improve the performance further.

5.2 False Negatives and False Positives

The most frequent cause of the false negatives was selection-type errors (mostly *a/the* confusion), accounting for 32.4% of all the 68 false negatives. As already explained, the mixture method is incapable of detecting selection-type errors. Thus, one needs some other techniques to detect this type of error. Extraneous definite articles come next (23.5%). The mixture method becomes incapable of detecting this type of errors when it uses rules from the countability prediction-based method. More importantly, extraneous definite articles are often hard to detect for all detection methods due to instances of generic reference [1]. To see this, let us consider the following example: "*I can see the two dogs. I hate* _ *dogs.*" Here, both *the* and ϕ can be put in the underline, the latter being a generic reference. To determine which is correct requires full knowledge of the context. Obviously, none of the methods in the experiments take this into consideration. A similar reason explains false negatives of which definite article is missing (23.5%). It often requires discourse and/or extra-textual information to determine whether the definite article should be used or not [5]. On the other hand, the usage of the indefinite article seems to be less problematic for the mixture method. Missing and extraneous indefinite articles only occupy 4.4% and 2.9% of the false negatives, respectively. POS-tagging or chunking errors also seem not to affect the performance so much (4.4%).

A major cause of the false positives is due to errors in number (38.9%). The mixture method (and also the other methods) in the experiments often mistakenly detected errors in number as determiner errors. For instance, it judged *beatiful place* in *I see beatiful place* as a missing determiner error where the error was actually an error in number (i.e., *beatiful palces*). To distinguish between errors in article and number often requires discourse knowledge. Another major cause is POS-tagging or chunking errors (30.2%). For instance, the word *kind* used as an adjective was often mistakenly POS-tagged as a noun as in *She is kind/NN.* In that case, the detection often resulted in a false positive because the word *kind* usually appears with a determiner (e.g., *a kind of*) in the corpora. Another cause is grammatical errors (8.0%). If words that are informative for determiner usage have grammatical errors, the mixture method tends to make a false positive.

5.3 Relation to Previous Work

Correlated with determiner-error detection is article generation. Researchers [15] first explored article generation methods based on hand-crafted rules, mostly aiming at machine translation. However, these methods are not directly applicable to determiner-error detection because they use the knowledge in the source language, which is not available in essay writing. Knight and Chander [9] took the first step to using a machine learning algorithm for article generation although their method only deals with *a/the* selection. Minnen et al. [14] extend this work to three-way classification. However, their method also depends on information such as functional tags in Penn Treebank which may not be reliable in essay writing.

As mentioned in Sect. 1, there has been a great deal of work on determiner-error detection although none of it explicitly exploits the learners' tendency. Only Han et al. [5] compared their three-way ME classifier-based method with its variant whose classification was between *a*/*the* and ϕ, and it achieved a slight improvement. It should be emphasized that the motivation for the use of the binary classification was that they thought the three-way classification was difficult for their ME classifier rather than they tried to exploit the learners' tendency. More importantly, their binary classification setting is essentially different from ours; ours is whether the NP in question requires a determiner or not.

6 Conclusions

This paper proposed a method for detecting determiner errors, which is one of the most frequent grammatical errors in learner English. This paper took the first step in exploiting the tendency of learners in English writing to often mistakenly omit and use determiners extraneously to achieve better performance. This paper attempted to combine this idea with the countability prediction-based method, which outperformed the conventional methods. The experimental results show that (i) it is difficult to predict correct determiners or even correct articles from the words surrounding the determiner in question especially when the target text is noisy (i.e., spelling and grammatical errors) and (ii) combining the strong tendency with countability prediction is effective in determiner-error detection.

In the future work, we will investigate a more sophisticated method for combining classification-based methods with the countability prediction-based method. We will also investigate how the proposed methods benefits language learning.

References

1. Bond, F.: Translating the Untranslatable. CSLI publications, Stanford (2005)
2. Burnard, L.: Users Reference Guide for the British National Corpus, version 1.0. Oxford University Computing Services, Oxford (1995)
3. Chodorow, M., Gamon, M., Tetreault, J.R.: The utility of article and preposition error correction systems for English language learners: feedback and assessment. Language Testing 27(3), 419–436 (2010)
4. Ferris, D., Roberts, B.: Error feedback in L2 writing classes: How explicit does it need to be? Journal of Second Language Writing 10(3), 161–184 (2001)
5. Han, N.R., Chodorow, M., Leacock, C.: Detecting errors in English article usage by non-native speakers. Natural Language Engineering 12(2), 115–129 (2006)
6. Huddleston, R., Pullum, G.K.: The Cambridge Grammar of the English Language. Cambridge University Press, Cambridge (2006)
7. Izumi, E., Saiga, T., Supnithi, T., Uchimoto, K., Isahara, H.: The development of the spoken corpus of Japanese learner English and the applications in collaboration with NLP techniques. In: Proc. of the Corpus Linguistics 2003 Conference, pp. 359–366 (2003)

8. Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T., Isahara, H.: Automatic error detection in the Japanese learners' English spoken data. In: Proc. of 41st Annual Meeting of ACL, pp. 145–148 (2003)
9. Knight, K., Chander, I.: Automated postediting of documents. In: Proc. of 12th National Conference on Artificial Intelligence, pp. 779–784 (1994)
10. Lapata, M., Keller, F.: Web-based models for natural language processing. ACM Transactions on Speech and Language Processing 2(1), 1–31 (2005)
11. Lee, I.: ESL learners' performance in error correction in writing: Some implications for teaching. System 25(4), 465–477 (1997)
12. Lee, J.: Automatic article restoration. In: Proc. of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pp. 31–36 (2004)
13. electronic dictionary research institute ltd, J.: EDR electronic dictionary specifications guide. Japan electronic dictionary research institute ltd. (1993)
14. Minnen, G., Bond, F., Copestake, A.: Memory-based learning for article generation. In: Proc. of 2nd Workshop on Learning Language in Logic and 4th Conference on Computational Natural Language Learning, pp. 43–48 (2000)
15. Murata, M., Nagao, M.: Determination of referential property and number of nouns in Japanese sentences for machine translation into English. In: Proc. of 5th International Conference on Theoretical and Methodological Issues in Machine Translation, pp. 218–225 (1993)
16. Nagata, R., Kawai, A., Morihiro, K., Isu, N.: A feedback-augmented method for detecting errors in the writing of learners of English. In: Proc. of 44th Annual Meeting of ACL, pp. 241–248 (2006)
17. Nagata, R., Nakatani, K.: Evaluating performance of grammatical error detection to maximize learning effect. In: Proc. of 23rd International Conference on Computational Linguistics, poster volume, pp. 894–900 (2010)
18. Nagata, R., Wakana, T., Masui, F., Kawai, A., Isu, N.: Detecting article errors based on the mass count distinction. In: Proc. of 2nd International Joint Conference on Natural Language Processing, pp. 815–826 (2005)
19. Robb, T., Ross, S., Shortreed, I.: Salience of feedback on error and its effect on EFL writing quality. Tesol Quartery 20(1), 83–93 (1986)
20. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: Proc. of 33rd Annual Meeting of ACL, pp. 189–196 (1995)

Analysis of Students' Learning Activities through Quantifying Time-Series Comments

Kazumasa Goda[1] and Tsunenori Mine[2]

[1] Kyushu Institute of Information Sciences, Saifu 6-3-1,
811-0117 Dazaifu, Japan
gouda@kiis.ac.jp
[2] Kyushu University, Motooka 744,
819-0395 Fukuoka, Japan
mine@ait.kyushu-u.ac.jp

Abstract. These days, many university teachers are concerned about the increasing number of students whose motivation is declining. Some of them fall into a situation that they cannot recover from by themselves, and require assistance, but they hesitate to call for help. In order to recognize such students quickly and give guidance to them in class, we have collected time-series comments in the classroom and analyzed them. In the analysis, we divided the comments into the three time slots: P (Previous), C (Current), and N (Next), and quantify them so that we can infer the learning behaviors between the previous and the current classes. We call this analysis method the PCN method. The PCN method is useful for grasping students' learning status in the class. Some of our case studies illustrate the validity of the PCN method.

Keywords: Learning activities, user model, time-series comment analysis, classroom improvement.

1 Introduction

In classrooms, teachers not only teach students, but also try to receive any information on their learning activities through careful observation of them. Most teachers empirically and roughly recognize their students' learning status by means of observations; they also know that giving feedbacks to the class is a very good way to improve students' learning attitudes. Only some teachers, however, attempt to do so because it is hard work to grasp all the class members' learning attitudes all over the periods in the semester; they pick up some cases according to their needs mainly based on their experience in the class.

This paper proposes a method of grasping all the class members' learning attitudes by analysis of their freestyle comments collected in the class. The method gives the procedure for analyzing and quantifying the comments and shows the evaluation viewpoints of the quantification. We call the method the PCN method. The PCN is the abbreviation of Previous, Current, and Next. The PCN method enables teachers to acquire a temporal learning status of each student as a form of triple (**P, C, N**): **P** (Previous) indicates the learning activity before the class time such as review of

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 154–164, 2011.

previous class and preparation for the coming class, **C** (Current) shows the understanding and achievements of class subjects during the class time, and **N** (Next) tells the learning activity plan until the next class.

The rest of the paper is organized as follows; Section 2 discusses related work and makes the difference from this work clear; Section 3 describes the PCN method; Section 4 discusses the case study; finally Section 5 concludes the paper and describes our future work.

2 Related Work

There exists a lot of work related to the topics touched on in this paper, such as adaptive learning, text mining of time-series data and so forth.

Here we discuss some work on adaptive learning. From behaviorism, PSI (Personal System of Instruction) is one of teaching methods, person to person education well-known for Keller Plan, which was proposed by Keller in 1960s (Keller 1968). Proctors use the predetermined set of objects for each student. Proctors play an important role in PSI and they should work very hard to grasp learning status of all the members in the class and manage the progress of the class, quality of which depends on their experiences. Since training proctors costs expensive and takes long time, PSI is applied to limited students requiring special aid. The PCN method provides data expressing learning status as values of P, C, and N. CSCL (Computer Supported Collaborative Learning) is a pedagogical research area on learning environment derived from CSCW (Computer Supported Cooperative Work) (Koschmann 1996). It provides the learning environment for collaborative learning across classes, schools, sometimes countries by computer connected to the internet. This breaks special barrier and students are located so wide in such environment that teachers encounter the difficulties in grasping learning status of all the students or even students in charge. The PCN method provides indexes expressing learning status of students and basic idea for a component of learning system supporting CSCL. Self-regulated learning is a learning style guided by metacognition (Zimmerman 1990). It is characterized three points, self-observation, self-judgment, and self-reactions. The PCN method provides indexes reducing the task for all of self-observation, self-judgment, and self-reaction. ID (Instructional Design) is the practice of maximizing the effectiveness of learning rooted in cognitive and behavioral psychology (Gagne 1965, Ito & Suzuki 2008), and there are many instructional design models but many of them are based on the ADDIE model with the five phases: analysis, design, development, implementation, and evaluation. The analysis process of ID needs the current learning status of the class. And the PCN can provide it. There exist so many user models concerning adaptive media systems (Brusilovsky 2001, Popescu et al. 2007) and they are roughly classified into three categories: the user model, the domain model, and the interaction model (Martins 2008). The PCN method helps the interaction model in inferring students' characters partly by PCN values.

Free style comments and essays are popular targets of text mining researches (Ishioka 2006, Burstein 2003). With respect to the content of the comments, most analyses of time-series comments are done for marketing such as CRM (customer

relationship management). Such comments include reputations, opinions, or requests expressing their preferences or characters. Our target domains are education, and the comments we use are gathered from students, and reflect their learning activities directly or indirectly. However, only a few researches of text mining using learning data (Romero 2007) have been done due to small numbers of data concerning learning status in time series. Previous work only analyzes time-series comments literary without concerning teachers' experiences. In this research, we periodically collected free style comments handwritten by students in the class. Our method analyses students' comments primarily literally, and secondly adjusts the results to more precise values with teachers' experience or memories. Since the comments include students' names, we can also track their learning activities easily. We have students' freestyle feedbacks as raw data relating to each writer by name, two classes, about 60 students for each class, about 14 periods of class period set.

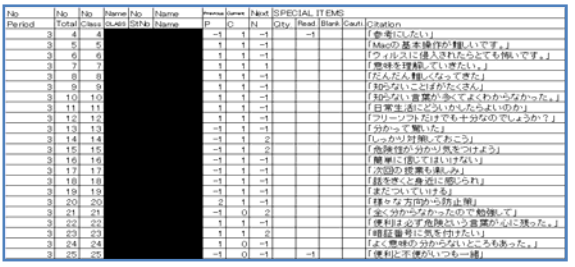
About data sources, learning activity analysis is mostly external to students such as server logs and teachers' observation. Our approach uses comments internal to students originally, or they actually write the comments. Most adaptive learning use recorded data such as server logs gathered automatically and almost unconscious of students related. We use recorded data but are consciously written by students because they are guided, told that the writings will be used for part of their score. Moreover, some students are sometime directly pointed and warned if the writings are not enough and lower level than the expected by teachers.

3 The PCN Method

3.1 Overview of the PCN Method

The PCN method provides the procedure to quantify students' freestyle comments and derives useful data to recognize their learning status. In this method, first, teachers read class comments written in natural language with free-style, and analyze them according to three time-series viewpoints: Previous, Current, and Next. The teachers evaluate the analyzed comments, convert them numerically, and record the result of the conversion. In numerical conversion, one value of (-1, 0, 1, 2) is given where the value has a higher grade as it becomes greater. So, we call the value as **Attention (-1), Bad (0), Fair (1), and Good (2)**, respectively.

Fig. 1 shows a working sheet for quantifying the comments to triple (P, C, N), and special items. The sheet also contains phrases that directly express learning status of students or show notes concerning to the students. As need arises, the teachers, further to PCN, can also record special items such as the things that should be told to all the class members or be dealt with one by one. These items are used for subsequent numerical conversion of the values of PCN. Each characteristic of PCN stands for **P**: Previous action before the class, **C**: Current evaluation within the class), and **N**: Next action plan for the next class, respectively. More precise explanation about PCN is described as follows:



No Period	No Total	No Class	Name CLASS	No StNb	Name Name	Previous P	Current C	Next N	SPECIAL ITEMS Qty	Read	Blank	Gauti	Citation
3	4	4				-1	1	-1		-1			「食堂にしたい」
3	5	5				1	1	-1					「Macの基本操作が難しいです。」
3	6	6				1	1	-1					「ウィルスに侵入されたらとても怖いです。」
3	7	7				1	1	1					「意味を理解していきたい。」
3	8	8				1	0	-1					「だんだん難しくなってきた。」
3	9	9				1	1	-1					「知らないことばがたくさん」
3	10	10				1	1	-1					「知らない言葉が多くてよくわからなかった。」
3	11	11				1	1	-1					「日常生活にどういかしたらよいのか」
3	12	12				-1	1	-1					「フリーソフトだけでも十分なのでしょうか？」
3	13	13				-1	1	-1					「分かって驚いた」
3	14	14				-1	1	2					「しっかり対策しておこう」
3	15	15				-1	1	2					「危険性が分かり気をつけよう」
3	16	16				-1	1	-1					「簡単に直してほいげない」
3	17	17				-1	1	-1					「次回の授業も楽しみ」
3	18	18				-1	1	-1					「話を常くと身近に感じられ」
3	19	19				-1	1	-1					「またついていける」
3	20	20				-2	1	-1					「様々な方向から防止策」
3	21	21				-1	0	2					「全く分からなかったので勉強して」
3	22	22				1	1	-1					「便利は必ず危険という言葉がんに残った。」
3	23	23				1	1	2					「暗証番号に気を付けたい」
3	24	24				1	0	-1					「よく意味の分からないところもあった。」
3	25	25				-1	0	-1		-1			「便利と不便がいつも一緒」

Fig. 1. Analyzing the comments (in Japanese)

P indicates the action between right after the previous period of the class and the current period, teachers quantify action such as the preparation for this class, review of the previous class, submission and completeness of homework, and so on.

C indicates the understanding and achievements of current class; teachers quantify their students' self-evaluation with their experience.

N indicates the action planning from right after current class to the next class; teachers enumerate declaration, statement of preparation, review, and so on.

After all the comments are quantified, teachers can adjust the values from other information as the questionnaire of the day, the memories concerning the students, and/or the experience of the class. The concrete criteria of rating values of PCN are described in the following subsections.

Absence is treated exceptionally and given as -5 to all of P, C, and N.

3.2 Quantifying P Values

P indicates the learning action between the previous class and the current class, such as reviews of the previous class or preparations for the current class. In the real comments, students describe this kind of action such as "I trained typing" or "I read chapter 3 of the textbook". For quantifying the value of P, one is selected from 4 levels: **Attention (-1), Bad (0), Fair (1), and Good (2)**.

Attention (-1) is rated if there are no expressions related to previous learning actions, in any form, directly or indirectly.

Bad (0) is rated if there is abstract expression concerning previous actions, but not in detail. Teachers can confirm the fact of the action but not detailed contents. For example, from real comments, "I trained typing" insists reality of actions, but does not explain in detail such as training time, or achievement level.

Fair (1) is rated if there are any concrete expressions concerning previous actions, but the action level implied from the expression does not reach the level expected in the class. For example, the comment, "I trained typing, and achieved the speed of 100 strokes per minute" describes the fact and detail on the previous action, but the described fact (100 strokes per minute) does not reach the expected level (150 strokes per minutes) of the class.

Good (2) is rated if there are any concrete expressions concerning previous actions and the action level implied from the expressions reach the level expected in the class. For example, the comment, "I trained typing, and achieved the speed of 200 strokes per minute" shows the fact and detail on the previous actions, and the described fact

(200 strokes per minute) goes beyond the expected level (150 strokes per minute) of the class.

It is so difficult to acquire comments relating to **P** at the first period of the class that we exceptionally rate Bad (0) as a default value.

3.3 Quantifying C Values

C indicates understanding and achievement of the current class. Teachers determine the value from their experience. For example, for the comments, "I finished the first exercise" or "I didn't finish all exercise because time is up," one value is empirically rated by the teachers.

For quantifying the value of C, one is selected from four levels: **Attention (-1), Bad (0), Fair (1), and Good (2)**.

Attention (-1) is rated if there are no expressions indicating the facts of students' understanding or achievements in the current class, in any form, directly or indirectly.

Bad (0) is rated if there are any expressions indicating the facts of students' understanding or achievements, but those expressions are too abstract for teachers to extract the students' understanding level. For example, the comment "I didn't understand it" or "It was difficult," shows facts about students' understanding, but their achievement level is not clear.

Fair (1) is rated if there are any concrete expressions that help teachers infer the students' understanding and achievement level, but the level is not so high. For example, the comment "I have done the first exercise," concretely shows the fact of student's achievements, but only "the first exercise" does not reach the expected level of the class.

Good (2) is rated if there are any concrete expressions that help teachers infer the students' understanding and achievement level, which goes beyond the expected level of the class, such as "Today I have done all exercises."

Since it is sometimes difficult to acquire comments related to **C** at the first two or three periods of the class, teachers request students to write comments related to **C** because comments are freestyle and students have not accustomed yet. In such cases, we rate **Bad (0)** as a default value and adjust them per each student with questionnaire of the day, and teachers' experience and memories for students.

3.4 Quantifying N Values

N indicates action plan after the class, and is guessed from comments of students. Teachers guess students' action plan from comments, and rate them numerically. For example, for comments "I will make preparation by next class," "I found necessity to train typing," teachers rate **Good (2)** or **Attention (-1)**.

Attention (-1) is rated if there are no expressions concerning action plan in the comments, in any form, directly or indirectly.

Good (2) is rated if there are any expressions concerning action plan in the comments, in any form: determination, declaration, or implication, such as "I found necessity to train typing," "I think my preparation is not enough," "I recognized that I should do exercise not only in mind but also by hand," and so on.

It is known facts from teachers' experience that motivation of students becomes weaker at the final period of the class after submission of their final reports. They feel so free that they write their plans, determinations, and declarations related to **N** more boldly and intrepidly than ever. We do not adjust the values of the final period at present.

3.5 Special Items

We currently record 5 special items: **Quantity**, **Readability**, **Blank**, **Caution**, and **Citation**. They are defined as follows:

- **Quantity** is quantified into an integer if extremely short or long.
- **Readability** is quantified if the letter and figure in the comment are extremely rough or polite.
- **Blank** is quantified if any item required in the comment is blank or not found.
- **Caution** is quantified if a phrase should be shared in the class such as common mistakes, good hints, inappropriate attitude, or laziness.
- **Citation** is sample sentences clipped from the comments.

The reasons of recording such items are to help teachers adjusting the results into more precise one. These items reflect the characters of students and reinforce the reliability of the same results as teachers' experience and memories to the students. In addition, they enhance and improve teachers' own experience if new facts are found. If the cells concerning special items are full, it means that the classroom is full of the students with many problems. Since there are few students with problems, the cells for special items are mostly blank. In very few cases, the cell is filled with -2 or less, only extremely bad situation.

4 Case Study

4.1 Correlation between PCN Values and Credits

As mentioned earlier, the PCN method quantifies learning activities described in freestyle comments. This enables teachers to visualize the tendency of each student's behavior in each period of the class. Teachers acquire the clues of understanding of students' learning activities if those are accidental or natural. Actually, **P** indicates preparation activity for the class. **N** indicates some activities related to reflection and motivation for the next class. Moreover, if we combine **N** and **P**, e.g., the m-th period value of N ($\mathbf{N_m}$ for short) and the (m+1)-th period value of P ($\mathbf{P_{m+1}}$ for short), we will find the relationship between the m-th preparation activity plan and the corresponding (m+1)-th real preparation activities.

Table 1. The correlation between PCN values and credit scores of students in a class. Pos. and Neg. present positive and negative values, respectively.

	P	C	N
Pos. and Neg.	0.742	0.786	0.655
Pos.	0.378	0.515	0.329
Neg.	0.769	0.776	0.748

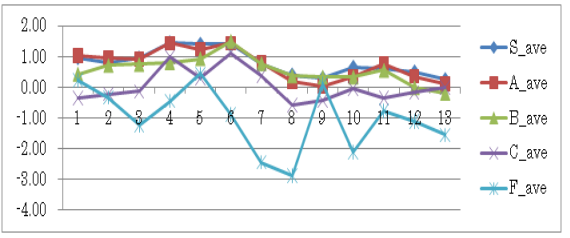


Fig. 2. C values for 5 groups: S, A, B, C, and F.

To apply the PCN method, we first analyzed the comments, and found that the following facts:

1. Many students tend to skip preparation activities to the class.
2. Many students describe the action plan to their next class.
3. Most of them do not make practice in real.

Next, we sum up **P**, **C**, and **N** of all the periods for each student, and calculate the correlation coefficient between the sum and the final score of each student's credit. We also sum up positive part and negative part of comments, and calculate the correlation coefficient between the sum of each part and the final score. The positive part of comments is the part that only non-negative values are summed up and negative values are treated as zero. The negative part of comments is the par that only non-positive values are summed up and positive values are treated as zero. The results are shown in Table 1. They depict strong correlation between the sum of negative parts of all of P, C, N values and the final score of credits. On the other hand, the sum of positive part of P, C, and N only shows weak or at most medium correlation. As references, the correlation between students' final score of credits and their report points is 0.634.

Fig. 2 shows the C value transitions of five groups divided by credit ranking: S (Outstanding, n=20), A (Excellent, n=25), B (Good, n=32), C (Fair, n=23), and F (Poor, n=9). Where X-axis is the period in the semester, and y-axis is the C value of the groups. The results show that we can clearly distinguish higher grade groups (S, A, B) from lower ones (C, F) with C value.

4.2 Comparison of Two Classes: PCN Values and Their Transition

Firstly, we calculated the average sum of PCN values for Class-A and Class-B at each period. From the results shown in Table 2, we found two singular points. For N_7, P_8 and P_9, two areas are distinguished from other areas. When considering N_7, the contents of the class changed marvelously. The class in the 6th period gives a lecture of computer literacy, which gives how to use some IT tools such as word processor, spread sheet, and presentation tool, and the lecture was changed to C programming from the 7^{th} period class. The computer literacy subject is educated compulsory and widely all over the senior high schools in Japan, and only a few contents differ their detail. On the other hand, programming with language C or others is not a required

subject until entering the university; In fact, most students are novices at programming. At the 7th period, the teacher explains the fundamental element and basic procedure of C programming slowly and precisely. Each student may feel that programming is very difficult, and feel necessity and importance of preparing the class. We regard it as natural that such the situations mentioned above greatly increase the value of N between the 6th and the 7th periods.

Table 2. The transition of PCN values by periods

		1	2	3	4	5	6	7	8	9	10	11	12	13
P	Ave	-1.3	-1.2	-0.4	-1.2	-0.4	-0.0	-0.1	-1.5	-0.1	0.3	0.5	0.5	0.3
	diff	X	0.1	0.8	-0.8	0.8	0.3	-0.0	-1.4	1.3	0.4	0.2	0.0	-0.2
C	Ave	0.5	0.5	0.5	1.0	0.9	1.2	0.4	-0.1	0.1	0.1	0.3	0.0	-0.1
	diff	X	0.0	-0.0	0.5	-0.1	0.3	-0.8	-0.6	0.2	0.0	0.2	-0.3	-0.2
N	Ave	-0.1	-1.2	-0.2	-0.2	-0.1	-0.4	1.2	0.8	0.7	-0.5	0.7	0.0	-0.8
	diff	X	-1.0	1.0	-0.0	0.1	-0.3	1.7	-0.4	-0.1	-1.3	1.2	-0.7	-0.8

Secondly, we consider P_8 and P_9. P_8 has the greatest decline at this period in the semester. On the other hand, P_9 goes up with the second biggest gap. It makes V curve between P_8 and P_9. The reasons can be explained as follows. As the subjects changed drastically in the 7th period from that in the 6th period, it made students feel so uneasy and feel necessity of preparing their class more than before. Thus, N_7 increased powerfully as mentioned before. On the other hand, the reason why P_8 falls down very much is that at the 7th period, the teacher spoke a lot to students so that they felt that programming is easy and fun. However, it was too much and made them underestimate the difficulties of programming, and caused them not to prepare the class. However, at the next period, P_9 rose again because students recognized and reflected that they should have prepared the class sufficiently.

Since we found the big difference between the 6th and 7th period of the class, we analyzed and compared the two segments, before seventh (first half) and after sixth (latter half); we also classified the students into positive and negative thinking groups. The results are shown in Table 3. As we inferred, latter half periods and the negative thinking groups showed the strongest correlation with their final scores of the class credits. This implies us two facts: 1) what and how they learned, or their learning process affects the final scores (credit score) not so greatly. 2) Negative action or no action on learning affects the scores greatly.

Next, we compared two classes by calculating the average of PCN values. The results concerning C are shown in Fig. 3. From the figure, we found two tendencies. In the first half period of the class, Class-B tends to be higher than Class-A. In the latter half period, Class-B seems to be more stable than Class-A. We considered the difference between Class-A and Class-B from the viewpoint of the comments. We read the comments again, and found that the comments of Class-B students tend to be more straight-forward and concrete than those of Class-A students. This implies that Class-B students tend to be more direct and talkative than Class-A students.

Table 3. Analysis by period group

	ALL			1st-6th			7th-13th		
	P	C	N	P	C	N	P	C	N
Pos.+Neg.	0.742	0.786	0.655	0.577	0.633	0.443	0.726	0.717	0.609
Pos.	0.378	0.515	0.329	0.154	0.455	0.149	0.401	0.365	0.339
Neg.	0.769	0.776	0.748	0.622	0.634	0.576	0.744	0.721	0.694

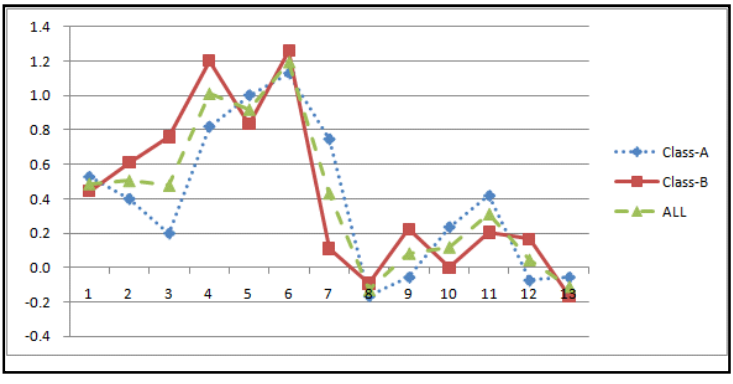


Fig. 3. C value transition of Class-A, Class-B, and All (Combined two classes)

We also compared the average of P values between Class-A and Class-B (Fig. 4 Left), and found Class-A tends to be higher than those of Class-B about P values. This means that Class-A students tend to make more preparation than Class-B, and also implies Class-A students tend to more serious than Class-B which is similar to the teacher's intuition. About C values Class-A tends to be lower than Class-B. This means that Class-A students tend to understand or achieve less than Class-B, and implies the average sum of credits of all the Class-A students is lower than those of Class-B, and this inference is against the result, or credits of the class. From this gap and teacher's feelings in the classroom, we infer Class-A students are pessimistic (or they write worse than real) and Class-B ones are optimistic (or they write better than real). Although we trust all the comments of each student as premise, some exaggerations cannot be avoidable and should be accepted.

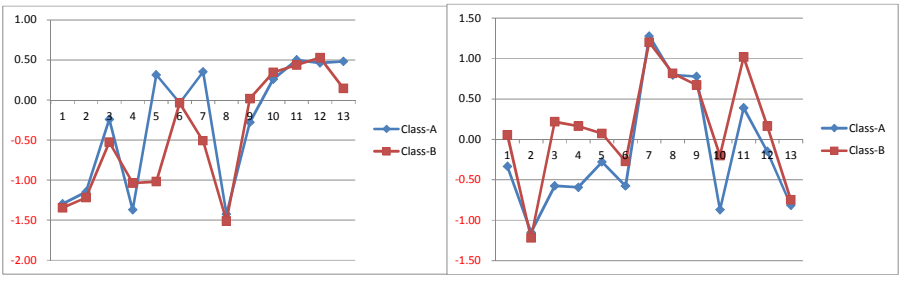


Fig. 4. P value transition (Left) and N value transition (Right) of Class-A and Class-B

Next, we focus on **N** value to compare the two classes (**Fig. 4** Right). We found that Class-A tends to be lower than Class-B, opposite tendency again C value transition. This implies Class-B students tend to declare their preparation or reviews explicitly but fail to do as they have written.

5 Conclusion

In this paper, we discussed the PCN method. The method quantifies the freestyle class comments and enables teachers to grasp the tendencies of students' learning activities in the class. The activities grasped are not only for the whole class members, but also for each member in the class. Concerning individual learning behaviors, we grasp the current status and trace the change of his/her activities with the PCN method. The PCN method provides the basis of improving both the class and students' learning activities. The statistical tests are left as the future work.

Although the costs of using the PCN method may not be cheap due to the workload of reading and quantifying the students' comments, some of techniques such as automations, natural language processing, or text mining would be reduce the costs. We are now trying to adopt text mining techniques for this objective.

References

1. Brusilovsky P.: Adaptive Hypermedia. User Modeling and User-Adapted Interaction 11, 87–110 (2001); Bull, S., Greer, J., McCalla, G., Kettel, L., Bowes, J.: User Modelling in I-Help: What, Why, When and How. In: Bauer, M., Gmytrasiewicz, P.J., Vassileva, J. (eds.) UM 2001. LNCS (LNAI), vol. 2109, pp. 117–126. Springer, Heidelberg (2001)
2. Burstein, J., Wolska, M.: Toward evaluation of writing style: Finding overly repetitive word use in student essays. In: Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics (2003)
3. Gagne, R.M.: The conditions of learning and theory of instruction, 1st edn. Holt, Rinehart & Winston, New York (1965)
4. Ishioka, T., Kameda, M.: Automated Japanese Essay Scoring System based on Articles Written by Experts. In: Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 233–240. Association for Computational Linguistics (2006)
5. Ito, T., Suzuki, K.: Development of an effective and sustainable system for ID training: Proposing a strategy model of Training of Trainer (ToT). Educational Technology Research 31(1-2), 13–24 (2008)
6. Keller, F.S.: Goodbye teacher. Journal of Applied Behavior Analysis 1, 79–89 (1968)
7. Van Kleek, M., Shrobe, H.: A Practical Activity Capture Framework for Personal Lifetime User Modeling (PLUM). In: Conati, C., McCoy, K., Paliouras, G. (eds.) UM 2007. LNCS (LNAI), vol. 4511, pp. 298–302. Springer, Heidelberg (2007)
8. Koschmann, T.: Paradigm Shifts and Instructional Technology: An introduction. In: Koschmann, T.D. (ed.) CSCL: Theory and Practice of an Emerging Paradigm, pp. 1–24. Lawrence Erlbaum, Hillsdale (1996)

9. Martins, A.C., Faria, L., Vaz de Carvalho, C., Carrapatoso, E.: User Modeling in Adaptive Hypermedia Educational Systems. Educational Technology & Society 11(1), 194–207 (2008)
10. Popescu, E., Trigano, P., Badica, C.: Towards a unified learning style model in adaptive educational systems. In: Proc. ICALT 2007, pp. 804–808 (2007)
11. Romero, C., Ventura, S.: Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications 33, 135–146 (2007)
12. Zimmerman, B.J.: Self-regulated learning and academic achievement: An overview. Educational Psychologist 25, 3–17 (1990)

Back-Review Support Method for Presentation Rehearsal Support System

Ryo Okamoto[1] and Akihiro Kashihara[2]

[1] Faculty of Science, Kochi University, Kochi, Japan
ryooka@is.kochi-u.ac.jp
[2] Department of Information and Communication Engineering,
The University of Electro-Communications, Tokyo, Japan
akihiro.kashihara@inf.uec.ac.jp

Abstract. We have been developed the peer review support system for a presentation rehearsal. The purpose of the systems is to refine presenter's knowledge and presentation through a review work of the peer review results as a feedback. For the review work, the system have to organize a lot of comments accumulated in the peer review process, and represent them in an easy-to-understand way. In this paper, we discuss an annotation groping method based on a presentation structure, which for a back-review process after a presentation rehearsal.

Keywords: presentation rehearsal, peer review support, presentation structure.

1 Introduction

The purpose of a presentation rehearsal is enabling a presenter to be aware of the insufficiency or incompleteness of his/her knowledge and to refine his/her knowledge, presentation materials (PowerPoint slides for example), and oral explanations. However, it is sometimes hard for peers to remember details of a particular part of the presentation precisely. This hinders the peers from giving explicit and practical comments to the presenter and causes some disagreements in the discussion. Therefore, the presenter cannot obtain instructive suggestions or resolutions from the peer review. Additionally, after receiving the peer review, the presenter should review the results of the peer review to refine his/her knowledge.

Unfortunately, the presenter cannot often remember the accurate situations of the discussion concern with each review points. To facilitate effectiveness of peer review in presentation rehearsal, we have pointed out four types of difficulties in a presentation rehearsal as follows [1].

(1) Pursuing the progress of a presentation concurrently
(2) Remembering details of the presentation precisely
(3) Tracing topics reviewed
(4) Remembering the result of the peer review precisely

Our approach to this issue is to propose utilization of hypervideo technique [2.] as a basis of a framework for supporting peer review works. In general, utilization of video

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 165–175, 2011.

technique is very popular to record and playback situations of a presentation rehearsal. Nevertheless, in the traditional way, recorded video sequences cannot playback partially to respond to the needs. Thus, we have proposed real-time-created hypervideo techniques [3] for a selective and partially playback of situations and attempted to realize time-shift of the peer review works.

We have been developing the prototype system of the presentation rehearsal support system for recent three years and made a test use of around 50 presentations. Now, we are in the term of focusing the process of after the rehearsal for revise the presentation materials such as slides for a next rehearsal. We call the process as “back-review”. It means a review of results of the review work in the presentation rehearsal. Most of the presenters rehearse more than once. Even after the second round, the issues, which are pointed out by the peers in the previous rehearsal, are often left unsolved or in unexpected state for the peers.

In the cycle of a knowledge refinement, the presentation is the later half of the processes which as knowledge publishing of pre-acquired knowledge to others. From this viewpoint, to support a process of a presentation rehearsal including back-review is not direct support of processes of acquiring or refining presenter’s knowledge. However, we believe to provide an effective and sufficient support for back-review work brings presenter awareness toward next phase of knowledge acquisition or refinement of insufficiency of knowledge.

In this paper, we propose the back-review support method and describe a design and prototyping of a back-review support system.

2 Presentation Rehearsal Support System

As an approach to solve the difficulties and get efficiency of a presentation rehearsal, we have been developing a prototype system of a presentation rehearsal support system [1], [3]. We have made improvements the system gone along with operational test since 2007 and recorded over 50 sets of rehearsal data. Figure 1 shows scenes of presentation rehearsals with the system for a graduation thesis.

In this chapter, we describe a process model for a presentation rehearsal and the system configuration as a basis of our approach to peer review and back-review support.

Fig. 1. Scenes of Presentation Rehearsal with Support System

2.1 Presentation Rehearsal Model

To solve the difficulties in a presentation rehearsal by computerized supporting system with hypervideo technique [3], we propose a presentation rehearsal model as shown figure 2. One of the main features of the model is to insertion some interval between a presentation period and a discussion period to up quality of comments from reviewers with hypervideo. And another is an extension of a support for an activity, which will be done by a presenter after presentation rehearsal. We call the process as a "back-review". It means a review of results of the review work in the presentation rehearsal. For the back-review, we have tried to use whole resources effectively which are acquired in the peer review process.

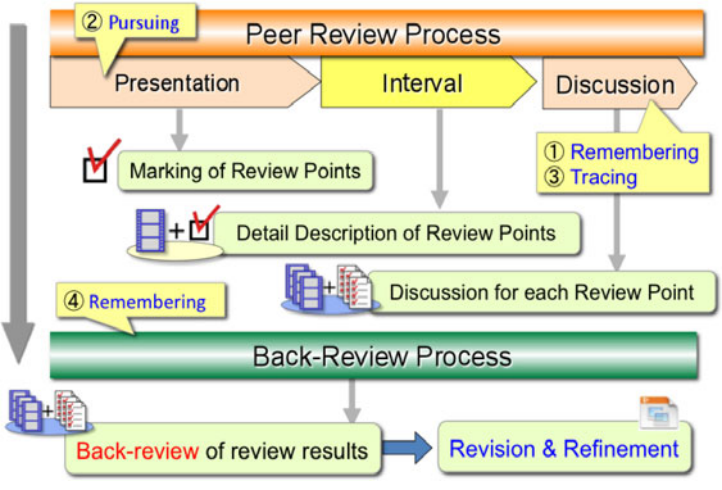


Fig. 2. Presentation Rehearsal Model

2.2 System Configuration

We have been developing a prototype system to solve the issues in chapter 1. Figure 3 is shows a configuration of the system. The system is basically developed from scratch with cross-platform environment, so they work on both of MacOSX and Windows platform. The system consists of five kinds of application software and each application is worked with other applications via network.

(1) Presenter Client

The presenter client works on a presenter's laptop computer with a presentation software (Keynote on MacOSX and PowerPoint on Windows). The client extracts information about presentation slides and timings of presentation progress from the presentation software and sends them to a review server simultaneously via network.

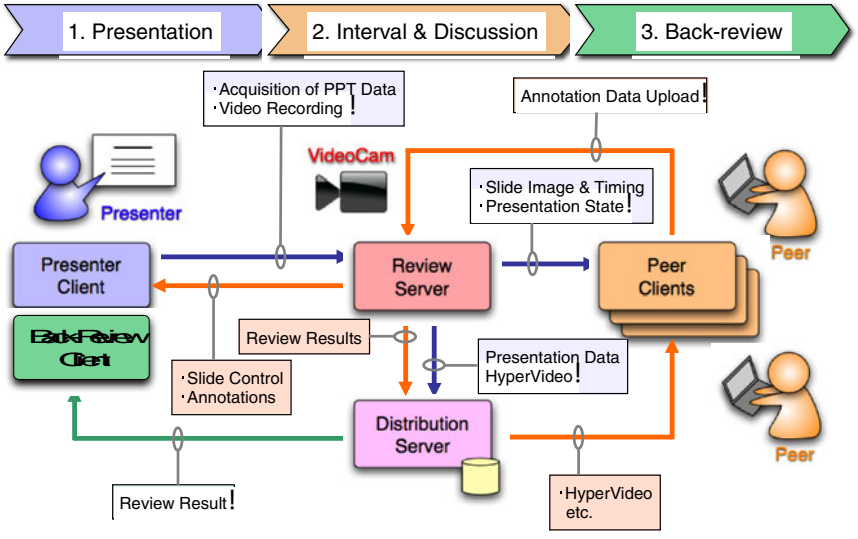


Fig. 3. System Configuration

(2) Review Server

This server is a center of the supporting environment and has many functions. There are three major functions, (a) Video recording function from connected video camera, (b) Collecting and distributing function of presentation information and (c) Management functions of peer review works. The server works as a hub in cooperation with other applications.

(3) Peer Client

This client works on peer's laptop computer. While presenter performs presentation, the client acquires information of the presentation, images and texts of each slides and transition timing of slides etc., from the review server in real time. The peers can make annotations in any points of the presentation as a peer review work, and the results of the work are transmitted to the review server at the same time.

(4) Distribution Server

All data received in the review server are transmitted and stored in the database of this server. In the process of discussion phase of peer review, peers access to the server with a peer client and retrieve annotation data for proceeding arguments. And also, the results of the peer review work are added to the database with relations to each annotation during the discussion.

(5) Back-Review Client

After a presentation rehearsal, the presenter connects to distribution server and reviews all data for back-review work to refine his/her knowledge and presentation materials through this client. The application is mainly used for organize comments which are accumulated in the rehearsal process.

3 Approach to Back-Review Work Support

In this chapter, we discuss about "Back-Review Work" by a presenter and some issues about it. Then, we propose our back-review method as a solution for the issues.

3.1 Back-Review Processes

The aim of the presentation rehearsal is to revise the contents or performance of the presentation based on the comments from peers who subscribe the rehearsal. Through the process the presenter would aware the insufficiency or incompleteness of him/her knowledge and revise their slides or oral presentations. These tasks are performed through the "back-review work", and the work is quite important to get enough effect of the rehearsal. The processes of the work are as follows.

(1) Take minutes of results of the peer review

The presenter should take part in a discussion, which is based on comments from peers, and take minutes of it. After that, the presenter brings the minutes back and revises presentation materials, such as slides.

(2) Specify the point at issue

As the first step of the back-review work, the presenter should make a thorough review of the points are listed in the minutes and specify the point at the issue in his/her presentation materials.

(3) Figure out the solution to the issue

The main purpose of a back-review work is to revise a presentation through concrete revisions of the issues. The presenter should be aware of the insufficiency or incompleteness of his/her knowledge and find out the solution to the issues.

3.2 Issues of Back-Review Work

Most of the presenters rehearse more than once. Even after the second round, the issues, which are pointed out by the peers in the previous rehearsal, are often left unsolved or in unexpected state for the peers. We have focused on these situations and pointed out two types of difficulties in a back-review work as follows.

(1) Difficulty of specification of all review points

In the back-review work, the presenter specifies review points from the minutes taken while the discussion of the rehearsal. In many cases, the types of comments earned from peers are very various, and it is quite difficult to take all the remarks within the discussion time.

(2) Incompleteness of remembering of an involved discussion

In the discussion phase, the presenter fully understands a comment at least once. But, in the back-review phase, there is no assurance that the presenter can remember all of review points. Sometimes, the presenter has to revise materials with no confidence.

3.3 Approach to Back-Review Support

To solve the issues mentioned above, we have considered approaches from three viewpoints as follows.

(1) Recording and reproduce of review results

For the reason above-mentioned, in the discussion phase, it is difficult for a presenter to take minutes of all remarks about various types of the comments. To settle the issue, a support environment to record all comments completely and reproduce them is required. In general, peers, as reviewers, take notes of points of the comments during a presentation is in progress and use them later in a discussion. If all of these notes are stored and referred in back-review work, the presenter would remember all review points.

(2) Specification of review points by an organization of stored comments

The presenter gets large number and various types of comments from peers, and the comments are redundant for the most. To specify the review points to be revised, the presenter should organize all the comments and exclude redundancies for clarity of review points.

As an approach for the issue, we propose the use of a presentation structure to organize the comments in association with certain elements of the presentation. The presentation structure is a hierarchical structure as shown in figure 4. By use of this structure in the back-review work, we are aiming to realize support functions.

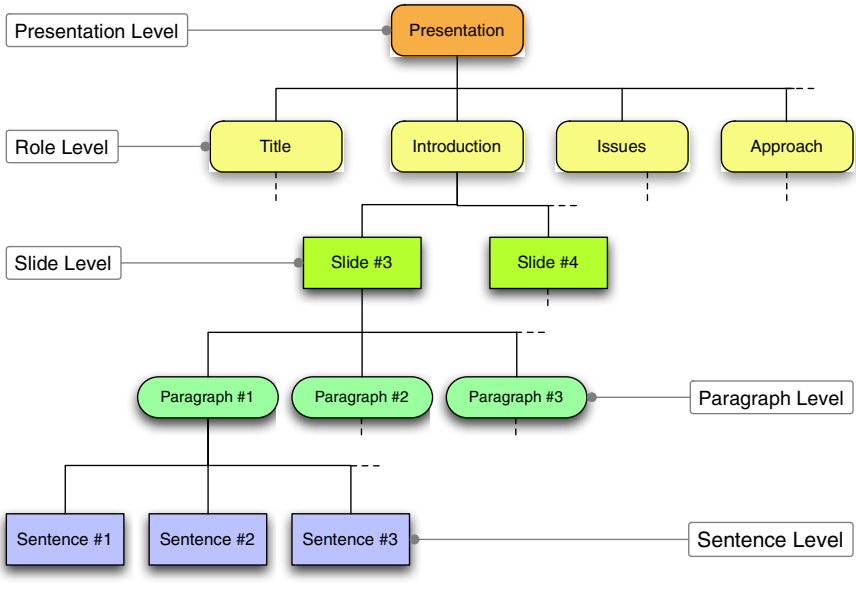


Fig. 4. Presentation Structure

(3) Revise Memo for Review Points Group

In a presentation structure, some of the groups of review points are appeared in certain place. We also propose the use of a revise memo for a corrective strategy to it. The work for making memos for all review points groups are preparations of a revision of presentation materials, and it will bring a definite back-review.

4 Implementation of Back-Review Work Support

In this chapter, we propose the framework of back-review support and a prototype system. The framework consists of three phases, "Comments Accumulation Phase", " Comments Organization Phase" and "Revise Memo Writing Phase".

4.1 Comments Accumulation Phase

This is a preparing step for the back-review work. In this phase, we use the presentation rehearsal support system developed in our preceding work for an accumulation of comments from peers.

During the presentation, peers use "Peer Client" to make annotations about curious or worrisome points. In this period, peers simply make annotations as a short memo, and after the presentation, peers can see information about the presentation. Figure 5 shows an example of the interface of the client application. The interface mainly consists of two types of display areas.

Fig. 5. Peer Client

The first is a display area for presentation data, which are located in the left. In this area, a peer can select certain slide timing from “Slides” list to load a related slide image and hypervideo segments. As a function for hyper-browsing to other hypervideo segment, keywords of the current slide are listed, and by clicking on these keywords, related slide information appear in the “Hyper Segments” list.

The second is a working area for an annotation work in lower right of the window. Peers can use four kinds of buttons labeled “Slide Contents”, “Slide Layout”, “Story” and “Others” temporarily to make a new record of a comment. Actually, the types of the comments are very various by a person, but complicated definition of categories of comments and to force peers complicated operation will disturb a smooth review work. Therefore we use this general classification here, and a presenter carries out a real classification of comments in “Comment Organization Phase” as a back-review work. A created comment is attached to a current slide is selected in “Slides” list, and if nothing is chosen in the list, the comment is treated as a general comment. The presenter in a back-review work also decides this attribute of an attachment. After presentation, all participants of the review work take few minutes interval to give the server necessary time to generate hypervideo. The peers go into details about the comments.

4.2 Comments Organization Phase

In this phase, the presenter organizes the comments accumulated in the previous phase by mapping the presentation structure. After presentation rehearsal, to enhance an efficacy of the peer review, it is necessary to give an opportunity for the presenter to remember the situations of the discussion with the peers after the rehearsal.

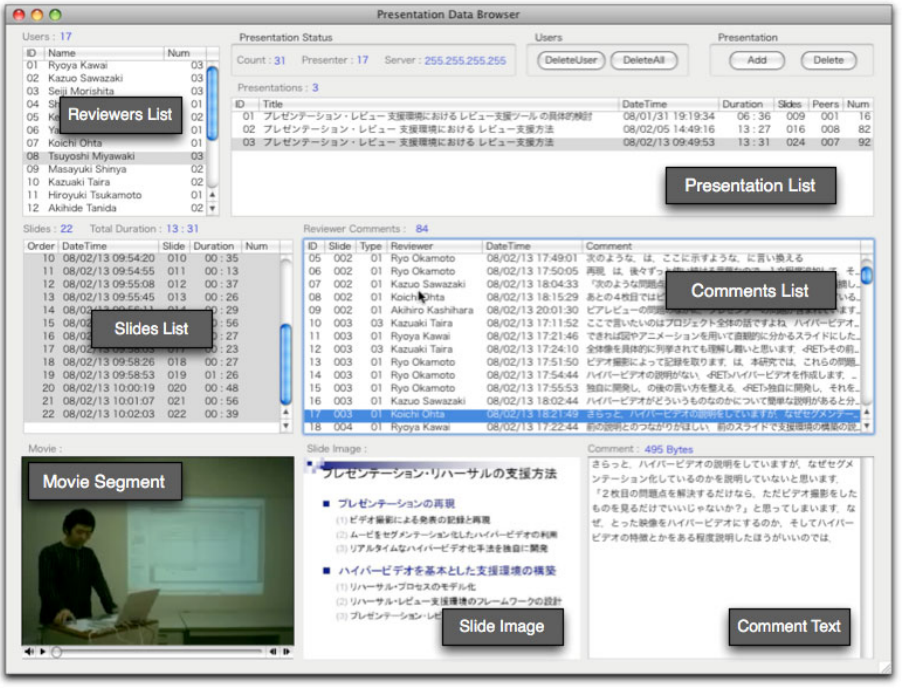


Fig. 6. Presentation Data Browser Mode of Back-Review Client

The “Distribution Server” works as “Back-Review Server” to distribute resources, which includes all kinds of data accumulated in the rehearsal. The presenter can access to the server to back-review his/her presentation any time by “Presentation Dater Browser Mode” of “Back-Review Client” as shown in figure 6.

Figure 7 shows an example of an interface of “Comment Organizer Mode” of “Back-Review Client”. The organization of the comment is performed according to the following procedure.

(1) Construction of the Presentation Structure

The presenter constructs the presentation structure at the start of the mode. At first, the presenter makes nodes of a role level of the slides under the root node labeled by the presentation title. Then the presenter drag and drop the slide’s thumbnail to appropriate location to make a thread of the slides. The appearance of the structure is not the same to the presentation structure shown in figure 2. In the interface, we adopted the design to display the level of three high ranks of the hierarchy structure for the reason of usability, but the lower-ranking level, paragraph level and sentence level, are not implemented this version of the prototype. To implement these levels, we are planning to apply a visual annotation technique in the next version.

(2) Grouping of Comments and Making Relation to the Presentation Structure

When the presenter loads the set of presentation data, some comments have already related to a certain slide, and others are related to the root node of the structure. The presenter can select any comments and investigate details of them and bundle similar comments from different reviewers to a group by manual drag and drop operations. The groups are displayed in “Presentation Structure Field ” as an icon of pins. The groups with a relationship to extend to plural nodes of slides or levels can draw a branch, if it is necessary.

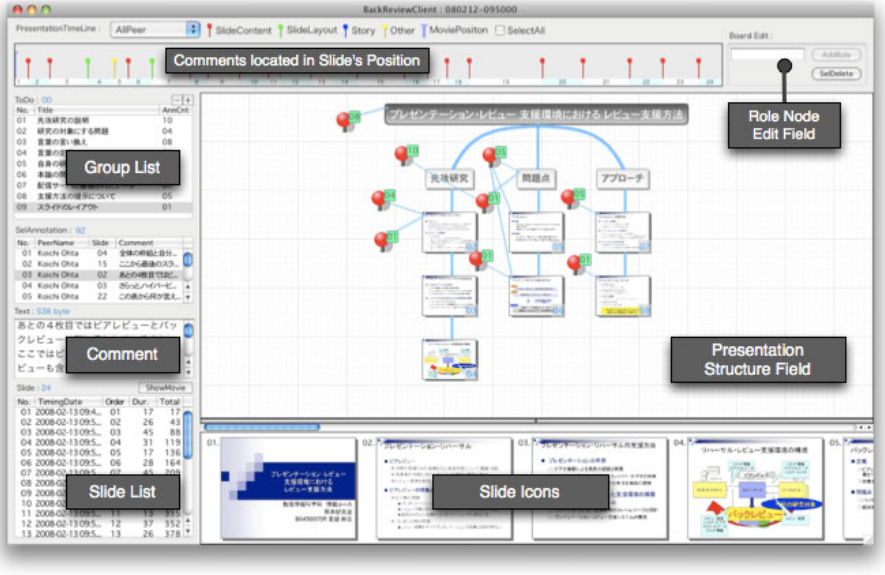


Fig. 7. Comment Organizer Mode of Back-Review Client

4.3 Revise Memo Writing Phase

Figure 8 shows an interface of “Revise Memo Writing Mode” of Back-Review Client. This interface basically consists of two area, memo writing area and comments browsing area. This mode is called when the presenter select the particular comment group from the interface of “Comment Organizer Mode”. The presenter finally refers these memos and revise presentation material.

Fig. 8. Revise Memo Writing Mode

5 Conclusion

In this paper, we described the design of a review support functions for a presentation rehearsal based on the configuration of our supporting system. And we also proposed the back-review support method and mentioned about the implementations of back-review client.

Presently, we made a test use of around 50 presentations, which includes from our collaborators, and the system works almost well. As our future work, we continue developing the system and consider about supports for a discussion process in a presentation rehearsal.

Acknowledgements. This research is supported in part by Grant-in-Aid for Scientific Research (C) (No.22500925) and (B) (No.23300297) from Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Okamoto, R., Kashihara, A.: Designing Presentation Review Environment with Realtime-created Hypervideo of Presentation Rehearsal. In: Proc. of 22th Annual Conference of Japanese Society for Educational Technology, pp. 703–704 (2006)
2. Sawhney, N., Balcom, D., Smith, I.: Hypercafe: Narrative and Aesthetic Properties of Hypervideo. In: Proc. of Hypertext 1996, pp. 1–10 (1996)
3. Okamoto, R., Kashihara, A.: Presentation Review Supporting Environment with Realtime-created Hypervideo Technique. In: Proc. of The 15th International Conference on Computers in Education (ICCE 2007), Hiroshima, Japan, pp. 433–440 (2007)
4. Kashihara, A., Hasegawa, S.: Unknown Awareness in Navigational Learning on the Web. In: Proc. of ED-MEDIA 2004, pp. 1829–1836 (2004)

Multiple Hypothesis Testing and Quasi Essential Graph for Comparing Two Sets of Bayesian Networks

Hoai-Tuong Nguyen, Philippe Leray, and Gérard Ramstein

LINA - Nantes Atlantique Computer Science Lab, CNRS 6241
COD - Knowledge and Decision Team
La Chantrerie - rue Christian Pauc
BP 50609 - 44360 NANTES Cedex 3, France
{hoai-tuong.nguyen,philippe.leray,gerard.ramstein}@univ-nantes.fr
http://www.lina.univ-nantes.fr/-COD-.html

Abstract. In machine learning, graphical models like Bayesian networks are one of important visualization tools that can be learned from data to represent pictorially a complex system. In order to compare two complex systems (or one complex system functioning in two different contexts), one usually compares directly their representative graphs. However, with small sample size data, it is hard to learn the graph that represents precisely the system. That's why ensemble methods (e.g. Bootstrapping, evolutionary algorithm, etc...) are proposed to learn from data of each system *a set of graphs* that represents more precisely this system. Then, for comparing two systems, one needs a mechanism to compare two sets of graphs. We propose in this work an approach based on multiple hypothesis testing and quasi essential graph (QEG) to compare *two sets* of Bayesian networks.

Keywords: Bayesian Network, Ensemble methods, Multiple hypothesis testing, Quasi Essential Graph.

1 Introduction

A graph is a representation of a set of objects, that consists mainly of a finite set of ordered pairs of the objects (called *nodes/vertices*) connected by links (called *edges/arcs*). In machine learning, graphical models like Bayesian networks can be learned from observational data to represent a large scale and complex system. Normally differences between graphs can predict the differences of systems. However, it is hard to describe the real differences. In fact, in many real applications, the sample size of available data is lower than the number of observed variables. For this reason, *ensemble methods* [12] (e.g. Bootstrapping [7,13], evolutionary algorithm [9,5,2,14], etc...) propose to learn *a set of graphs* (used shortly *"a set"* in the next sections). This set can represent as precisely as possible the complex system. Then, for comparing two systems, one needs a mechanism to compare two sets of graphs.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 176–185, 2011.

The main contribution of this paper is an approach based on multiple hypothesis testing to compare *two sets* of Bayesian networks (BNs). Two sets of simulated BNs are experimented to demonstrate the performance of proposed approach.

In this work, we have to deal with the three following issues:

Firstly, one of major issues for graph comparison approaches using BNs is the Markov equivalence. In fact, some edges can be inverted without changing the underlying independence model. In other words, two structurally different graphs can correspond to the same dataset. So a direct comparison of two graphs is impossible. One solution (cf. Section 2.1) is using one property of Markov equivalence: all the equivalent graphs can be summarized by an *essential graph* [10]. That means in order to compare two graphs we have to compare the essential graphs of this two original graphs.

Secondly, there is no parametrization for graph distribution. Therefore, in order to compare two sets of graphs, we have to transpose the problem into edge comparisons. As there are numerous edges in each graph, we propose to apply a multiple test. When many hypotheses are tested, the chance of committing some Type I errors (false positives) probability increases, often sharply, with the number of hypotheses. A p-value of 0.01 no longer corresponds to a significant finding. Thus, we must correct the significance threshold α. The choice of test and α correction are presented in Section 2.2 and 2.3.

Thirdly, in order to limit the number of tests, we proposed an approach to eliminate noisy edges that are unuseful to test (cf. Section 3). These edges have a small probability of occurrence. This approach resumes statistically each set of BNs into a "most" representative named *quasi essential graph* (QEG) [11]. QEG allows us to find the most relevant edges in each original set of graphs.

In the Section 4, we present our experiments and results.

2 Multiple Test for the Comparison of Two Sets of Bayesian Networks

2.1 Bayesian Networks and Markov Equivalence

Bayesian networks (BNs) are probabilistic graphical models which have been widely used for modeling knowledge, prediction or classification tasks in various domains. A BN, $\mathcal{B}$, use directed acyclic graph (DAG), $\mathcal{G}(V, E)$ *where* V *is a set of vertices and* E *is a set of edges*, as a core of model to represent the relationship between variables. The important advantage of the use of BNs in this work is that we can identify different aspects occurred in the real graph-based application: (1) model learned from observational data; (2) taking into account conditional dependence/independence between variables.

However, as mentioned above, in order to use the BNs we have to deal with the problem of Markov equivalence. Two DAGs are Markov equivalent iff they have the same links (edges without regarding for direction) and v-structures [1] [4].

[1] V-structure is a triplet (X, Y, Z) where $X \rightarrow Z \leftarrow Y$ and (X,Y) are not adjacent.

If we do not take into account the problem of Markov equivalence, we risk to consider two equivalent graphs as two different graphs.

One solution for the problem of Markov equivalence allows to use *essential graph* (EG) [10]. An EG is defined as a graph with directed edges corresponding to every *compelled* edge (present in all DAGs of the equivalence class) and undirected edges corresponding to every *reversible* edge (not compelled) in this equivalence class. [4] provides an efficient algorithm for determining the essential graph $EG(\mathcal{G})$ of a given DAG G.

After transforming DAGs to essential graphs, we can apply hypothesis tests on these obtained essential graphs. However, in this work, we limit to demonstrate only the proposed approach on the skeletons (undirected graph obtained by ignoring the direction of the directed edges). We will deal with essential graphs (including directed edges) in future work. The next section presents the choice of test for the problem of skeleton comparison of *two sets* of graphs.

2.2 Choice of Test

As mentioned above, the goal of this work is to propose a test for comparing two sets of graphs. As there is no parameter for a graph distribution. Therefore, in order to compare graphs, we have to transpose the problem into edge comparison. Moreover, the relationship between nodes is very important in the real-world applications. For example, in biological network, the change of the relationship between a pair of genes can radically alter the biological properties of all the network. Therefore, we focus on the change of each *edge* relationship between two sets of graphs.

We continue to identify the kind of data. The observational variable is the occurrence of edges that takes two values *"absent"/"present"* (*qualitative* variable). In the context of comparing two sets of graphs, we propose to study on the *frequency* of the occurrence of edges in each set of graphs. That's why, in this work, we chose *the tests on frequency data* in which *Fisher's exact test* is an appropriate test for our problem. In fact, we do not have any informations about the distribution of data, we can not use binomial test. On one hand, Fisher's test always gives the exact P value. In the other hand, Chi-square test is simpler to calculate but yields only an approximate P value. Therefore, we prefer Fisher's exact test to ensure the accuracy of results. It uses the hypergeometric distribution to calculate the p-value P_i (cf. eqn 1) for edge e_i where n_i (resp. $\overline{n_i}$) is the number of times the corresponding edge is present (resp. absent) in the first set (of size n), n'_i (resp. $\overline{n_i}'$) in the second set (of size n').

$$P_i = \frac{(\overline{n_i} + \overline{n_i}')!(n_i + n'_i)!n!n'!}{\overline{n_i}!\overline{n_i}'!n_i!n'_i!(n + n')!} \tag{1}$$

In our research context, we need to determine the difference of the frequencies of occurrence of edges between two sets of graphs. That means, we use one Fisher's exact test for each possible edge e_i. As there are many edges in a graph, we have to apply a multiple test. This problem is discussed in the next sections: (1) the

classical problem of the correction for the significance threshold for each test (cf. Section 2.3) and (2) how to reduce of the number of tests (cf. Section 3).

2.3 Significance Threshold Correction

Normally, the most common approach of the multiple testing problem consists of two steps: (1) computing a test statistic T_i (or its p-value P_i) for each test i; (2) applying a multiple testing procedure to determine which hypotheses to reject while controlling a suitably defined Type I error rate (false positive error rates) at level of significance α. However, if we apply one test with $\alpha = 0.05$ typically, the probability of getting a false positive result is 0.05 and the probability of not getting a false positive result for a single test is $1 - \alpha = 0.95$. Now suppose that, we perform $m = 10$ tests, each with $\alpha = 0.05$. The probability that we will get at least one false positive result is $= 1 - 0.95^m = 1 - 0.95^{10} = 0.4$. The problem is the probability of at least one false positive result is near *certain* if we do 1000 tests with $\alpha = 0.05$. That means we *can not* find any evidence to reject the null hypothesis. Thus, we must correct α *less conservatively* in order to increase the possibility of rejecting null hypothesis. This procedure is called the control of making Type I error or α correction [6].

In the literature, almost all α correction methods are based on Family-wise error rate, FWER (Bonferroni's correction and Bonferroni-Holm's correction) and False discovery rate, FDR (Benjamini-Hochberg's correction for instance). Bonferroni correction [1] proposes to divide the target α by the number of tests being performed. For precedent example, if we want apply $k = 1000$ test, the local $\alpha_i = \alpha/k = 0.05/1000 = 0.00005$. If the p-value is less than the Bonferroni-corrected target α, then reject the null hypothesis. Bonferroni correction is called a "single-step" method (α_i is corrected once time for all test). Based on Bonferroni's correction, [8] proposed a "stepwise" method (α_i is corrected step-by-step for each test), the Bonferroni-Holm's correction. It examines each hypothesis in an ordered sequence, and the decision to accept or reject the null hypothesis depends on the results of the previous hypothesis tests (beginning with the smallest p-value, and continuing until it fails to reject a null hypothesis). The Bonferroni-Holm's correction is more powerful and less conservative than simple Bonferroni, since with simple Bonferroni, you compare all p-values to α/k. With the Bonferroni-Holm's method, we therefore have more opportunities to reject null hypotheses. However, these approaches are still very conservative. That's why [3] proposed the Benjamini-Hochberg correction, less conservative than above methods, that's based not only on the probability of at least one false positive (cf. FWER), but also on the proportion of false positives among the rejected null hypotheses (cf. FDR). This proportion can be pre-defined expectingly by user.

In the context of multiple test, we expect FDR lower than a global threshold α. This requires to correct each local α_i in order to protect against making a false positive conclusion in each test. Benjamini-Hochberg proposed a FDR method that the true null hypotheses p-values are independent $uniform(0, 1)$ random variables. This is one of the first developed and is widely used method. The procedure of Benjamini-Hochberg's correction is presented in table 1.

Table 1. Benjamini-Hochberg's correction algorithm

Require: A list of *p-value* $P = \{P_1, ..., P_k\}$, k is number of tests and α, significance threshold.
Ensure: A list of indexes of rejected null hypotheses.

1. $listIndex \leftarrow \emptyset$;
2. $P' \leftarrow order(P, ASC)$;
3. $Index \leftarrow getIndex(P', P)$;
4. **for** $i = 1$ to k **do**
5. **if** $P'[i] \leq \alpha * i/k$ **then**
6. $listIndex[i] \leftarrow Index[i]$;
7. **else**
8. Break; {*Stop and fail to reject (accept) any others hypotheses*}
9. **end if**
10. **end for**
11. **return** $listIndex$;

Notations:
$order(P, ASC)$: function returning the list of ascend ordered *p-value* of P
$getIndex(P', P)$: function returning the real indexes of tests according to ordered list of *p-value*

For each local hypothesis H_i, calculate the corresponding p-value P_i from the test statistic. k is the number of null hypotheses which are simultaneously tested Then, we order the p-values $P_1, ..., P_k$ from smallest to largest and the corresponding hypotheses $H_1, ..., H_k$. For a expected FDR, compare the ordered p-value p_i to the critical value $\frac{\alpha * i}{k}$ to reject H_i.

3 Reducing the Number of Tests by Using Quasi Essential Graph

As mentioned above, the number of tests causes the major difficulty to multiple hypothesis testing. To deal with this problem, we have not only to correct the significance threshold α, but also to decrease the number of tests. A naive approach can be applied by considering only the edges found in at least one of two sets of graphs. That means we do not test on $\frac{n(n-1)}{2}$ possible undirected edges, where n is the number of variables. In this section, we introduce another approach that can be used for reducing the number of tests. In fact, in a recent work, we proposed a new object named *quasi essential graph* (QEG) [11] to resume statistically each set of BNs into a "most" representative graph by eliminating noisy edges that have a small probability of occurrence. It helps to reduce the number of tests for the multiple hypothesis testing approach on the edges of graph.

A *quasi essential graph* (V, G, w_u, w_a) is a weighted graph defined by: (1) $V = \{X_1, ..., X_n\}$, a set of discrete random variables; (2) a DAG G, where each node represents a variable from V; (3) a set of weights w_u associated to each (undirected) edge in G skeleton; (4) a set of weights w_a associated to arrows of each directed edge in G.

Given a set of BNs $\mathcal{B}$ and threshold $\beta > 0.5$ (ensuring acyclic), QEG Q is a representative of $\mathcal{B}$ iif: (1) Q has the same set of variables with all BNs; (2) the

Table 2. Multiple hypothesis testing with or without QEG filtering for comparing two sets of BNs : proposed algorithm

Require: Two sets of graphs POP_1, POP_2 and type of approach (with or without QEG).
Ensure: A list of tested edges with their p-values.

1. $List \leftarrow \emptyset$; // list of tested edges with their p-values
2. **if** $(QEG ==' Yes')$ **then**
3. $UG_1 \leftarrow POP2QEG(POP_1)$;
4. $UG_2 \leftarrow POP2QEG(POP_2)$;
5. **else**
6. $UG_1 \leftarrow POP2UG(POP_1)$;
7. $UG_2 \leftarrow POP2UG(POP_2)$;
8. **end if**
9. $UnionUG \leftarrow UG_1 \cup UG_2$;
10. **for** $k = 1$ to $UnionUG.num\_edge()$ **do**
11. $List \leftarrow List.add(UnionUG.edge(k), Fisher(UnionUG.edge(k)))$;
12. **end for**
13. **return** $List$;

Notations:

$POP2QEG(POP_i)$: function returning the QEG of set i

$POP2UG(POP_i)$: function returning the weighted skeleton of set i

$Fisher(UnionUG.edge(k))$: function returning p-value with Fisher's exact test on edge k-th of the graph $UnionUG$

probability of occurrence in $\mathcal{B}$ of each undirected edges of Q is greater than β; (3) the probability of occurrence in $EG(\mathcal{B})$ of each directed edges (arrow) of Q is greater than β.

Our goal is to find the most relevant edges by using the representative QEG of each set of BNs. As mentioned in the section 2.2, in this work, we test only on the occurrence of undirected edges. Therefore, this procedure is realized only on the undirected part of QEG: we construct first a union graph U of the skeleton of two QEG Q_1 and Q_2; Then, for each undirected edge e_i found in U, we compare its weight in the skeleton of Q_1 and Q_2 by calculating $D_i = w_u^i(Q_1, e_i) - w_u^i(Q_2, e_i)$. If $|D_i| \geq 0$, e_i is marked as "relevant" edge.

After the elimination step, we can apply a multiple test (cf. Section 2.2) for all relevant edges on the essential graphs of two sets of BNs. This global algorithm is described in table 2.

4 Experiments

4.1 Generation of Experimental Data

The experimental study was designed on simulated BNs. In order to evaluate the proposed method, we generated different couples of two sets of BNs. For each couple, different configurations of variables can be varied (cf. Table 3).

It is important to note that if we try to fix some variables for example nG, nV, nE or Λ, then $nRndE$ and Δ play important roles to make differences between two sets graphs. This also allows to control the rate of committing errors of the results of tests presented next (cf. Section 4.3).

Table 3. List of variables for graph generation procedure

Name	Definition
nG	Number of graphs per set
nV	Number of vertices per graph
nE	Number of edges per graph
$nRndE$	Number of random edges for adding operations
Δ	Mean of the Poisson distribution used for randomly generating the number of differences between two initial graphs of two sets by edge adding/deleting operations
Λ	Mean of the Poisson distribution used for randomly generating the number of differences between the initial graph and the other graphs in a set by edge adding/deleting operations

In general, each couple consists of two different sets of random DAG. Each graph is constructed randomly by changing (adding/removing) some edges (Λ) from the initial DAG. And the initial DAG of each set must be also different. To ensure this, with each pair of set of DAGs, the initial DAG of the first set can be a random DAG. But, the initial DAG of the second set must be generated randomly from the first one by randomly changing (adding/removing) some edges (Δ). In order to ensure the real (not random) difference between two sets of graphs, we applied only the adding operation of an edge from a "fixed" edge list. That means, in order to generate a new graph from the initial graph, the adding operation must choose one of edges in a fixed list of valid edges. This list can be generated randomly by taking a fixed number of edges ($nRndE$) that excluded the presented edges of the initial graph. In this work, we chose $nG = 100$, $nV = 100$, $nE = 100$, $nRndE = \{10, 20, 50, 100\}$, $\Delta = \{3, 7, 15\}$ and $\Lambda = 5$.

4.2 Experimental Protocol and Result Evaluation Methods

In order to limit the number of tests, we implemented QEG algorithm that allows us to get a union skeleton with only relevant edges of two sets BNs (cf. Table 2). We implemented also Fisher exact test (eqn 1) to verify the independence between relationship of nodes and the source of each set of BNs. As with our simulated data each BN have about 100 edges, we implemented Benjamini-Hochberg's and Bonferroni's α correction to identify the list of null hypotheses can be rejected. These implementations have been coded in C++ with the Boost library (`http://www.boost.org`) and APIs provided by the ProBT platform (`http://bayesian-programming.org`).

Our first objective is to compare the performance and respective interest of Bonferroni's and Benjamini-Horchberg's correction. Then, we need to observe the different behaviors of the multiple test with respect to the choice of the "tested" edges (all edge present in one of both sets or all edges present in one of both representative QEGs).

4.3 Results

Table 4 describes the ability of each testing procedure (without α correction, or with Bonferroni's or Benjamini-Horchberg's correction ; with or without QEG filtering) in several experimental contexts.

Table 4. Multiple test results. **EXP:** Experimental context ($nRndE, \Delta$); **Real:** Number of real differences between the initial graph of each set; **BH:** Benjamini-Horchberg's correction; **BON:** Bonferroni's correction; **tp:** True Positive (reject H_0 when an edge exists in one initial graph but not in the other); **tn1:** True Negative 1 (accept H_0 when an edge exists in both initial graphs); **tn2:** (accept H_0 when an edge does not exist in both initial graphs); **fp1:** (reject H_0 when an edge exists in one initial graph but not in the other); **fp2:** False Positive 1 (reject H_0 when an edge exists in one initial graph but not in the other); **fn:** (accept H_0 when an edge exists in one initial graph but not in the other).

EXP	Real	Without QEG																		N
		without correction						BH						BON						
		tp	tn1	tn2	fp1	fp2	fn	tp	tn1	tn2	fp1	fp2	fn	tp	tn1	tn2	fp1	fp2	fn	
100-15	17	17	89	129	2	30	0	17	91	156	0	3	0	17	91	159	0	0	0	267
50-15	18	18	116	88	5	40	0	18	90	136	0	23	0	18	90	159	0	0	0	267
20-15	19	19	87	4	3	26	0	19	88	6	2	24	0	19	90	14	0	16	0	139
100-7	5	5	95	148	2	22	0	5	97	170	0	0	0	5	97	170	0	0	0	272
50-7	6	6	94	49	3	40	0	6	96	76	1	13	0	6	96	89	1	0	0	192
20-7	9	9	93	1	2	34	0	9	93	3	2	32	0	9	95	13	0	22	0	139
20-3	5	5	93	2	4	35	0	5	96	3	1	34	0	5	97	15	0	22	0	139
10-3	1	1	98	0	1	19	0	1	99	0	0	19	0	1	99	1	0	18	0	119
Total	80	80	765	421	22	246	0	80	750	550	6	148	0	80	755	620	1	78	0	1534

(a) without QEG

EXP	Real	With QEG																		N
		without correction						BH						BON						
		tp	tn1	tn2	fp1	fp2	fn	tp	tn1	tn2	fp1	fp2	fn	tp	tn1	tn2	fp1	fp2	fn	
100-15	17	17	89	0	2	0	0	17	91	0	0	0	0	17	91	0	0	0	0	108
50-15	18	18	116	0	5	0	0	18	90	0	0	0	0	18	90	0	0	0	0	108
20-15	19	19	87	0	3	0	0	19	89	0	1	0	0	19	90	0	0	0	0	109
100-7	5	5	95	0	2	0	0	5	97	0	0	0	0	5	97	0	0	0	0	102
50-7	6	6	94	0	3	0	0	6	97	0	0	0	0	6	97	0	0	0	0	103
20-7	9	9	93	0	2	0	0	9	95	0	0	0	0	9	95	0	0	0	0	104
20-3	5	5	93	0	4	0	0	5	97	0	0	0	0	5	97	0	0	0	0	102
10-3	1	1	98	0	1	0	0	1	99	0	0	0	0	1	99	0	0	0	0	100
Total	80	80	765	0	22	0	0	80	755	0	1	0	0	80	756	0	0	0	0	836

(b) with QEG

We can observe the following general behavior : all the approaches can detect the real differences between the two initial graphs (tp) but differ in their error rates. The "no correction" approach has a very high false positive rate. This value decreases first with BH approach and then with the most conservative BON procedure. Using QEG filtering highly decrease the false positive whatever the testing procedure.

In a recent research, [6] proved that FDR based procedures present a promising alternative to approaches that control the FWER. It is also one of the most used approaches in real application where thousands of tests are performed simultaneously, such as the differentiation of gene expression, etc... Meanwhile, in with our simulated graphs, the experimental results show that Bonferroni's correction (FWER based) outperforms Benjamini-Horchberg's correction on the rate of committing errors (cf. Table 4). This is a normal result of tests where there are many (real) differences between two samples. So, it is easy to reject a null hypothesis whatever the correction methods applied. In fact, because of not

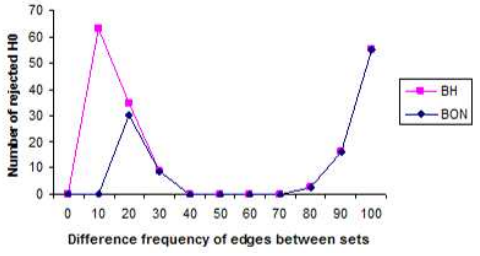


Fig. 1. Comparison between Bonferroni's correction vs. Benjamini-Horchberg's correction (without QEG) on the ability of rejecting null hypothesis according to the difference of edge frequency in the two sets

completely random data generation (cf. Section 4.1), if an edge has been chosen for generating the difference between sets, its frequencies in two sets are really different. That's why p-value for the test on this edge is small and the rate of null hypothesis reject is high.

The experimental results also show that Bonferroni's correction accept more frequently the null hypothesis (no difference) for edges with a small frequency difference between sets ($< \frac{20}{100}$, cf. Figure 1) than Benjamini-Horchberg's one. In fact, smaller $\frac{50}{100}$ is a rate that we are not sure for a difference. This is also the reason why we prefer the tests with QEG that reject all smaller 0.5 probability of occurrence edges. Therefore, the multiple test with QEG reduces a number of useless tests. In fact, Table 4 shows that for 8 experiments, we need apply only totally 836 tests in stead of 1534. Moreover, the tests do not commit almost any errors.

5 Conclusion

We proposed in this work a new approach for the comparison of two sets of graph-based models. This approach is based on the application of multiple test on edges of graph. We described in this paper the solution for dealing with crucial issues cause by the comparison of two sets of Bayesian networks. We also present the quasi essential graph (QEG) that summaries statistically each set of Bayesian networks into a "most" representative graph. The main inspiration of this work is the combination of robustness of ensemble method and statistical significance of QEG.

From this point, this approach have to be extended theoretically and experimentally. We want to compare not only the undirected skeleton but also the directed part of each essential graph and experiment this approach for the real sets of Bayesian networks that are built from a real application.

Acknowledgment. This research has been supported by the grant BIL (Bio-Informatics Ligérienne, Region of Pays-de-la-Loire, France).

References

1. Abdi, H.: Bonferroni and Sidak corrections for multiple comparisons. Encyclopedia of measurement and statistics 1, 103–107 (2007)
2. Auliac, C., d'Alché Buc, F., Frouin, V.: Learning transcriptional regulatory networks with evolutionary algorithms enhanced with niching. In: Masulli, F., Mitra, S., Pasi, G. (eds.) WILF 2007. LNCS (LNAI), vol. 4578, pp. 612–619. Springer, Heidelberg (2007)
3. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society 57(1), 125–133 (1995)
4. Chickering, D.: Learning equivalence classes of Bayesian-network structures. Journal of Machine Learning Research 2, 445–498 (2002)
5. Delaplace, A., Brouard, T., Cardot, H.: Two evolutionary methods for learning bayesian network structures. In: International Conference on Computational Intelligence and Security, pp. 288–297 (2007)
6. Dudoit, S., Shaffer, J.P., Boldrick, J.C.: Multiple hypothesis testing in microarray experiments. Statist. Sci. 18(1), 71–103 (2003)
7. Friedman, N., Goldszmidt, M., Wyner, A.J.: Data analysis with bayesian networks: A bootstrap approach. In: Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence, vol. 24, pp. 206–215. Morgan Kaufmann Publishers, San Francisco (1999)
8. Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6(2), 65–70 (1979)
9. Larranaga, P., Poza, M., Yurramendi, Y., Murga, R.H., Kuijpers, C.M.H.: Structure learning of bayesian networks by genetic algorithms: A performance analysis of control parameters. IEEE Transactions on Pattern Analysis and Machine Intelligence 18, 912–926 (1994)
10. Madigan, D., Andersson, S., Perlman, M., Volinsky, C.: Bayesian model averaging and model selection for markov equivalence classes of acyclic graphs. Communications in Statistics: Theory and Methods 25, 2493–2519 (1996)
11. Nguyen, H.T., Leray, P., Ramstein, G.: Summarizing and visualizing a set of bayesian networks with quasi essential graphs. ASMDA (Applied Stochastic Models and Data Analysis) International Society, Rome, Italy (to appear, 2011)
12. Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research 11, 169–198 (1999)
13. Rodin, A.S., Boerwinkle, E.: Mining genetic epidemiology data with bayesian networks i: Bayesian networks and example application (plasma apoe levels). Bioinformatics 21(15), 3273–3278 (2005)
14. Wang, T., Yang, J.: A heuristic method for learning bayesian networks using discrete particle swarm optimization. Knowl. and Info. Sys. 24, 269–281 (2010)

Financial Performance Analysis of European Banks Using a Fuzzified Self-Organizing Map

Peter Sarlin and Tomas Eklund

Turku Centre for Computer Science – TUCS, Department of Information Technologies,
Åbo Akademi University, Joukahaisenkatu 3-5, 20520 Turku, Finland
{Peter.Sarlin,Tomas.Eklund}@abo.fi

Abstract. Due to the recent wave of bank failures, stress tests have been conducted on banks within the European Union. The stress tests, however, only consider the adequacy of a bank's capital ratios, whereas the general financial performance of individual banks is disregarded. In this paper, we use the Self-Organizing Map (SOM) to perform a visual multidimensional and temporal financial performance analysis of European banks. We address several different problems concerning financial performance analysis. We deal with the problem of selecting suitable financial ratios by performing dimensionality reduction using PCA. We also deal with difficult data using outlier trimming and normalization techniques, and use the SOM for imputing missing values. We use a decision-framework for choosing the final model, based upon a set of map and clustering quality measures. Finally, we implement a second-level fuzzified Ward clustering for visualization purposes and for assessing the crispness of the solution. The result is a visual SOM model for financial performance analysis of European banks.

Keywords: financial performance analysis, European banks, visualization, Self-organizing map, fuzzified clustering.

1 Introduction

Due to recent waves of bank failures in Europe, several stress tests have been conducted on banks within the European Union. Stress tests are important risk management tools used as part of banks' internal risk management assessments and by supervisors to assess the shock absorption capacity of financial institutions and the system in general. In essence, stress tests assess adverse and unexpected outcomes related to a variety of risks, and provide an indication of the absorption sufficiency of the banks' Tier 1 and 2 capital ratios. The stress tests, however, only consider the adequacy of a bank's capital ratio, whereas the general financial performance of individual banks is disregarded.

In today's global and competitive markets, it is important for a bank to recognize its competitors, and to know their strengths and weaknesses. Fortunately, with the rapid advances in information technology, access to financial information about banks has also increased exponentially. Today, through financial databases, such the US Securities and Exchange Commission's Edgar database and commercial financial

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 186–195, 2011.

databases such as Bankscope and Reuters, financial institutions, investors and regulators have access to massive amounts of fairly well-standardized financial information of banks.

However, analyzing the financial performance of banks is not completely unproblematic. First, there are of course comparability issues because of accounting differences. The issues are both differences between US GAAP and IFRS [1] as well as differences in implementation of IFRS [2]. Second, financial performance can be disentangled into several subdimensions, e.g. asset quality, capital, operations and liquidity ratios, and visualizing these multidimensional data both over time and across banks is difficult when using spreadsheets or conventional statistical tools. A further problem is the statistical properties of financial ratios. It has been shown in several studies (e.g., [3–5]) that these ratios are often significantly skewed with large amounts of extreme values, and thus do not exhibit normal distributions. Delicate questions for financial performance analysis are choosing the most representative ratios out of the available database and dealing with correlations and non-linear relations. These traits of financial ratio data make financial performance analysis with conventional tools a difficult process.

Financial performance analysis using the Self-organizing map (SOM), an unsupervised projection and clustering tool, has been proposed in, e.g., [6–10]. In this paper, we use the SOM to perform a financial performance analysis of European banks. When addressing problematic issues regarding financial performance analysis, we follow the approach in [8]. We deal with the problem of selecting suitable financial ratios by performing dimensionality reduction using PCA. The problem of difficult data is addressed with outlier trimming and normalization techniques. We also deal with the issue of incomplete data by using the SOM for imputing missing values. We follow the approach in [11] by presenting a fuzzy membership degree of each data point (node) to each cluster for visualization purposes and for assessing the crispness of the solution. However, instead of using fuzzy *c*-means (FCM) clustering, as in [8,11], the memberships are computed using distances between data points (nodes) and cluster centers. This is particularly interesting as this applies on any crisp clustering and as the SOM aims at preserving neighborhood relations rather than distances between data. Finally, to assess the quality of a SOM, we use a decision-framework for choosing the final model. The aim of the paper is an easily interpretable SOM model for visual performance analysis of European banks. As a complement to stress tests, this model can be used (1) to investigate the performance of a bank in relation to its competitors based on a battery of financial ratios, and (2) to assess mean profiles of banks with, e.g., best and worst shock absorption capacity.

The paper is structured as follows. In Section 2, we present the fuzzified two-level SOM model, the data and the pre-processing. We discuss the training framework in Section 3 and present a visual performance analysis in Section 4. Section 5 concludes.

2 The Two-Level SOM and Fuzzified Ward Clustering

The SOM is a method with simultaneous clustering and projection capabilities first developed by Kohonen [12]. As the SOM algorithm is well-known, we do not present details of it here – for further reference see [13]. The Viscovery SOMine 5.1 package

is used in this study mainly for its superior visual representation. The training process starts with reference vectors ordered based on the two principal components. Then the training algorithm has two steps: (1) finding the best-matching units (BMUs) and (2) adjusting the reference vectors. The first step compares all input data vectors x_j (where j=1,2,…,N) with the network's reference vectors m_i (where i=1,2,…,M) to find the best match m_b:

$$\left\| x_j - m_b \right\| = \min_i \left\| x_j - m_i \right\|. \tag{1}$$

Then the second step adjusts each reference vector m_i with the batch updating algorithm [13, p. 111]:

$$m_i(t+1) = \frac{\sum_{j=1}^{N} h_{ib(j)}(t) x_j}{\sum_{j=1}^{N} h_{ib(j)}(t)}, \tag{2}$$

where t is a discrete time coordinate and $h_{ib(j)}$ a decreasing function of neighborhood radii and time. The rest of the parameters in SOMine are the following: map size (the number of nodes), map format (the ratio of X and Y dimensions), and the training schedule (number of training cycles).

The quality of the map can, for example, be measured in terms of quantization error or a distortion measure (see e.g. [14]). For its simplicity, as is below stated, the QE is used in this paper. It represents the fit of the map to the data, i.e., an average of the distances between all input vectors x_j and their corresponding BMUs m_b.

The nodes of the map can further be divided into clusters of similar nodes. In our case, the clusters and their centers are computed using Ward's [16] hierarchical clustering. The following modified Ward's criterion is used as a basis for measuring the distance between two candidate clusters:

$$d_{kl} = \begin{cases} \dfrac{n_k n_l}{n_k + n_l} \cdot \left\| c_k - c_l \right\|^2 & \text{if } k \text{ and } l \text{ are adjacent} \\ \infty & \text{otherwise} \end{cases}, \tag{3}$$

where k and l represent clusters, n_k and n_l the cardinality of clusters k and l, and $\left\| c_k - c_l \right\|^2$ the squared Euclidean distance between the cluster centers of clusters k and l, and the distance between non-adjacent clusters is infinite. When clusters k and l are merged to cluster h, the cardinality is the sum of the cardinalities of k and l and the centroid the mean of c_k and c_l weighted by their cardinalities.

In addition to dividing the nodes into crisp clusters, as is commonly done, we follow the approach in [11] by also presenting a fuzzy membership degree of each node to each cluster. However, instead of using fuzzy c-means clustering and applying it on the nodes, the membership degrees are computed using Euclidean distances between data points and the centroids of the crisp clusters. For this, any crisp clustering method, as appropriate, is applicable. This approach resembles that in [15],

but differs by being implemented on a second-level clustering instead of directly on the nodes, by not assuming inverse exponential distances, and by introducing a fuzzification parameter. As we also implement the fuzzification on the nodes, it can as well be used for assessing the topological ordering of the grid. The crisp clustering is fuzzified by computing the inverse distance between data point x_j (or each reference vector m_i) and each cluster center c_k (where $k = 1,2,\ldots,C$):

$$u_{jk} = \frac{1}{\left\| x_j - c_k \right\|^{\frac{2}{\mu-1}}}, \tag{4}$$

where $\mu \in (1,\infty)$ is the fuzzy exponent (i.e., the fuzzifier) which controls the extent of overlapping between the clusters. However, we normalize the similarity matrix u_{jk} to the following cluster membership matrix for each node:

$$v_{jk} = \frac{u_{jk}}{\sum_{k=1}^{C} u_{jk}}. \tag{5}$$

to fulfill the probabilistic constraint $\sum_{k=1}^{C} v_{jk} = 1$. For equal weighting of the distances for each variable, this is of course performed on normalized data. The extent of overlapping between the clusters is set by the fuzzy exponent μ. When $\mu \to 1$, the fuzzy clustering converges to a crisp clustering, while when $\mu \to \infty$ the cluster centers tend towards the center of the data set. $\mu = 2$ and $\mu = 3$ can be seen as benchmarks, since they give squared and simple Euclidean distances.

2.1 Data and Pre-processing

The dataset consists of annual financial figures for banks from European countries, and is retrieved from the Bankscope financial database. The set consisted of an initial total of 1,236 banks for the period from 1992:12–2008:12. From this set we include only banks from the European Union for a more homogeneous set of banks.

We chose to use all financial ratios provided in the database. From the initial set of 38 ratios, only 24 were selected because of limited data availability for the remaining ratios. Ratios with more than 25% missing data and banks with more than 1/3 of missing values for a given year were removed. Finally, we were left with a resulting 9,655 rows of data, and a total of 855 banks.

A major problem in choosing financial ratios is selecting the correct combination to cover the appropriate classes of financial performance. Among the 24 ratios selected, there was a great deal of correlation, yet choosing which ones to drop is a delicate question. Therefore, PCA was used for dimension reduction. This enabled reducing the dimensionality of the problem, and thus increasing interpretability, without losing explainability.

However, although the SOM is tolerant to missing values, as it allows training by solely considering the indicators that are available (see e.g. [17]), PCA requires complete data. In order to train with and map data rows with missing data, it was thus

necessary to first impute missing values for the PCA extraction. Following [8], we chose to use the SOM for imputing missing values. Since the SOM allows projection of data rows onto a map by solely considering the indicators that are available, the simplest way is to impute the missing from the BMU. The data rows with no missing values were used for training the SOM model, whereafter the missing values are imputed from the BMU of each data vector.

PCA reduces dimensions, but neither removes outliers nor normalizes the data. Moreover, although the SOM is tolerant towards outliers (see e.g. [13]) they still affect the interpretability of the map greatly, and thus a method for outlier removal was required. In this case, in order to not lose significant amounts of data, a trimming with replacement of outlier data was performed. The replacement was based upon modified boxplots, where extreme values were defined as those outside 1.5 times the interquartile range (Q3-Q1). Instead of eliminating values, we used replacement with the trimmed minimum or maximum value, as is typical in Winsorizing, a trimming method that replaces a specified percentile from each tail. The modified boxplot is preferred over Winsorizing since the boxplot accounts for the particular distribution of each variable. The trimming resulted in replacement of a total of 7.39% of the data, distributed as needed per variable and tail.

Using PCA, seven principal components (PC1–7) with eigenvalues greater than 1, which explain 84.44% of the variability in the data, were then extracted from the 24 variables.[1] From studying the component loadings, we were able to identify that the components reflect the following dimensions: *capital ratios* (PC1), *loan ratios* (PC2), *profitability* (PC3), *interest revenue (PC4), non-operating items (PC5), subordinated debt (PC6)* and *loan loss provisions (PC7).* Using the varimax rotated principal component loadings, the 24 variables were transformed to PCA scores. In general, high values are better, with the exception of loan ratios, which reflect the inverse of capital to net loans and the liquidity of the bank's assets, and loan loss provisions, which measure the ratio of risk to interest rate margins. However, since the set of banks is not completely homogeneous, the optimal level of the indicators may vary depending on the nature of the bank. Note also that PC6 actually reflects the inverse of the percentual subordinated debt, and thus high values are better. Finally, as is needed for equal weighting for the SOM, columnwise normalization by variance was used for standardized weighting of inputs.

3 Training

For evaluating maps trained with different parameter values, we mainly follow the decision-framework developed in [8]. However, the first decision to be taken, the number of clusters, differs from that study. Instead of using heuristics for finding the optimum number of clusters, we simply assign the number of clusters equal to the number of principal components, i.e., seven. We use a set of measures to gauge the quality of the map. First, we measure the clustering tendency and span over the map by computing how many of the seven clusters attract at least one max or min value of the supposedly uncorrelated principal components. Second, the fit of the maps to the

[1] The PCA matrix and definitions of financial ratios are available from the authors on request.

data distribution is assessed based on quantization error (QE). For the set of experiments, the maps with QE in the 50th percentile are evaluated as accurate. Percentiles are preferred over absolute thresholds, since the measures depend on the used data sample. Third, the topographic ordering of the maps is evaluated using Sammon's mapping [18], which is a non-linear mapping from a high-dimensional input space to a two-dimensional plane. We use Sammon's mapping to visualize the nodes on 2D and 3D planes. A map is defined to have an adequate topographic ordering if it does not have several non-adjacent nodes as neighbors in the data space and is not twisted at any point. Fourth, the maps are evaluated based on visual interpretability for a decision maker, since the choice of the best map for visual analysis is in the end a subjective task. To sum up, the criteria are the following:

1. Quantization error in the 50^{th} percentile
2. The number of clusters that attract max and min values in the 50^{th} percentile
3. Topographic ordering
4. Subjective interpretability.

The constructed map is trained using 9,655 rows of data with a dimensionality of 7 – one for each principal component. Although the Tier 1 capital ratio and the total capital ratio (Tier 1 + Tier 2) were dropped due to the share of missing values, we include their distribution on the trained map for assessing the shock absorption capacity. During the course of the experiment, several maps were trained using different parameter values (tension, cycles of training, number of clusters, number of nodes and map format). In the final experimental stage, a set of parameters are, however, kept constant. The map format is chosen to be 75:100, since [19] recommends that the map ought to be rectangular in order to achieve a stable orientation in the data space. The number of clusters is, as argued above, chosen to be seven. The parameters that have been varied are tension and number of nodes. We train models for $M \in \{50,100,150,200,250,300,400,500,600,1000\}$ and $\sigma \in \{0.0001,0.3,0.5,0.75,1.0,1.5,2.0\}$; however, since our evaluation results depend on the rank in the sample, we exclude parameter values which for certain provide useless results. Since small tension values consistently result in unordered maps, we start from 0.5 and go to a maximum tension value of 2. The number of nodes is defined to range from 100 to 300, since we assess that neither smaller nor larger maps are adequate for our purpose.

The SOM experiments are shown in Table 1, where the fulfillment of the above quality measures are shown with X marks. Obviously, increasing the number of nodes leads to higher quantization accuracy but poorer topographic ordering, while increasing the tension value leads to poorer quantization accuracy but better topographic ordering. The interpretability measure shows that the best maps are obtained with average values for both parameters, while there is no clear pattern between the parameters and the clustering tendency measure. Based on the above decision framework, we chose the four best maps for in-depth analysis. These models were not only compared in terms of real decision-making relevance, but also thoroughly tested for twistedness with Ward clustering, Sammon's mapping, and by evaluating the topological location of BMUs for each node. We finally chose a model with 137 nodes on a 13x11 grid and a tension of 0.7. However, differences between the quality of these maps were small.

Table 1. Results of the SOM experiments. Top 4 models are shown with a grey highlight, the chosen map is shown in bold and X marks in columns 1–4 indicate quality measure fulfillment.

Tension / No.of nodes	0.5				0.75				1				1.5				2			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
100 (85)	X		X				X	X		X	X	X			X	X			X	
150 (137)	X		X		**X**	**X**	**X**	**X**		X	X	X			X	X			X	
200 (188)	X	X			X	X	X	X	X	X	X	X			X	X			X	
250 (247)	X				X				X		X	X			X	X			X	
300 (331)	X	X			X	X			X	X		X	X	X	X	X			X	

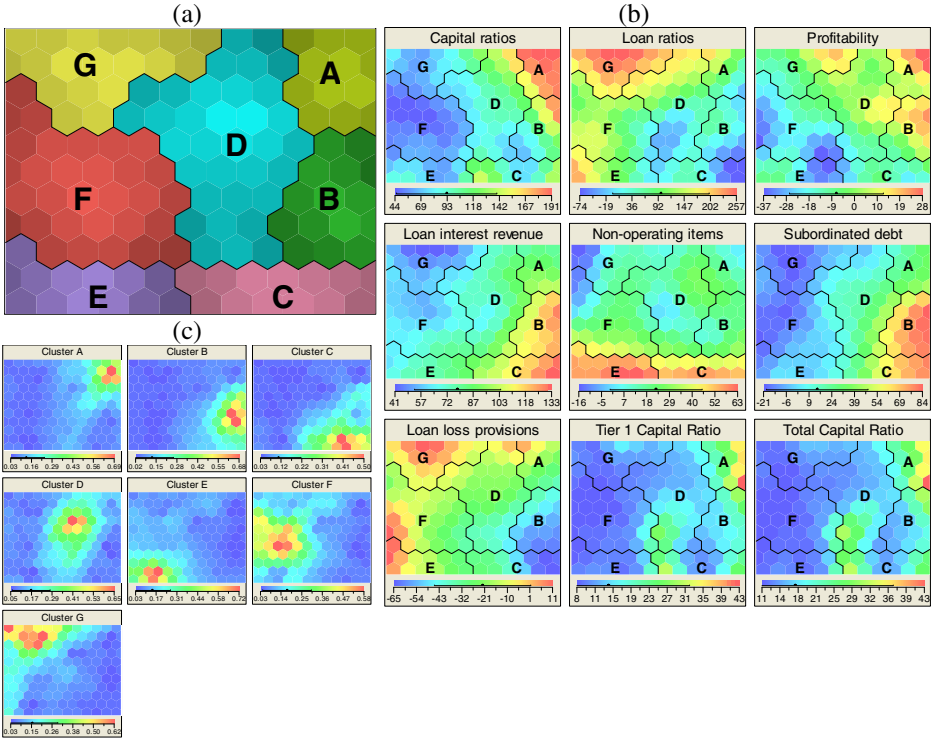


Fig. 1. The SOM grid (a), its feature planes (b) and the nodes' cluster membership degrees (c)

The crisp clustering of the above SOM model was fuzzified using Euclidean distances. The fuzzifier μ was tested for values between 1 and 10. Based upon these experiments, a benchmark μ-value of 2.0 provided an adequate fuzzification of the map. It introduces a fuzziness degree large enough to show relationships between clusters, but not large enough to completely eliminate cluster borders.

As the map describes a multidimensional space on a two-dimensional grid, all information on the map cannot be visualized in two dimensions. To further enhance the visualization, the variables are separately shown on their own grids. Each feature plane displays the distribution of that principal component on the map, with cold colors (blue) indicating low values and warm (red) indicating high values. As the

feature planes are different views of the same map, one unique point represents the same node on all planes. Thereby, the characteristics of the SOM model can be identified by studying the underlying feature planes (Fig. 1a and 1b). The clusters in Fig. 1a and the contours on each plane show the crisp Ward clustering, while Fig. 1c visualizes the nodes' cluster membership degrees.

The banks in clusters A, B and C can be seen as the best in class performers. Cluster A is the best cluster with high capital ratios, loan ratios, profitability and loan loss provisions and average values for the rest, while Cluster B differs by showing lower loan ratios and loan loss provisions and higher loan interest revenue, subordinated debt and Tier 1 and 2 ratios. Cluster C is similar to cluster B, but with slightly lower capital ratios, profitability, loan loss provisions and Tier 1 and 2 ratios and somewhat higher loan interest revenues, non-operating items and subordinated debt. The banks in cluster D can, in general, be seen as average performers, as they show average values for all variables. The banks in cluster E, F and G can be seen as the poorest groups, with low values for capital ratios, profitability, loan interest revenue, subordinated debt and Tier 1 and 2 ratios, and high values for loan ratios and loan loss provisions. Cluster E differs from the rest by showing lower loan ratios and loan loss provisions, and cluster G by showing higher loan ratios and profitability and lower loan interest revenue and non-operating items. Cluster F can be seen as the poorest performer by being an average of clusters E and G, except for lower values for capital ratios, loan ratios, profitability and loan loss provisions.

4 Results and Analysis

In this section, we perform a visual financial performance analysis for the largest bank in five Western European countries. We project data onto the map using Eq. 1 and connect consecutive data points using trajectories.

Fig. 2 and 3 show trajectories and cluster membership degrees for the top five Western European banks. Fig. 2 shows that, during the period from 2002–08, both

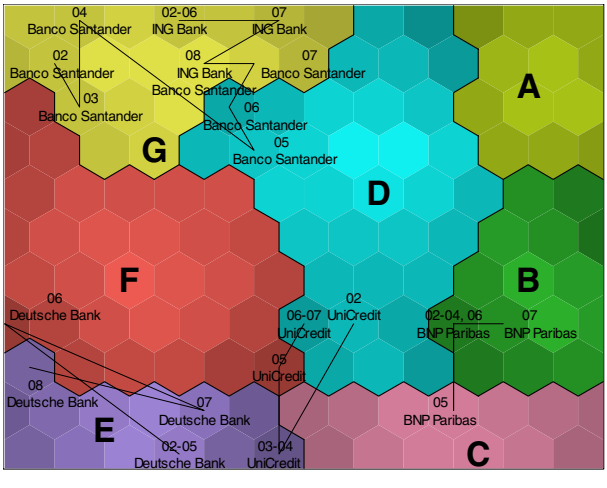


Fig. 2. A visual financial benchmarking of the top five largest European banks from 2002–08

Spanish Banco Santander and Dutch ING Bank hovered on the border between the poor cluster G and average cluster D. As can be seen from the feature planes in Fig. 1c, a movement from cluster G to cluster D indicates an improvement in all measures. Fig. 3 shows that the cluster memberships for Santander switch between G and D, but that ING's distance to D is persistently high. Even though ING's performance is stable, it displays currently a trend towards the poor cluster G. French BNP Paribas, on the other hand, exhibits best in class performance during the period, by moving between clusters C and B with a trend towards B. During the entire period, German DB is projected into the poor clusters E and F with small variations over time. Fig. 2 shows that UniCredit Banca is projected in between BNP and Deutsche Bank and varies on the border of cluster C, D, E and F. The membership degrees in Fig. 3 confirm that the company resembles all four clusters to a close to equal extent ($\max(v)=0.2$).

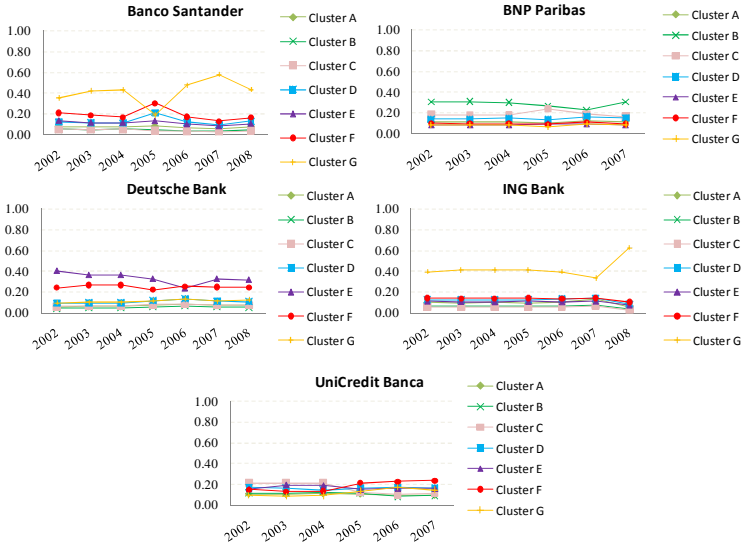


Fig. 3. Cluster memberships for the top five European banks during 2002–09

5 Conclusions

In this paper, we have used the SOM to perform a multidimensional and temporal visual financial performance analysis of European banks. Following the approach in [8], we deal with problematic issues when performing financial performance analysis, including how to deal with difficult data while preserving variability and relationships, how to select the battery of financial ratios, and how to assess the validity of the clustering solution. We have also introduced an approach for fuzzifying any crisp second-level clustering of the SOM. The result is a visual SOM model for financial performance analysis of banks.

References

1. Van der Meulen, S., Gaeremynck, A., Willekens, M.: Attribute differences between U.S. GAAP and IFRS earnings: An exploratory study. International Journal of Accounting 42(2), 123–142 (2007)
2. Nobes, C.: The survival of international differences under IFRS: towards a research agenda. Accounting and Business Research 36(3), 233–245 (2006)
3. Lev, B.: Financial Statement Analysis: A New Approach. Prentice-Hall Inc., N.J (1974)
4. Foster, G.: Financial Statement Analysis. Prentice-Hall, Inc., Englewood Cliffs (1978)
5. Deakin, E.B.: Distributions of Financial Accounting Ratios: Some Empirical Evidence. The Accounting Review 51, 90–96 (1976)
6. Back, B., Sere, K., Vanharanta, H.: Managing Complexity in Large Data Bases using Self-Organizing Maps. Accounting Management and Information Technologies 8(4), 191–210 (1998)
7. Eklund, T., Back, B., Vanharanta, H., Visa, A.: Using the Self-Organizing Map as a Visualization Tool in Financial Benchmarking. Information Visualization 2(3), 171–181 (2003)
8. Eklund, T., Sarlin, P., Jokipii, A.: Visual Financial Benchmarking using the Self-Organizing Map: A Revisit. Turku Centre for Computer Science Technical Report, No. 1009 (2011)
9. Sarlin, P.: Visual monitoring of financial stability with a self-organizing neural network. In: 10th IEEE International Conference on Intelligent Systems Design and Applications, pp. 248–253. IEEE Press, Cairo (2010)
10. Resta, M.: Early Warning Systems: an approach via Self Organizing Maps with applications to emergent markets. In: 18th Italian Workshop on Neural Networks, pp. 176–184. IOS Press, Amsterdam (2009)
11. Sarlin, P., Eklund, T.: Fuzzy Clustering of the Self-Organizing Map: Some Applications on Financial Time Series. In: 8th International Workshop on Self-Organizing Maps, pp. 40–50. Springer, Helsinki (2011)
12. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 66, 59–69 (1982)
13. Kohonen, T.: Self-Organizing Maps. Springer, Berlin (2001)
14. Vesanto, J., Sulkava, M., Hollmén, J.: On the decomposition of the self-organizing map distortion measure. In: International Workshop on Self-Organizing Maps, pp. 11–16. Springer, Hibikino (2003)
15. Cottrell, M., Letrémy, P.: Missing values: Processing with the kohonen algorithm. In: Applied Stochastic Models and Data Analysis, pp. 489–496, Brest (2005)
16. Ward, J.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236–244 (1963)
17. Samad, T., Harp, S.A.: Self-organization with partial data. Network: Computation in Neural Systems 3, 205–212 (1992)
18. Sammon, J.W.: A Non-Linear Mapping for Data Structure Analysis. IEEE Transactions on Computers 18(5), 401–409 (1969)
19. Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J.: SOM_PAK: The Self-Organizing Map Program Package. Technical Report, Helsinki University of Technology (1996)

Self-Organizing Map in Process Visualization

Miki Sirola and Jaakko Talonen

Aalto University, Department of Information and Computer Science, Finland

Abstract. Our research group is studying data-analysis based techniques in decision support and visualization. We have had co-operation with a Finnish nuclear power plant Olkiluoto within a long industrial research project. We have developed many decision support schemes and visualizations based on self-organizing map (SOM) method combined with other methodologies. In this paper, we discuss about SOM method in the process visualization of dynamic systems. With a case example produced with the Olkiluoto plant data we show the information value of this method. Some comparisons to other methodologies are made and the assessment of the information value and the definition of the assessment criteria are discussed. The measurement of the information value is a challenging task.

Keywords: Self-Organizing Map, Data Analysis, Visualization, Neural Methods.

1 Introduction

The visualization in process industry is a complex issue. The need for presenting the information content in the control rooms have changed along the developing technology and with time. In nuclear industry, many modernization projects have been carried out. For instance, wide monitoring screens set up many new requirements for the presentation techniques.

Early fault detection is an important research issue in the nuclear industry. The earlier the abnormal behavior in the process is detected, the better possibilities there are to identify the problem in time and handle the recovery procedure properly. We have developed tools for helping operators in their work, and to help experts to understand better different phenomena in the process [1].

Prototyping has been one research methodology used in our research group. In many prototypes a neural method self-organizing map is used and combined with other more or less traditional methods [1]. We have also done traditional data analysis with nuclear power plant data and training simulator data, and developed methods and tools for helping decision support in the nuclear field. Visualization is an important part of this research. Many tools and methods could be easily generalized or modified to cover other application areas as well.

Process failure detection with complex data analysis methods is a widely studied research area. Also about process presentation and visualization other studies are made. For instance, in nuclear field [2] and other industrial branches [3], [4] many techniques have been developed. Decision support visualizations [5], [6] are also found from the literature.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 196–202, 2011.

In this paper we study the use of the self-organizing map in visualization of process data. Also SOM method in dynamic systems is discussed. In a case example with the Olkiluoto nuclear power plant data, we show the information value of the method.

The structure of the paper is the following. In Section 1, we introduce the problem, view a little the background of the topic, make a quick look into the related studies, and introduce shortly the paper structure. In Section 2, we introduce the Self-Organizing Map (SOM) method and show with an example the possibilities to use the method in dynamic systems. In Section 3, we introduce in detail an industrial case example, where the SOM method is used with real data. In Section 4, we discuss the methods and make some comparisons. In Section 5, we summarize the paper, make conclusions and figure out possible future work.

2 SOM Method in Dynamic Systems

Self-organizing map (SOM) is an effective method in neural computing for the analysis and visualization of multidimensional data. The SOM algorithm [7] resembles vector quantization (VQ) algorithms. The difference with regard to VQ techniques is that the neurons are organized on a regular grid and along with the selected neurons also its neighbours are updated. The SOM performs an ordering of the neurons. The SOM is a multidimensional scaling method projecting data from input space to a lower, typically 2-dimensional output space.

A SOM consists of neurons organized in an array. The number of neurons may vary. Each neuron is represented by an n-dimensional weight vector, $m = [m_1, \ldots, m_n]$, where n is equal to the dimension of the input vector. The neurons are connected to adjacent neurons by a neighbourhood relation, which defines the structure of the map. Rectangular and hexagonal neighbourhoods are the most used topologies.

The SOM is trained iteratively. In each training step, one sample vector x from the input data set is chosen randomly and the distance between it and all the weight vectors of the SOM are calculated using some distance measure. The neuron c whose weight vector is closest to the input vector x is called the Best-Matching Unit (BMU):

$$\|x - m_c\| = \min_i \{\|x - m_i\|\} \tag{1}$$

where $\| \cdot \|$ is the distance measure.

Since BMU is found, the weigh vectors of SOM are updated so that the BMU is moved closer to the input vector in the input space. The topological neighbours of the BMU are treated in a similar way. The adaptation procedure stretches the BMU and its topological neighbours toward the sample vector. The SOM update rule for the weight vector of the unit i is:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \tag{2}$$

where t is time. The x(t) is the input vector randomly drawn from the input data set t and $h_{ci}(t)$ the neighbourhood kernel around the winner unit c at time t. The neighbourhood kernel is a non-increasing function of time and the distance of unit i from the winner unit c. It defines the region influence that the input sample has on the SOM.

Originally the SOM algorithm was not designed for changing time. The SOM is able to analyze ideally only static data sets. Many attempts to use the SOM method in the analysis of dynamic data have been done. It has been used in many time-related problems especially in process modelling and monitoring. These issues are discussed for instance in [8].

One possibility to describe dynamical behaviour is the visualization of trajectories, which link together the adjacent winner neurons (BMU) in the SOM grid. The SOM trajectories have such features as linked BMUs, where each BMU represents a certain instant of time. The operator can learn to adjust the control variables according to the visual impression so that the process stays in the desired regions of the map.

An example of using trajectory expression in a dynamic system is in Figure 1. Here the trajectory of the U-matrix shows visually how an imaginary accident scenario proceeds in a nuclear power plant. The data come from the Finnish Olkiluoto nuclear power plant training simulator. In normal operation the trajectory stays in a certain region in the U-matrix, but when the transient becomes big enough the trajectory moves out to another region. In the example of Figure 1 there is a leak in the main circulation. Different scenarios are somewhat separable in the U-matrix [9].

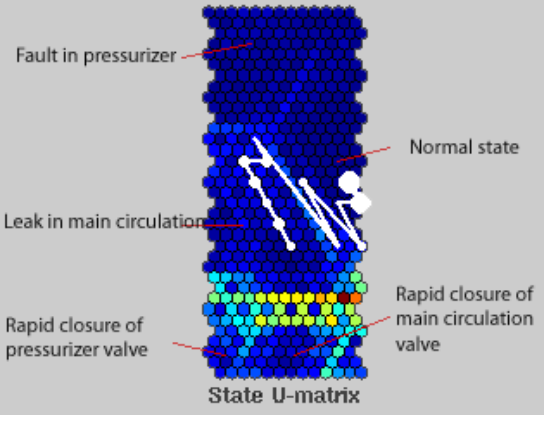


Fig. 1. Dynamical behavior in the process is show by U-matrix trajectory

More examples about handling spatio-temporal problems with the SOM method are written in [1]. In the next section we go through an industrial case example where the SOM method is used with data from a Finnish nuclear power plant.

3 A Case Example with Industrial Data

In our newest research, real data from the reactor unit 1 of Olkiluoto nuclear power plant (NPP) is used. In April 2009, more than 700 signals were stored, every tenth second. In a six-hour period, a change in a valve position was performed. Changes of the process signals in the reheater section and other parts of the NPP were captured in

the recorded data. The position of the control valve at the reheater was changed. At 8 - 10 p.m. process was controlled manually, at 10 - 12 p.m. after the first part of the measurements the process was stabilized. Then the control valve was opened for two hours.

In this example it is shown how to use the SOM method to observe changes between the process signals. Which are the signal values in each state? Which signals depend on the others? In the variable selection phase all signals from the reheater section were selected, totally 125 signals. In our visualization eleven signals were selected from the reheater area and three were selected elsewhere for the analysis, see Table 1.

Table 1. Explanations for the signal measurements and their units

'412K571'	'temperature after 412T23'	'C'
'412K584'	'steam temperature before V5'	'C'
'412K576'	'412T22 emergency dump V22'	'C'
'412K466'	'level 412T15'	'%'
'412K720'	'V312 position'	'%'
'412K301'	'412T9 and 445T302 flow'	'kg/s'
'412K490'	'level 412T15'	'%'
'412K188'	'412T11 pressure'	'BAR G'
'412K705'	'V41 position'	'%'
'412K517D'	'rate of change 412E1 phase 2'	'C/MIN'
'412K513D'	'rate of change 412E2 phase 1'	'C/MIN'
'431K457'	'condenser 431E1 level'	'm'
'431K551'	'condenser 431E1 temperature'	'C'
'413V501xM'	'average of piston positions'	'%'

The control valve test in the reheater part did not affect to the reactor pressure and steam flows. They are situated before the reheater and many other process parts. The condenser is located after the reheater and before the reactor. Three signals shown in the last rows of Table 1 were selected, because signal measurements are after the reheater. Last signal is derived from the redundant measurement. Four measurements are averaged. Next step in the analysis is the visual inspection of the U-matrix, component planes and labeling, see Fig.2

From the U-matrix visualization, it can be seen that there are essentially three clusters (process was controlled manually, the stabilization period and the control valve is open). The component planes show the limits for current process signal values. Also '412K571', '412K584' and '412K576' in the reheater part have high linear correlation with the condenser and the vacuum system part signal '431K551'. Other interesting remark is that another variable '431K457' from this area has the highest values in the end of the stabilization period. In other words, the level of the condenser is the highest after four hours the experiments were started, although the highest temperature was detected at the end of the experiments.

More exact analysis can be done by the principal component projection, see Fig. 3. For example, '412K466' and '412K490' get higher values when the control valve is open than when it is controlled manually.

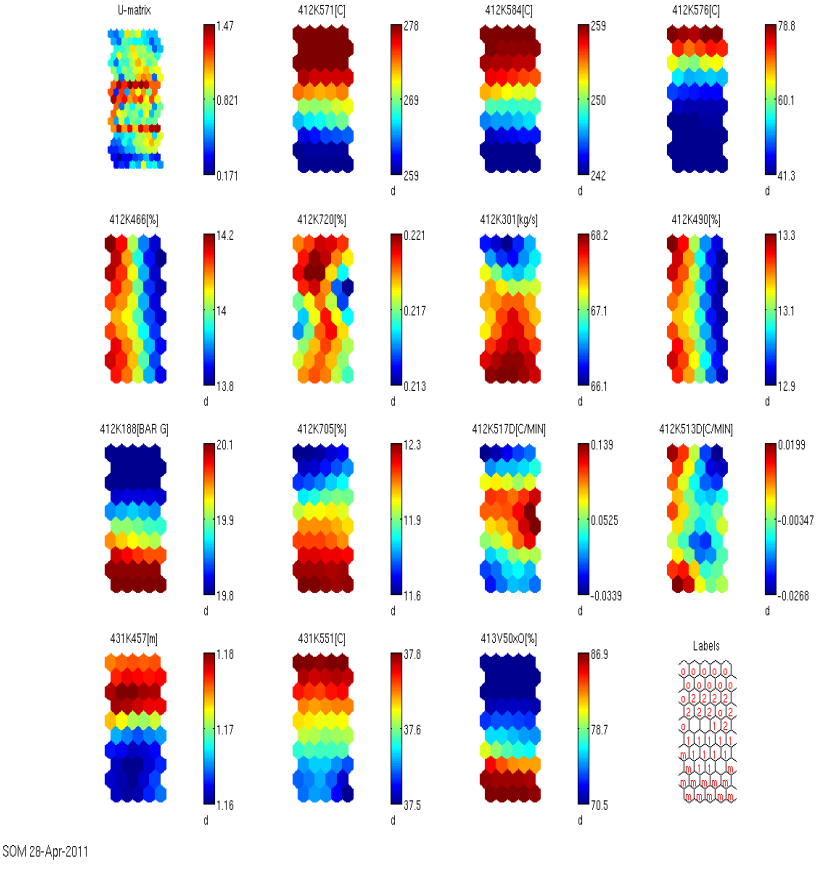


Fig. 2. Case example. In this case 14 process signals were monitored. In the labels, m: the process was controlled manually, o: the control valve is open. 1: the first hour of stabilization period and 2: the second hour of the stabilization period. The U-matrix reveals that the process states 'm' and 'o' are clustered very clearly. The SOM component planes show the values for each process state.

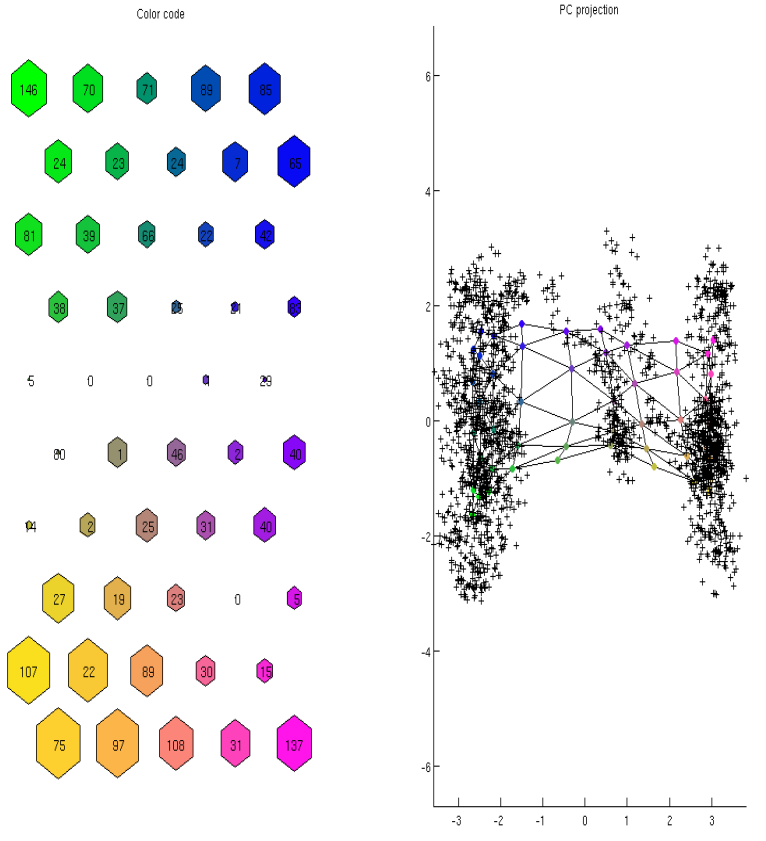


Fig. 3. A principal component projection (PCP). The color map is another type of visualization for U-matrix. Distance matrix information is shown as zero hits for each part of the map (number inside the object). The longer distances are visualized by smaller objects. From the PCP three different clusters can be detected. The first and the last points of the stabilization cluster are situated near to the other two clusters.

4 Discussion

Comparing the SOM method with the PCA (Principal Component Analysis) [10] method it can be noticed that with the SOM method the non-linear behavior is seen better than with the PCA method, which is able to show only the linear dependence.

With the SOM method it is possible to see the dynamical development of the process by using the U-matrix trajectories, and the clustering structure of the data with the U-matrix itself [8]. The correlations of certain variables are very clearly seen with the component plane SOM maps. The faulty development in the data is easy to detect for instance with the quantization error.

The shape of the SOM map can also reveal important things about the distribution of the data, if the shape of the map is not restricted or prohibited. Detecting the pre-stage of the fault is possible with various ways [1]. The visualization of the process

and its progression with SOM maps, and leak detection with an adaptive process model are also studied in [1].

5 Conclusion

We have shown with a case example by using industrial data the information value of the self-organizing map in the process visualization. With some verbal comparisons, we have tried to differentiate this method from some other commonly used methods. The information value can be clearly seen, although the use of this method in a real control room would need special attention and capabilities from the operators. The operator training would therefore meet new challenges.

The measurement of the information value with some concrete way is a very difficult task. We are writing another paper, where the user-interface assessment and visualization assessment are studied more in detail [11]. Some useful criteria can be found to estimate these values. One branch in this field is the psychological studies.

There are a lot of open questions and demanding challenges for studying this issue further in the future. The SOM method alone is not enough for finding out all necessary information out of the process, but it can add additional information value compared with some more traditional methods. The best results can be achieved by using many different methodologies in a well-selected combination.

References

1. Sirola, M., Talonen, J., Parviainen, J., Lampi, G.: Decision support with data-analysis methods in a nuclear power plant. TKK Reports in Information and Computer Science, TKK-ICS-R29, Espoo (2010)
2. Paulsen, J.: Design of process displays based on risk analysis techniques. PhD thesis, Technical University of Denmark and Risø National Laboratory, Roskilde (2004)
3. Vesanto, J.: Data explaration process based on the self-organizing map. PhD thesis, Helsinki University of Technology (2002)
4. Laine, S.: Using visualization, variable selection and feature extraction to learn from industrial data. PhD thesis, Helsinki University of Technology (2003)
5. Kazancioglu, E., Platts, K., Caldwell, P.: Visualization and visual modeling for strategic analysis and problem solving. In: Proceedings of International Conference on Information Visualization (IV 2005). IEEE, Los Alamitos (2005)
6. Kwon, O., Kim, K.-Y., Lee, K.C.: MM-DSS: integrating multimedia and decision-making knowledge in decision support systems. Expert Systems with Applications (2006)
7. Kohonen, T.: The self-organizing map. Springer, Heidelberg (1995)
8. Barreto, G., Araujo, A., Ritter, H.: Time in self-organizing maps: an overview of models. International Journal of Computer Research (2001)
9. Sirola, M., Lampi, G., Parviainen, J.: Failure detection and separation in SOM based decision support. In: Workshop on Self-Organizing Maps (WSOM), Bielefeld (2007)
10. Hair, J., Anderson, R., Tatham, R., Black, W.: Multivariate data analysis, 5th edn. Prentice-Hall, Englewood Cliffs (1998)
11. Sirola, M., Talonen, J.: New visualization techniques and their assessment (to be published in 2011)

Kalman Filter vs. Particle Filter in Improving K-NN Indoor Positioning

Jaegeol Yim[1], Jinseog Kim[2], Gyeyoung Lee[1], and Kyubark Shim[2]

[1] Dept. of Computer Engineering, Dongguk University
Gyeongju, Korea
{yim,lky}@dongguk.ac.kr
[2] Dept. of Statistics and Information Science, Dongguk University
Gyeongju, Korea
jinseog.kim@gmail.com, shim@dongguk.ac.kr

Abstract. The Kalman filter has been widely used in estimating the state of a process and it is well known that no other algorithm can out-perform it if the assumptions of the Kalman filter hold. For a non-Gaussian estimation problem, both the extended Kalman filter and particle filter have been widely used. However, no one has performed comparison test of them. In the consequence, they arbitrarily choose one of them and apply it on their estimation process. Therefore, we have compared the performance of the Kalman filter against the performance of the particle filter. One of the practical fields on which these filters have been applied is indoor positioning. As the techniques of manufacturing mobile terminals have made a big progress, the demand for LBS (location based services) also has rapidly grown. One of the key techniques for LBS is positioning, or determining the location of the mobile terminal. Outdoor positioning is not a big burden to system developers because GPS (Global Positioning System) provides pretty accurate location information of a mobile terminal if the line of sight is not blocked. On the contrary, there is no practical solution for the indoor positioning problem. We can obtain exact location of a mobile terminal if we invest large amount of money, but this is economically not practical. One of the most practical candidate solutions for the indoor positioning problem is the WLAN (Wireless Local Area Network) based positioning methods because they do not require any special devices dedicated for indoor positioning. One of the most significant shortcomings of them is inaccuracy due to the noise on measured data. In order to improve the accuracy of WLAN based indoor positioning, both the Kalman filter and the particle filter processes have been applied on the measurements. This paper introduces our experimental results of comparing the Kalman filter and the particle filter processes in improving the accuracy of WLAN based indoor positioning so that indoor LBS developers can choose appropriate one for their applications.

Keywords: Indoor Positioning, K-NN, Kalman filter, Particle filter, fingerprinting method.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 203–213, 2011.

1 Introduction

A location based service (LBS) provides very useful services such as navigation, troops control, logistics, and so on to the users based on the geographic locations of the users and items. Therefore, positioning users and items is the most essential technique in development of a location-based service system [1]. There are numerous positioning systems including GPS [2], wide-area cellular-based systems [3], infrared-based system [4], radio frequency (RF) + ultrasonic-based systems [5,6], physical contact systems [7], various computer vision systems [8], and RF based systems [9,10].

LBS are so useful that providing indoor LBS is very desirable. For instance, many huge buildings in metropolitan area, large scale companies, factories, universities, huge (underground) shopping malls, and so on are especially demanding LBS. Among the existing positioning systems, wireless local area network (WLAN) based positioning techniques are most interesting in the field of indoor LBS because of the following reasons. GPS signal is not available inside of a building so GPS system cannot be an indoor positioning and the others require special equipments dedicated for positioning. On the other hand, WLAN positioning systems do not require additional hardware dedicated for positioning and wireless network is being serviced everywhere including college campuses, airports, hotels and homes, so it is more economical and less time consuming than other indoor positioning techniques.

A WLAN-based positioning system determines a user's position referring to the received signal strengths (RSS) of the signals from access points (APs). However, it has a serious shortcoming. That is, RSS is influenced by so many environmental parameters and even K-NN (Nearest Neighbor) method which is the most accurate method of RSS based positioning is not accurate enough to be used in a practical application system.

Kalman filter and particle filter are good candidate tools in dealing with such noisy data as RSS. In fact, [11,12] has already introduced extended Kalman filter method for indoor positioning and [13] has already applied the particle filter to improve WLAN based indoor positioning. The process of Kalman filter iteratively predicts and corrects the prediction with measurements until some termination criteria met. The particle filter models the hidden state as a spatial posterior probability density function (PDF). Then, similarly to the Kalman filter, a particle filter also iterates prediction and correction, but it predicts and corrects the PDF not the state itself.

It is well known that the Kalman filter yields the best result if the assumptions (independent, white, normal distribution, and so on) of the Kalman filter hold. However, no one knows which one is better for a non-Gaussian estimation problem. If we know which one performs better than the other then we would not bother implementing both filters in developing an indoor LBS. Therefore, this paper introduces our experimental results of comparing the Kalman filter and the particle filter processes in improving the accuracy of K-NN WLAN based indoor positioning.

2 Related Works

We are comparing the Kalman filter against the particle filter in improving K-NN WLAN based indoor positioning. Therefore, K-NN WLAN based indoor positioning, the Kalman filter and the particle filter are the subjects related to our work.

2.1 K-NN WLAN Based Indoor Positioning

In the K-NN process, we build a look-up table in the first phase, or off-line phase. The entire area is covered by a rectangular grid of points called candidate points. At each of the candidate points we measure the RSSIs many times. Let RSSIij denote the j-th received signal strength of the signal sent by APi. A row of the look-up table is an ordered pair of (coordinate, a list of RSSIs). A coordinate is an ordered pair of integers (x, y) representing the coordinates of a candidate point. A list of signal strengths consists of five integers, RSSI1, RSSI2, ..., where RSSIi is an average of signal strengths RSSIij received at (x, y) and sent by APi. An example of look-up table is shown in Table 1.

In the second phase, or real-time phase, the positioning program gathers RSSIs the user receives at the moment. If the positioning program is running on the user's handheld terminal, then the terminal itself will collect RSSIs. Let $X = (RSSI1, RSSI2, \ldots)$ be the vector of the collected RSSIs, K-NN then searches the look-up table to find the K closest candidate points and returns the average of the K coordinates of them as the user's current location.

Table 1. An example look-up table of K-NN (C.P stands for Candidate Points, CP_i is the coordinates of i-th C.P, APi is the MAC address of i-th AP)

C.P	AP1	AP2	AP3	AP4	AP5
CP1	-39	-55	-56	-70	-67
CP2	-40	-56	-55	-69	-66
CP3	-44	-42	-62	-45	-61
...	...	...	...	...	...

2.2 Kalman Filter

There are hundreds of papers on the Kalman filter, most of which involve its application to autonomous or assisted navigation [14,15,16,17], whereas there have been a few reports on its application to indoor positioning or navigation [18]. The Kalman filter process is summarized in Fig. 1. A state of a moving object is represented as x in the process. The vector x consists of the moving object's x, y coordinates and the velocity as shown in Fig. 2. The matrix A is to project the next state from the current state and the value of A is shown in Fig. 2 where in A represents the time elapsed from the current to the next states. The measurement is represented as the vector z consisting of x and y coordinates. In order to select x-y coordinates from x, H should be as shown in Fig. 2. The Kalman process iteratively predicts and corrects the prediction with measurement. During the prediction, it projects the next state and the

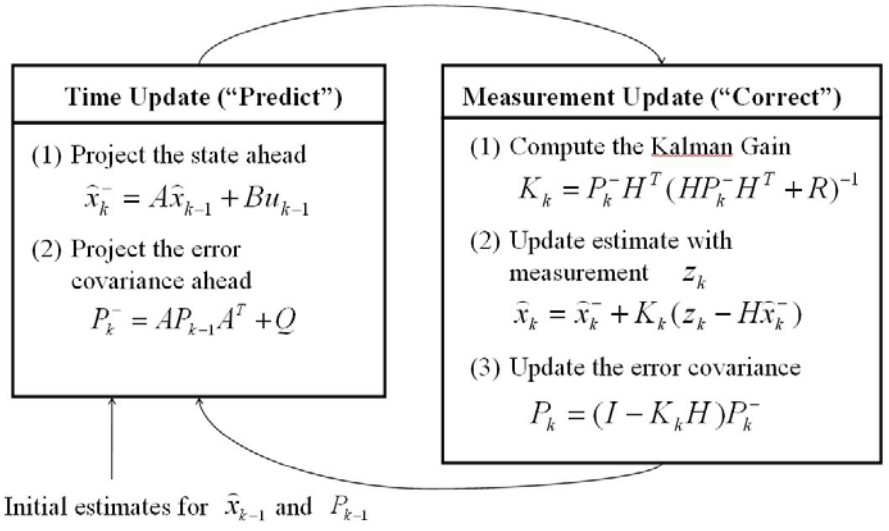


Fig. 1. The Kalman filter process

$$\hat{x}_k = \begin{pmatrix} x_k \\ y_k \\ v_{xk} \\ v_{yk} \end{pmatrix}, , A = \begin{pmatrix} 1 & 0 & \triangle t & 0 \\ 0 & 1 & 0 & \triangle t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Fig. 2. The matrices used in the Kalman filter

error covariance, P. During the correction, it computes the Kalman gain, update estimate with measurement, and update the error covariance.

2.3 Particle Filter

The particle filter process is also a repetitive process. After initialization, it repeats importance sampling step and resampling step as shown in Fig. 3[19].

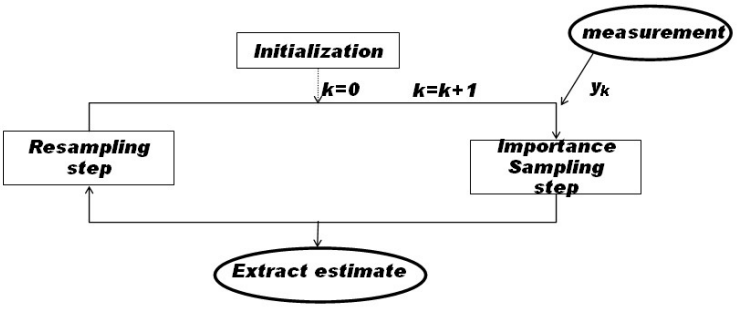


Fig. 3. A schematic diagram of the particle filter process

During the initialization, it populates particles, or samples, in the probability distribution function of the initial state of the dynamic system we are concerned. During the importance sampling step, considering the measurement at the moment, it assigns weights to the particles. Then, it determines the state at the moment as the weighted average of the particles. During the resampling step, it generates the next generation of particles.

3 Implementation of Kalman Filter

We have implemented our Kalman filter process as shown in Table 2 in C# using Microsoft visual studio. We initialize matrices Q, R, P, and X as shown in Table 4. Q and R are supposed to be the 'process noise covariance' and 'the measurement noise covariance' matrices respectively, but they are unknown. Besides, we know that it is the ratio of Q over R that really affects the performance of the Kalman filter [18]. Since the measurement noise is so big, we initialized R with much bigger number than Q's elements. X in the Kalman filter is corresponding to the state of the process which is the location of the mobile terminal in our experiment. Therefore, we initialize X with the first measurement. The matrix P represents the estimate error matrix. Since we already know our measurement error is so big, we initialize P with a big number. R and Q are the design parameters of the process and we have executed the process on the K-NN results shown in Table 1 with various values for R and Q. After many experimental executions, we have found the appropriate values for them as shown in Table 4.

Table 2. Our Kalman Filter Process

1. Initialize:
   For ($i = 2$; not EOF; $i++$) {
2. Step (1) of Time Update
3. Step (2) of Time Update
4. Step (1) of Measurement Update
5. Step (2) of Measurement Update. Use the i-th measured location as z. Print $\hat{x}_k$.
6. Step (3) of Measurement Update
   }

$$Q = \begin{pmatrix} 0.001 & 0 & 0 & 0 \\ 0 & 0.001 & 0 & 0 \\ 0 & 0 & 0.001 & 0 \\ 0 & 0 & 0 & 0.001 \end{pmatrix}, \; R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \; P_0 = \begin{pmatrix} 300 & 0 & 0 & 0 \\ 0 & 300 & 0 & 0 \\ 0 & 0 & 300 & 0 \\ 0 & 0 & 0 & 300 \end{pmatrix}, \; \hat{x}_k = \begin{pmatrix} 85175.4 \\ 12225.5 \\ 0 \\ 0 \end{pmatrix}$$

Fig. 4. Initial values of our Kalman filter process

4 Implementation of the Particle Filter

We have implemented our Particle filter process as shown in Table 3 in C# using Microsoft visual studio. Assuming that the moving object moves along the x-axis

at the speed of 788 mm/time unit, we initialize and to 0 and 788, respectively. We also assumed that the ranges of X and Y coordinates of the true positions are as follows:

$77290 \leq X \leq 81993$ and $9748 \leq Y \leq 17657$.

Therefore, we set the area for the initial particles as follows:

$74800 \leq X \leq 84490$ and $7500 \leq Y \leq 19850$.

According to [13], the appropriate number of particles is 400. So, we set the distance between a pair of adjacent particles parallel to the $X(Y)$ axis be 510(650) pixels for the initial particles.

At the moment, assuming that the mobile terminal can move to any direction, we use "RandomNormal($0, 1.0\pi$);", and that the speed can also change dramatically, we use "RandomNormal(788, 700);" in the procedure of generating particles, where RandomNormal returns a random number in Normal(mean,

Table 3. Our particle filter process

1. Initialization. Generate initial particles (1 particle/m^2). For each particle, let 1 be its weight.
$$\theta = 0 \text{ (in rad)}, V = 788(mm/s)$$
2. for ($i = 1$; not EOF; $i++$) {
   (a) z_i = the i-th measured location;
   (b) for each particle x_i , compute its weight w_i as follows:
$$w_i = w_i p[z_i|x_i], \text{ where } p[z_i|x_i] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z_i - x_i)^2}{2\sigma^2}\right)$$
   where $\sigma = 20,000$.
   (c) TotalOfWeight = $\sum_{j=1}^{n} w_j$, where n is the number of particles
   (d) Normalize the weights so that the total of them is 1.
   (e) Print the weighted sum, (X, Y, θ, V), of all the particles
   (f) Particle = $Propagation(X, Y, \theta, V)$
   }
   Propagation$(X, Y, \theta, V, TotalOfWeight)${
   if ($TotalOfWeight < 0.000001$)
   return(1 particle/m^2); // For each particle, let 1 be its weight
   else { // generate numberOfParticles (400, for example) particles
   for (i=0; I ¡ numberOfParticles; i++) {
   $Temp\theta = +RandomNormal(0, 1.0\pi)$;
   $TempV = V + RanddomNormal(0.0, 300)$;
   Particle[i].X = X + TempV*cos(Tempθ);
   Particle[i].Y = Y + TempV*sin(Tempθ);
   Particle[i].θ = Tempθ ;
   Particle[i].V = TempV;
   } // end of for
   return(Particle);
   } // end of else
   }

standard deviation) distribution. Now, the remaining parameter is the weight function for step 2.(b) in Table 3. As the weight of a particle, we use the inverse of the distance between the particle and the posterior generated in the previous iteration. The following is the actual sentence used in our program:

```
particles[j].weight
    = 1/(Math.Sqrt((x - particles[j].X) * (x - particles[j].X)
      + (y - particles[j].Y) * (y - particles[j].Y)));
```

5 Experiments

We are comparing the Kalman filter and the particle filter processes in improving the accuracy of WLAN based indoor positioning. Therefore, we have performed the K-NN positioning while we are walking through the path shown in Fig. 5. We have walked through the path 4 times. A part of our test results is shown in Table 4. The unit of the measurements is AutoCAD unit. In our experiments,

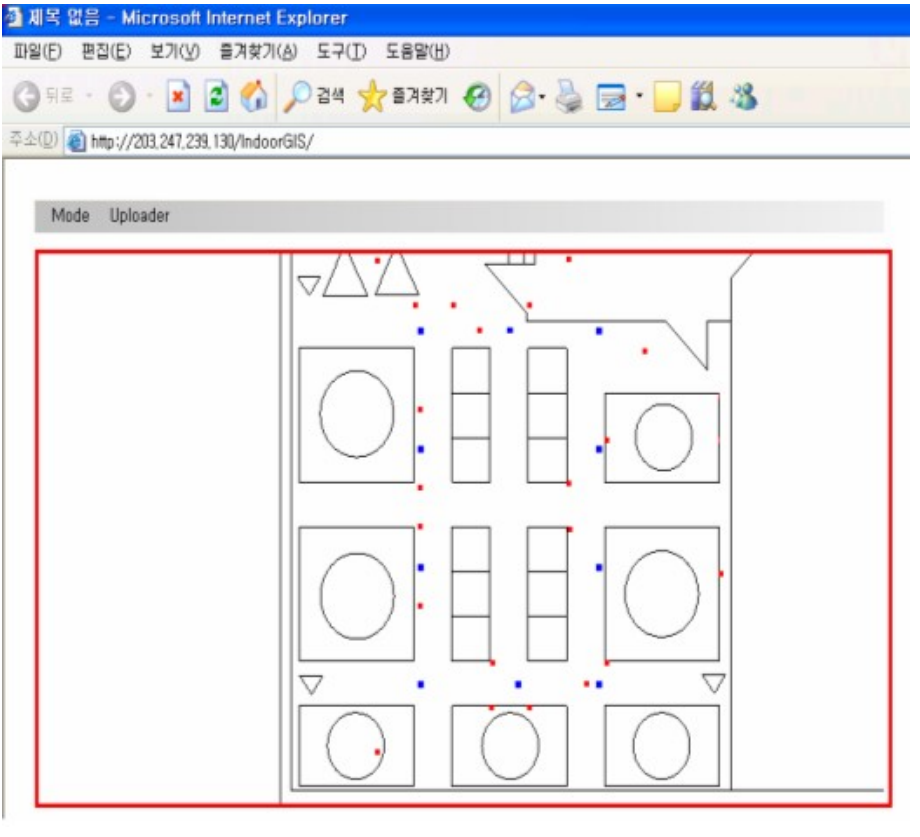


Fig. 5. A typical test result of K-NN indoor positioning

Table 4. A part of the 1-NN results (average error is 6,240.03 and standard deviation is 2,732.15)

	True X	True Y	Measured X	Measured Y	Error
Point 1	79,626	17,657	85,175	12,225	7,765
Point 2	80,414	17,657	77,187	7,444	10,711
Point 3	81,203	17,657	82,190	10,207	7,515
...	...	...	...	...	...
Point 29	78,069	17,657	76,187	19,217	2,444
Point 30	78,847	17,657	79,172	18,222	651

0.5 meter is equivalent to about 1,000 units. That is, 6420, the average error of K-NN obtained by the experiment, is equivalent to about 3.21 meters.

Then we have run the Kalman filter process on the data shown in Table 4 with various R and Q. Our test results are summarized in Table 5. R and Q matrices represent measurement noise covariance and process noise covariance, respectively. We found that our measurement noise covariance is 7025691.4 from Table 4. However, process noise covariance is hard to obtain. After many experiments, we have found that the accuracy of the Kalman filter process depends on the ratio of R value over Q value. Therefore, we have fixed R to 1 and changed the value of Q. After many experiments, we have found that $R = 1$ and $Q = 0.0001$ is appropriate. We have walked around the path on Fig. 5 4 times. The row labeled "For all" of Table 5 represents the average of all the 4 laps while "first lap" represents the average of the first lap. In summary, our Kalman filter process improves the accuracy of our K-NN process by $6421/3323 = 1.93$ times. In other words, the error is reduced by almost half if we use the Kalman filter process. The unit of the figures shown in Table 5 is AutoCAD pixel and 3323 is equivalent to about 3.323 meters.

Table 5. A summary of our Kalman filter test results

R	1	1	1	1	1	1	7025691.4	7025691.4
Q	0.1	0.01	0.001	0.0001	0.000055	0.00001	700	1.0
For all	4975	4233	3459	3323	3355	3435	3376	4071
first lap	4855	4304	4207	4160	4156	4152	4028	5883

In the first experiment of running our particle filter, we have tried the following two weight functions: 1/(distance**2) and 1/(distance**4). Our test results are summarized in Table 6. The test results suggest 1/(distance**2) as the weight function. Considering the average error of K-NN, 6421, the particle filter improves the accuracy by 14% if 1/(distance**2) is used as the weight function. However, considering the average error of the Kalman filter, 3323, there should be a room to improve the particle filter process.

Table 6. Test results for simple weight functions

Weight	K-NN	1/distance	1/(distance**2)	1/(distance**4)
Average error	6240	5854	5534	12415

Therefore, we have tried the weight function shown in Table 3 with various values for σ, the standard deviations of RandomNormal, and found that the values shown in Table 3 are appropriate and our test results are summarized in Table 7.

Table 7. Test results for another simple weight functions

Weight	K-NN	$\sigma = 12,000$	$\sigma = 20,000$	$\sigma = 30,000$
Average error	6,240	3,716	3,648	3,703

Chao et al.[13] assigned the weight of 0 to the particles which are located at invalid area, for example, located beyond a wall. With this strategy, we could improve our particle filter process a little further as shown in Table 8.

Table 8. Test results for location-constrained weight function

Weight	K-NN	Kalman filter	Particle filter $\sigma = 20,000$	Location-constrained weight
Average error	6,240	3,323	3,648	3,572.4

Chao et al.[13] also suggested the strategy of removing a particle if its location is invalid during the Propagation. We have tried this strategy with various sizes of the valid area. However, this strategy did not improve the accuracy of our particle filter as shown in Table 9.

Table 9. Test results for location-constrained particle generation

Weight	K-NN	Kalman filter	Particle filter $\sigma = 20,000$	Size=3m	Size=2m	Size=0.6m
Average error	6,240	3,323	3,648	3,980	4,000	4,052

6 Conclusions

Chao et al. [13] and Yim et al.[18] introduced particle filter application and Kalman filter application on K-NN WLAN based indoor positioning. We have compared these two methods and found out that their accuracies are almost same. Chao et al. showed that the accuracy of the particle filter can be almost doubled if we make use of location-constraint. We have tried the location-constrained weight strategy and found out it can improve the accuracy of our

particle filter process a little. We have also tried the location-constrained particle generation strategy. However, it did not contribute any in improving our particle filter process. Instead, as we decreased the size of the valid area, the running time of our process significantly increased. This implies that the performance of the particle filter changes drastically as the characteristics of measured data change.

The particle filter takes a long time for populating the particles. However, the process is intuitive and easy to implement. It is also easy to modify an implemented particle filter to adapt it to the new set of data. On the other hand, the Kalman filter is easy to tune R and Q to get the better results. However, it is not so easy to modify an implemented Kalman filter to adapt it to a different kind of measurements. In the case of Kalman filter, if we assign small numbers to the elements of R, it is guaranteed for the Kalman filter to yield at least the measurements. On the other hand, in the case of particle filter, the result can be worse than the measurements if we use wrong formula for weight calculation or population generation. In conclusion, if both are tuned to get their best results, then their performance is almost same and each of them has its own strength and weakness. Therefore, in practice, we should choose appropriate one considering the strength and weakness, characteristics of measured data and the situation of application.

Acknowledgments. This research was supported by Basic Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0006942).

References

1. Yim, J.: Introducing a decision tree-based indoor positioning technique. Expert Systems with Applications 34(2), 1296–1302 (2008)
2. Enge, P., Misra, P.: The Global Positioning System. Proceedings of the IEEE Special Issue on GPS, 3–172 (1999)
3. Tekinay, S.: Special Issue on Wireless Geolocation Systems and Services. IEEE Communications Magazine (April 1998)
4. Want, R., Hopper, A., Falcao, V., Gibbons, J.: The Active Badge Location System. ACM Transactions on Information Systems 10(1), 91–102 (1992)
5. Harter, A., Hopper, A.: A New Location Technique for the Active Office. IEEE Personal Communications 4(5), 43–47 (1997)
6. Priyanthat, N., Chakraborty, A., Balakrishnan, H.: The Cricket Location-Support System. In: Proc. of 6th ACM International Conference on Mobile Computing and Networking, Boston, MA (August 2000)
7. Orr, R.J., Abowd, G.D.: The Smart Floor: A Mechanism for Natural User Identification and Tracking. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI 2000), The Hague, Netherlands, pp. 1–6 (April 2000)
8. Krumm, J., et al.: Multi-camera Multi-person Tracking for Easy Living. In: Proceedings of the 3rd IEEE Int'l Workshop on Visual Surveillance, Piscataway, NJ, pp. 3–10 (2000)

9. Bahl, P., Padmanabhan, V.: RADAR: An in-building RF-based user location and tracking system. In: Proceeding of INFOCOM 2000, pp. 775–784 (March 2000)
10. Ladd, A.M., Bekris, K.E., Rudys, A., Kavraki, L.E., Wallach, D.S., Marceau, G.: Robotics-based Location Sensing Using Wireless Ethernet. In: Proceedings of the Eighth Annual International Conference on Mobile Computing and Networking (MOBICOM 2002), New York, pp. 227–238 (2002)
11. Kotanen, A., Hannikainen, M., Leppakoski, H., Hamalainen, T.D.: Experiments on local positioning with bluetooth. In: Proceedings of International Conference on Information Technology: Coding and Computing (Computers and Communications) (ITCC 2003), pp. 297–303 (April 2003)
12. Qasem, H., Reindl, L.: Unscented and Extended Kalman Estimators for non Linear Indoor Tracking Using Distance Measurements. In: Proc. of the 4th Workshop on Positioning, Navigation and Communication, WPNC 2007, pp. 177–181 (March 2007)
13. Chao, C., Chu, C., Wu, A.: Location-Constrained Particle Filter human positioning and tracking system. In: Proceedings of IEEE Workshop on Signal Processing System, pp. 73–76 (2008)
14. Teo, T., Chai, J., Yao, W.: Design of a positioning system for AGV navigation. In: Proc. of the 7th International Conference on Control, Automation, Robotics and Vision (ICARCV 2002), pp. 637–642 (2002)
15. Peroutka, Z.: Design considerations for sensorless control of PMSM drive based on extended Kalman filter. In: Proc. of 2005 European Conference on Power Electronics and Applications, pp. 1–10 (September 11-14, 2005)
16. Boussak, M.: Implementation and experimental investigation of sensorless speed control with initial rotor position estimation for interior permanent magnet synchronous motor drive. IEEE Transactions on Power Electronics 20(6), 1413–1422 (2005)
17. Bolognani, S., Tubiana, L., Zigliotto, M.: Extended Kalman filter tuning in sensorless PMSM drives. IEEE Transactions on Industry Applications 39(6), 1741–1747 (2003)
18. Yim, J., Park, C., Joo, J., Jeong, S.: Extended Kalman Filter for Wireless LAN Based Indoor Positioning. Decision Support Systems 45, 960–971 (2008)
19. Copsey, K.: Tutorial on Particle filters. Pattern and Information Processing Group DERA Malvern, K.Copsey@signal.dera.gov.uk

On Possibility of Conditional Invariant Detection

Hani Fouladgar, Behrouz Minaei-Bidgoli, and Hamid Parvin

School of Computer Engineering, Iran University of Science and Technology (IUST), Tehran, Iran
{fouladgar,b_minaei,parvin}@iust.ac.ir

Abstract. Software engineering includes some different process such as designing, implementing and modifying of software. All of these processes are done to have fast developed software as well as reach a high quality, efficient and maintainable software. Invariants help programmer and tester to do most steps of software engineering more easily. Invariants are mostly always true but of course with a specific confidence. Since some invariants are produced on some conditions of program execution and not always, conditional invariants can show the behavior of program so much better. For producing this kind of invariants, it might be used some technique of data mining such as association rule mining or using decision tree to obtain rules. So the paper will introduce a new perspective to dynamic invariant detection. Also the feasibility of conditional invariant detection is examined and a framework to extract them is proposed.

Keywords: Daikon, Invariant, Association Rules, Variable Relations, Decision tree, Program point, Data mining, Software engineering, Predicate, Verification.

1 Introduction

In recent years, invariant plays an important role in software engineering such as software testing and verification. Invariants are properties of program variables and relationships between these variables in a specific line of code which called program point. Generation of invariants is a significant key in program verification. These properties and relationships among the program variables or constants are always true; thus programmer or tester can estimate the behavior of program in different program points. Invariant also is used in generating software behavioral model [1] so invariant can also be useful in software engineering in this way. With the help of software behavioral model we can lightly perform design, validation, verification, and maintenance. As seems, one of the most significant contributions of invariants is modifying of code where properties help programmer to verify the code. Software testing takes a considerable time in software development life cycle. Although software testing is done automatically in present day, but traditionally the onus of software testing was human's obligation [2]. Testing is divided in two categories; functional testing and structural testing. Functional testing, which is also called black box testing, performs testing without considering of the program logic but by checking the program output against the input. This kind of testing does not take into

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 214–224, 2011.

account program inner workings. On the other hand, structural testing or white box testing analyze program according to actual code checking and knowing about its logic. Invariants are also detected by static and dynamic approaches [4].

In static approaches, runtime behavior and syntactic structures of program are analyzed without actual running of code [5]. Static analysis completely is done automatically. One analysis which traditionally has been used in compilers for collecting necessary information in optimization is Data-flow. Indeed, Data-flow analysis detects some essential invariants in each program points and employs these invariants to find out the behavior of program. This kind of behavior can be used in compilers for optimization. Abstract interpretation is a theoretical framework for static analysis [6].

On the other hand, Dynamic approaches extract program properties and information by the help of actual executing of the program code [7]. In the other words, by executing the program with different inputs, called test suits, it is possible to detect invariants dynamically. Dynamic invariants extraction emerges to software engineering in recent years with the advent of Daikon [4]. Program properties of certain point of program are reported by use of invariant inference system via different test suits through different executions. Invariant mostly is checked in the entries and exits of each function.

This paper concentrates on the dynamic extraction of conditional invariants. Conditional invariants are the invariants which are revealed in specific form of conditional proposition, throughout all this paper. These invariants emerge dynamically and all of the steps are fully automatic. We are going to improve the quality of discovered invariants dramatically by using association rule mining. In association-rule-based invariant extraction system, invariants are represented through the variables' condition. For inferring invariants, they are two prominent issues [2]: first we would be able to determine the beneficial invariants and then to exert inference on program context. In this paper we handle these two parts. The rest of this paper begins with related work (section 2). After that some frequently used definitions are appeared (section 3) and the paper continues with section 4 about what conditional invariant is. Then we propose condition invariant framework (section 5). At last we conclude and talk about future work (section 6).

2 Related Works

In this section, we peruse some implementations of dynamic invariant detection. We discuss some implementations which are more relevant to our job but it is worth to mention there are many valuable efforts in this topic.

Dynamic invariant detection is introduced by *Daikon* [3]. Daikon is the most favored tool for detecting dynamic invariant and until now, comparing with other dynamic invariant detection methods [3]. However this software has some problems out of which the most serious one is being time-consuming.

DySy provided a dynamic symbolic execution technique to improve the quality of inferred invariant [8]. Besides executing test cases like other dynamic invariant inference tools DySy coincidentally performs a symbolic execution. For each test

unit, DySy results in program's path conditions. At the end, all path conditions are combined and build the result.

Agitar presented Agitator which is a commercial testing tool and is inspired by Daikon [9]. Software agitation joins the results of research in test-input generation and dynamic invariant detection. The results are called *observations*.

The DIDUCE [10] helps programmer by detecting errors and determining the root causes. Besides detecting dynamic invariant, DIDUCE checks program behavior against extracted invariants up to each program points and reports all detected violations. DIDUCE checks simple invariants and does not need up-front instrument.

While there are many related work in the dynamic invariant detection, there is lack of any considerable related work about dynamic invariant detection. This makes this paper first attempt to deal with the dynamic detection of conditional invariants.

3 Terminology

In this section we discuss about notions which we repeatedly use trough this paper. The aim of the section is to help readers to obtain a better perception of the paper.

Definition 1. *Invariants* can be defined as prominent relation among program variables. Invariants in programs are formulas or rules that are emerged from the source code of a program and remain unique and unchanged with respect to the running phase of a program with different parameters.

Definition 2. *Program points* are specific points in a program, such as the *Enter* or *Exit* point of a function, which serve as report points for variable relations and invariants. Most frequent program points in use are the Enter and Exit points of sub-programs and functions.

Definition 3. *Pre-conditions* of a program point are the conditions, relations and invariants that hold immediately before approaching to that program point. In the case of sub-programs or a function *Enter* point of a sub-program or a function acts as its pre-condition.

Definition 4. *Post-conditions* of a program point are the conditions, relations and invariants that hold immediately after leaving from that program point. In the case of sub-programs, a function *Exit* point of a sub-program or a function is considered as its post-condition of it. Typically, post-condition also contains relations between the original value of a variable and its modified one (before and after that program point). In other words, invariants in post-conditions contain relations between variables in pre-condition and post-condition.

4 Conditional Invariant

Most of invariant extraction systems concentrate on perfect invariants and they are unable to express invariants which are appeared in special situation. This means, the invariants which are reported by invariant extraction system are true with the specific confidence but they do not figure out invariants which are true in a special condition.

To clarify the matter, consider Fig. 1. (This example is artificial and illustrates several points we are going to discuss.)

In this example we assume variables x and y are global. An appropriate unit test for this function might be x<y and its complement. In an ordinary invariant extraction system the *post-condition* invariant which could be detected for this function is:

- x>y

This invariant shows after leaving compute() the x values are always are greater than y values. This invariant is adequate but it does not present a complete behavior of this function. This means this mere invariant can not be useful neither in *formal specification* nor *assert statement*.

```
void compute()
{
    if (x < y)
    {
        int temp = x;
        x = y;
        y = temp;
    }

}
```

Fig. 1. Example method whose invariant we want to infer

This deficiency puts us to think to have a set of invariants which can appropriately show the program behavior. In the other words we need a set of invariants which tell us compute() swaps x and y values when y value is greater. The final outcome of post-condition of compute() invariants (or compute():::Exit in our method) are:

```
1 orig(x)>orig(y)  -> x=orig(x)
2 orig(x)>orig(y)  -> y=orig(y)
3 orig(x)<orig(y)  -> x=orig(y)
4 orig(x)<orig(y)  -> y=orig(x)
```

Fig. 2. Related invariants in our method

In upon invariants, *orig(var)* shows *var* value just before entrance of compute(). This approach removes the weakness of previous dynamic invariant inference. As seen, Fig. 2 completely describes the function behavior.

Over all our work contains following parts:

- We introduce the idea of using association rule mining for invariant inference. We believe our method makes up the next generation of dynamic invariant inference tools. We believe our approach opens a new ways to perform dynamic invariant inference in not far future.
- We describe our approach by flowcharts.

5 Proposed Conditional Invariant Detection Framework

In this section, we propose our idea in details. First, we provide *predicates* for each execution of program point and then invariant detector uses these predicates to extract the rules. Program points are usually function entries and exits. Function entries and exits are called *Enter point* and *Exit point* of function. For Enter point, all values of global variables and parameters participate while for exit all values of global variables and parameters as well as their prior values participate. With having more variety of invaluable predicates, more beneficial invariants are produced. Extracted rules show behavior of program point in conditional form. In following we discuss the classes of predicates and clarify all predicates.

To better understanding of the process, Fig. 3 schematically shows the algorithm flowchart of employing association rule mining in extracting conditional invariants step by step. Each datatrace file in Fig. 3 contains possible predicate of in a program point.

5.1 Classes of Predicate

Here we present all classes of predicates which might be used by invariant detector. By the help of an association rule mining tool, we can extract conditional invariants. We try to provide a terse set of predicates to have an acceptable potential result but definitely there are some predicates which are missed. The following lists classes of predicates which are used in our approach, where x and y are variables:

- Predicates over any numeric variable:
  - IsNonZero: when the variable is never set to 0
  - IsOne: when the variable is always equal to 0
  - IsMinesOne: when the variable is always equal to -1
  - IsEven: when the variable is always even
  - IsPowerOfTwo: when the variable is always power of two
- Predicates over any string variable:
  - IsNull: when the variable is always null
  - IsEmpty: when the variable contains no characters
- Predicates over two numeric variable:
  - Ordering comparison: $x < y$, $x \leq y$, $x > y$, $x \geq y$, $x = y$, $x \neq y$
  - functions: $y = fn(x)$ or $x = fn(y)$, for fn a built-in unary function (absolute value, negation, bitwise complement)
- Predicate over two string variable:
  - Equality: $x = y$ when two strings are equal
  - Substring: $y=sub(x)$ when y is substring of x
  - Reversal: $y=rev(x)$ or $y=rev(x)$ when x is the reverse of y
- Predicates over a array:
  - Element relationship: when the array elements are equal or sorted by (=>,>,<,<=)
  - IsNonZero: when none of array elements are equal to 0

- Predicate over an array and a numerical variable:
  - Membership: x∈ y (x and y are common type arrays)
- Predicate over two arrays:
  - comparison: x < y, x ≤ y, x > y, x ≥ y, x = y, x ≠ y
  - Sub-array : y=sub(x) when y is sub-array of x
  - Reversal: y=rev(x) or y=rev(x) when x is the reverse of y

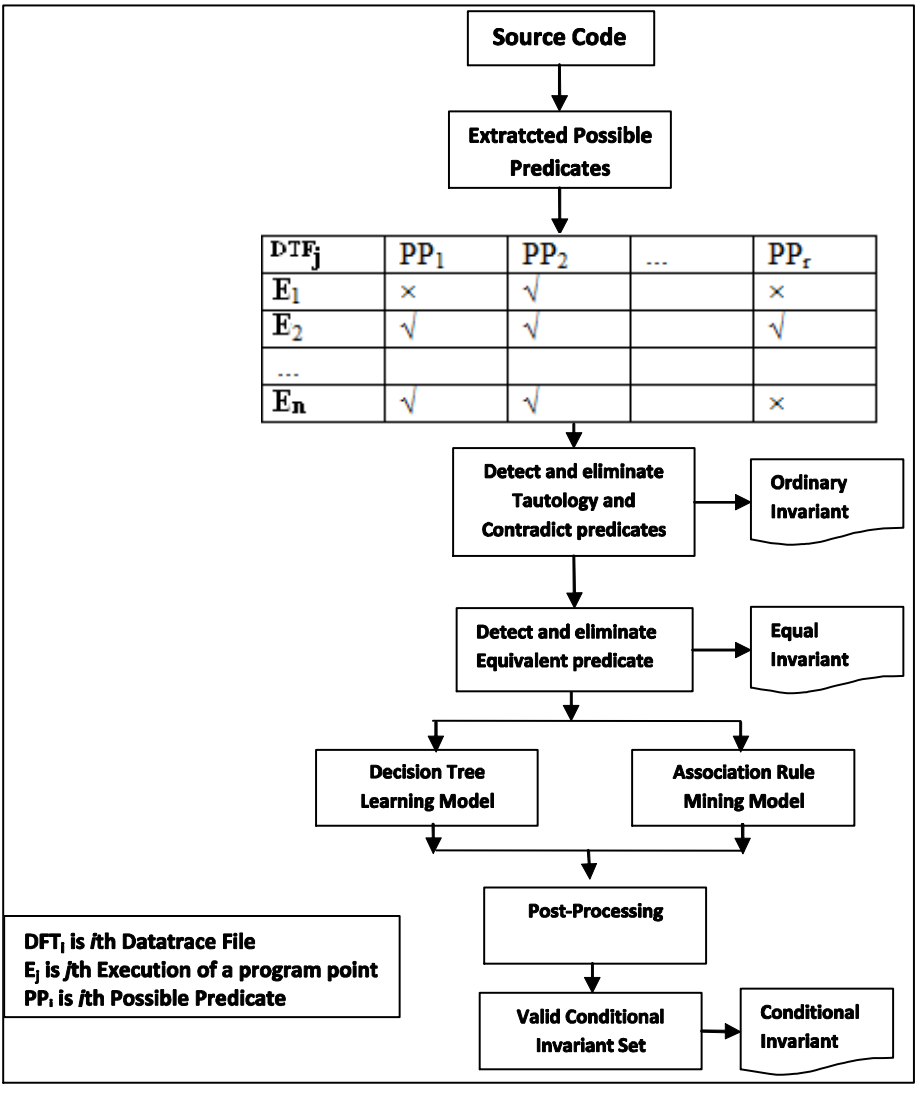


Fig. 3. Algorithm flowchart of employing association rule mining in extracting conditional invariants

Presented predicates are produce for each program point. Each presented predicate has Boolean value. In other words this predicates might be *true* or *false*. By performing association rule mining on these predicate we would have some rules with specific support and confidence. In following we discuss about association rule mining Model and the domination of this technique whether support our aim.

5.2 Using Association Rule Mining on Defined Predicates

In subsection 5.1 we defined all predicates which are interfered for each program point variable. In the other words we bring forward any possible predicates in a specified program point. The obtained predicates, all, have Boolean values. These values can easily be used for mining association rules. Each time for each predicate as consequent we check other predicates which make relations with it. Consider we have predicates P_1, P_2, P_3, … , P_q. We start with P_q we check all predicates if they have relation with P_q. It means we check if P_1 as the antecedent can result P_q otherwise we conjunct P_1 and P_2 and check if now they result P_q and so forth. Then we will perform these steps for P_{q-1}.

For closer looking at our paper tendency consider presented function in Fig. 1. We discussed about this function and its Exit program point conditional invariants. Now we demonstrate the steps to create these rules. First we must prepare our database and transactions. Each record shows one executing of function. We instrument the code so that in each execution, predicates between all variables are stored in a file. The result is presented in Fig. 4.

Transaction	orig(x)>orig(y)	orig(x)<orig(y)	x=orig(x)	y=orig(x)	x=orig(y)	y=orig(y)
T_1	true	false	true	false	false	true
T_2	false	true	false	true	true	false
T_3	false	true	false	true	true	false
T_4	true	false	true	false	false	true
T_5	true	false	true	false	false	true
T_6	false	true	false	true	true	false

Fig. 4. Related transaction for Fig. 1

Fig. 4 shows neither all transaction nor all predicates but it presents just some of them to manifest the method. The minimum support and minimum confidence respectively are 50% and 100%. Two large itemsets which are inferred from Fig. 4 is:

- orig(x)>orig(y), x=orig(x),y=orig(y)
- orig(x)<orig(y), y=orig(x), x=orig(y)

And following rules are archived:

- orig(x)>orig(y)⇒ x=orig(x)
- orig(x)>orig(y)⇒y=orig(y)
- orig(x)<orig(y)⇒y=orig(x)
- orig(x)<orig(y)⇒x=orig(y)
- x=orig(x)⇒orig(x)>orig(y)
- x=orig(x)⇒y=orig(y)

- y=orig(y)⇒orig(x)>orig(y)
- y=orig(y)⇒x=orig(x)
- y=orig(x)⇒orig(x)<orig(y)
- y=orig(x)⇒x=orig(y)
- y=orig(y)⇒ orig(x)<orig(y)
- y=orig(y)⇒ y=orig(x)

All presented rules are true and obey minimum support and minimum confidence but only four first one are tangible and others must be filtered. The four first rules are the same as rules we represent in Fig. 2. These to conditional invariant describe the behavior of compute(). One thigh which is important to say is in this method rules' Consequent part contains only one predicate and we do not have compound consequences. Whole the process of *Association Rule Mining Model* box in the Fig. 3 is illustrated in the Fig. 5.

5.3 Time Order

Here in subsection we check our approach time order. It is necessary check time order because we want to see if it is affordable. Assume we have m variables in a program point. Each two variables make a predicate so overall we have q predicates. q is obtained via equation (1):

$$\binom{m}{2} = q \tag{1}$$

So we have predicates P_1, P_2, P_3, … , P_q. To have a rule with P_q as the consequent, our association rule mining tool must check if each of P_1, P_2, P_3, … , P_{q-1} has relationship with P_q then it has to check if two of P_1, P_2, P_3, … , P_{q-1} have relationship and so forth. Consequently for having a rule with P_q as the consequent, our tool has to handle (2) number of checks:

$$\binom{q-1}{1} * 2 + \binom{q-1}{2} * 2^2 + \cdots + \binom{q-1}{q-1} * 2^{q-1} \tag{2}$$

Totally, the association rule mining tool must handle (3) number of checks:

$$\sum_{i=1}^{q-1} \sum_{j=1}^{i} \binom{i}{j} * 2^j \tag{3}$$

For example assume we have 7 variables in one of our program points. The total number of checks might be 2097152. If we have 10 variables the total number of checks might be 3518437208832. As seen the time order is exponential. This time order is not acceptable at all. Of course we should pay attention that all these check is not handled because if for example P_1 has relationship with P_q other sets of predicates which contains P_1 will not be checked anymore and will not be interfered but it does not affect the time order so much and overall time order is exponential.

Another issue which is worthwhile to emphasize again is that, in generating rules left-hand part or antecedent must be in the shortest state. For example if $P_1 \Rightarrow P_n$ rules such as P_1, $P_2 \Rightarrow P_n$ is not valuable.

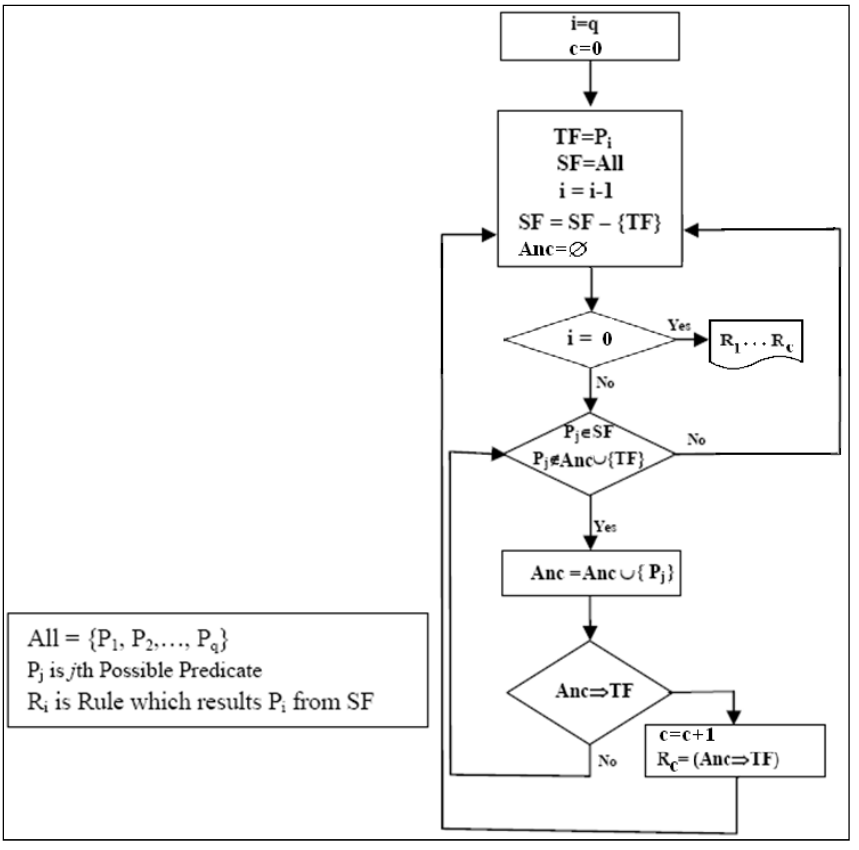


Fig. 5. Association Rule Mining Model

5.4 Using Decision Tree Learning Model

In decision tree we can find a relationship between one attribute called *goal* or *class* and other attributes. On other words we can predict the goal by having other attributes. Two properties of decision tree are:

- Approximately lowest number of antecedents
- Feasible highest confidence of the rule

These two properties might be so much helpful for generating association rules by the contribution of decision tree because our main purpose is to have some rules with lowest number of antecedents with high confidence. For employing this technique we should consider each predicate as the goal and try to capture the predicates which result our goal. Consider we have predicates P_1, P_2, P_3, … , P_q. We start with a predicate such as P_1 as our goal or class. We make the decision tree for P_1 and then we try to figure out other predicates which defined P_1's result. If we go upward from leaves to root in obtained tree, they can be rules which show when P_1 is true and when it is false. Then we obtain the P_2's tree and so forth.

Using decision tree for obtaining the rules is so much faster than normal association rule mining. Because by employing decision tree we do not have to check all the predicates to each other but we split our source into some smaller subsets and then classify each predicate lonely.

Fig. 6 demonstrates process of *Decision Tree Learning Model* box in the Fig. 3.

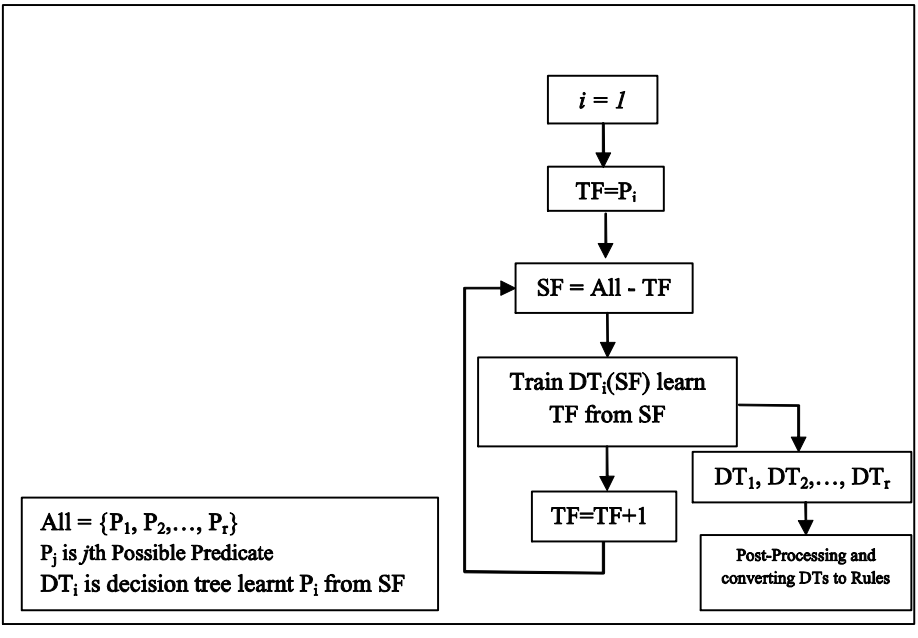


Fig. 6. Decision Tree Learning Model

6 Conclusion and Future Work

Here in this subject, different properties of program are checked in different program points. These properties usually show the behavior of that program point. Invariants are always true with a specified confidence so they can not represent those behaviors which are true with assuming of a condition. Conditional invariant solve this problem because via these kinds of invariant, we would have predicate which are true while another predicate is also true. We tried to generate the association rules by ordinary association rule mining and by repeated checks. The time order in via these methods is exponential and is not acceptable at all. So we brought up the decision tree and try to obtaining rules with this technique. We check each predicate as a goal try to find the related predicates which result the goal.

For future work, we can try to obtain the rules by the help of *Bayesian network*. Bayesian network is another data mining technique which creates a model to predict the value of the goal based on other input variables. Bayesian networks are very

efficient when the features (or predicates in our work) do not have correlation. By Bayesian network and its related methods we can detect the conditional invariant from presented predicates in each program point.

References

1. Krkay, I., Brunx, Y., Popescuy, D., Garciay, J., Medvidovic, N.: Using dynamic execution traces and program invariants to enhance behavioral model inference. In: ICSE NIER (2010)
2. Vanmali, M., Last, M., Kandel, A.: Using a neural network in the software testing process. International Journal of Intelligent Systems 17(1), 45–62 (2002)
3. Ernst, M.D., Cockrell, J., Griswold, W.G., Notkin, D.: Dynamically discovering likely program invariants to support program evolution. IEEE TSE 27(2), 99–123 (2007)
4. Ernst, M.D., et al.: Dynamically discovering likely program invariants to support program evolution. In: Proc. ICSE 1999, pp. 213–224. ACM, New York (1999)
5. Weiß, B.: Inferring invariants by static analysis in KeY. Diplomarbeit, University of Karlsruhe (March 2007)
6. Jones, N.D., Nielson, F.: Jones and Flemming Nielson. Abstract interpretation: A semanticsbased tool for program analysis. In: Abramsky, S., Gabbay, D.M., Maibaum, T.S.E. (eds.) Handbook of Logic in Computer Science, vol. 4, pp. 527–636. Oxford University Press, Oxford (1995)
7. Ernst, M.D., Perkins, J.H., Guo, P.J., McCamant, S., Pacheco, C., Tschantz, M.S., Xiao, C.: The Daikon System for Dynamic Detection of Likely Invariants. Science of Computer Programming (2006)
8. Csallner, C., et al.: DySy: Dynamic symbolic execution for invariant inference. In: Proc. of ICSE (2008)
9. Boshernitsan, M., Doong, R., Savoia, A.: From Daikon to Agitator: Lessons and challenges in building a commercial tool for developer testing. In: ISSTA, pp. 169–179 (2006)
10. Hangal, S., Lam, M.S.: Tracking down software bugs using automatic anomaly detection. In: ICSE, pp. 291–301 (2002)

On Temporal Gödel-Gentzen Translation

Norihiro Kamide

Waseda Institute for Advanced Study, Waseda University,
1-6-1 Nishi Waseda, Shinjuku-ku, Tokyo 169-8050, Japan
drnkamide08@kpd.biglobe.ne.jp

Abstract. Temporal logics and their intuitionistic counterparts are of growing importance in Computer Science. These intuitionistic counterparts, called intuitionistic (or constructive) temporal logics, are known to be useful for formalizing functional programming. To show a clear relationship between temporal logics and their intuitionistic counterparts has thus been required. In this paper, a theorem for embedding first-order linear-time temporal logic into its intuitionistic counterpart is proved using Baratella-Masini's temporal extension of the Gödel-Gentzen negative translation of classical logic into intuitionistic logic.

1 Introduction

The *Gödel-Gentzen negative translation* is well-known to be a fundamental translation that can obtain a theorem for embedding classical logic into intuitionistic logic (see e.g., [8,3]). This embedding theorem is known to be important for obtaining a clear relationship between classical logic and intuitionistic logic. We wish to extend this embedding theorem to *temporal logics*, since classical and intuitionistic (or constructive) temporal logics, which have independently studied, are of growing importance in Computer Science. In this paper, a theorem for embedding first-order linear-time temporal logic into its intuitionistic counterpart (i.e., intuitionistic first-order linear-time temporal logic) is proved using Baratella-Masini's temporal extension of the Gödel-Gentzen negative translation.

Linear-time temporal logics (LTLs) [7], which are usually based on classical logic, are well-known to be useful for verifying and specifying concurrent systems. The intuitionistic counterparts of LTLs, called *intuitionistic (or constructive) LTLs*, are known to be useful for formalizing functional programming and staged computation. A Gentzen-type sequent calculus LT_ω for LTL, which is based on classical logic, was introduced by Kawai, and the cut-elimination and completeness theorems for this calculus were proved [6]. Two constructive and bounded versions of LT_ω, which are based on intuitionistic logic, were introduced by Kamide and Wansing [5]. In this paper, an alternative intuitionistic counterpart ILT_ω (of LT_ω), which was also discussed in [5], is introduced for obtaining the objective embedding theorem.

An intuitionistic temporal natural deduction system PNJ for a *logic of positions* was introduced by Baratella and Masini, and the strong normalization

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 225–234, 2011.

theorem for PNJ was proved [1]. The system PNJ is based on the notion of *position formulas*, and has an induction inference rule concerning a time induction axiom. They also introduced a temporal extension of the Gödel-Gentzen negative translation, and proved a theorem for embedding a classical temporal natural deduction system PNK into PNJ, using the extended translation. In this paper, we use the Baratella-Masini translation with some extensions for the quantifier cases.

2 Sequent Calculus

The following list of symbols is adopted for the language of the underlying logics: free variables $a_0, a_1, ...$, bound variables $x_0, x_1, ...$, functions $f_0, f_1, ...$, predicates $p_0, p_1, ...$, logical connectives $\rightarrow$ (implication), $\wedge$ (conjunction), $\vee$ (disjunction), $\neg$ (negation), $\forall$ (any), $\exists$ (exists), G (globally), F (eventually) and X (next). The numbers of free and bound variables are assumed to be countable, and the numbers of functions and predicates are also assumed to be countable. It is also assumed that there is at least one predicate. A 0-ary function is an individual constant, and a 0-ary predicate is a propositional variable. Lower-case letters $p, q, ...$ are used to denote atomic formulas, Greek lower-case letters $\alpha, \beta, ...$ are used to denote formulas, and Greek capital letters $\Gamma, \Delta, ...$ are used to represent finite (possibly empty) sets of formulas. For any $\sharp \in \{\neg, \mathrm{G}, \mathrm{F}, \mathrm{X}\}$, an expression $\sharp\Gamma$ is used to denote the set $\{\sharp\gamma \mid \gamma \in \Gamma\}$. The symbol $\equiv$ is used to denote the equality of sets of symbols. The symbol ω is used to represent the set of natural numbers. An expression $\mathrm{X}^i\alpha$ for any $i \in \omega$ is inductively defined by $\mathrm{X}^0\alpha \equiv \alpha$ and $\mathrm{X}^{n+1}\alpha \equiv \mathrm{X}^n\mathrm{X}\alpha$. Lower-case letters i, j and k are used to denote any natural numbers. An expression of the form $\Gamma \Rightarrow \Delta$ is called a *sequent*. An expression $L \vdash S$ is used to denote the fact that a sequent S is provable in a sequent calculus L.

Kawai's sequent calculus LT_ω [6] for LTL is presented below.

Definition 1 ($\mathbf{LT}_\omega$). *The initial sequents of* LT_ω *are of the form:*

$$\alpha \Rightarrow \alpha.$$

The structural rules of LT_ω *are of the form:*

$$\frac{\Gamma \Rightarrow \Delta, \alpha \quad \alpha, \Sigma \Rightarrow \Pi}{\Gamma, \Sigma \Rightarrow \Delta, \Pi}\ (\mathrm{cut}) \qquad \frac{\Gamma \Rightarrow \Delta}{\alpha, \Gamma \Rightarrow \Delta}\ (\mathrm{we\text{-}left}) \qquad \frac{\Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta, \alpha}\ (\mathrm{we\text{-}right}).$$

The logical inference rules of LT_ω *are of the form:*

$$\frac{\Gamma \Rightarrow \Sigma, \mathrm{X}^i\alpha \quad \mathrm{X}^i\beta, \Delta \Rightarrow \Pi}{\mathrm{X}^i(\alpha \rightarrow \beta), \Gamma, \Delta \Rightarrow \Sigma, \Pi}\ (\rightarrow\mathrm{left}) \qquad \frac{\mathrm{X}^i\alpha, \Gamma \Rightarrow \Delta, \mathrm{X}^i\beta}{\Gamma \Rightarrow \Delta, \mathrm{X}^i(\alpha \rightarrow \beta)}\ (\rightarrow\mathrm{right})$$

$$\frac{\mathrm{X}^i\alpha, \Gamma \Rightarrow \Delta}{\mathrm{X}^i(\alpha \wedge \beta), \Gamma \Rightarrow \Delta}\ (\wedge\mathrm{left1}) \qquad \frac{\mathrm{X}^i\beta, \Gamma \Rightarrow \Delta}{\mathrm{X}^i(\alpha \wedge \beta), \Gamma \Rightarrow \Delta}\ (\wedge\mathrm{left2})$$

$$\frac{\Gamma \Rightarrow \Delta, \mathrm{X}^i\alpha \quad \Gamma \Rightarrow \Delta, \mathrm{X}^i\beta}{\Gamma \Rightarrow \Delta, \mathrm{X}^i(\alpha \wedge \beta)}\ (\wedge\text{right}) \qquad \frac{\mathrm{X}^i\alpha, \Gamma \Rightarrow \Delta \quad \mathrm{X}^i\beta, \Gamma \Rightarrow \Delta}{\mathrm{X}^i(\alpha \vee \beta), \Gamma \Rightarrow \Delta}\ (\vee\text{left})$$

$$\frac{\Gamma \Rightarrow \Delta, \mathrm{X}^i\alpha}{\Gamma \Rightarrow \Delta, \mathrm{X}^i(\alpha \vee \beta)}\ (\vee\text{right1}) \qquad \frac{\Gamma \Rightarrow \Delta, \mathrm{X}^i\beta}{\Gamma \Rightarrow \Delta, \mathrm{X}^i(\alpha \vee \beta)}\ (\vee\text{right2})$$

$$\frac{\Gamma \Rightarrow \Delta, \mathrm{X}^i\alpha}{\mathrm{X}^i\neg\alpha, \Gamma \Rightarrow \Delta}\ (\neg\text{left}) \qquad \frac{\mathrm{X}^i\alpha, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta, \mathrm{X}^i\neg\alpha}\ (\neg\text{right})$$

$$\frac{\mathrm{X}^i\alpha(t), \Gamma \Rightarrow \Delta}{\mathrm{X}^i\forall x\alpha(x), \Gamma \Rightarrow \Delta}\ (\forall\text{left}) \qquad \frac{\Gamma \Rightarrow \Delta, \mathrm{X}^i\alpha(a)}{\Gamma \Rightarrow \Delta, \mathrm{X}^i\forall x\alpha(x)}\ (\forall\text{right})$$

$$\frac{\mathrm{X}^i\alpha(a), \Gamma \Rightarrow \Delta}{\mathrm{X}^i\exists x\alpha(x), \Gamma \Rightarrow \Delta}\ (\exists\text{left}) \qquad \frac{\Gamma \Rightarrow \Delta, \mathrm{X}^i\alpha(t)}{\Gamma \Rightarrow \Delta, \mathrm{X}^i\exists x\alpha(x)}\ (\exists\text{right})$$

$$\frac{\mathrm{X}^{i+k}\alpha, \Gamma \Rightarrow \Delta}{\mathrm{X}^i\mathrm{G}\alpha, \Gamma \Rightarrow \Delta}\ (\text{Gleft}) \qquad \frac{\{\ \Gamma \Rightarrow \Delta, \mathrm{X}^{i+j}\alpha\ \}_{j\in\omega}}{\Gamma \Rightarrow \Delta, \mathrm{X}^i\mathrm{G}\alpha}\ (\text{Gright})$$

$$\frac{\{\ \mathrm{X}^{i+j}\alpha, \Gamma \Rightarrow \Delta\ \}_{j\in\omega}}{\mathrm{X}^i\mathrm{F}\alpha, \Gamma \Rightarrow \Delta}\ (\text{Fleft}) \qquad \frac{\Gamma \Rightarrow \Delta, \mathrm{X}^{i+k}\alpha}{\Gamma \Rightarrow \Delta, \mathrm{X}^i\mathrm{F}\alpha}\ (\text{Fright})$$

where a in ($\forall$right) *and* ($\exists$left) *is a free variable which must not occur in the lower sequents, and t in* ($\forall$left) *and* ($\exists$right) *is an arbitrary term.*

An intuitionistic version ILT_ω of LT_ω is defined below. For the inference rules of ILT_ω, we use the same names as those of LT_ω. The sequents of ILT_ω are of the form $\Gamma \Rightarrow \Delta$ where Δ is restricted to a single formula or an empty set. This restriction of sequents is the difference between ILT_ω and LT_ω. Due to this restriction, ILT_ω is a subsystem of LT_ω.

Definition 2 (ILT_ω). *In the following definition, Δ means a single formula or an empty set.*

The initial sequents of ILT_ω *are of the form:*

$$\alpha \Rightarrow \alpha.$$

The structural rules of ILT_ω *are of the form:*

$$\frac{\Gamma \Rightarrow \alpha \quad \alpha, \Sigma \Rightarrow \Delta}{\Gamma, \Sigma \Rightarrow \Delta}\ (\text{cut}) \qquad \frac{\Gamma \Rightarrow \Delta}{\alpha, \Gamma \Rightarrow \Delta}\ (\text{we-left}) \qquad \frac{\Gamma \Rightarrow}{\Gamma \Rightarrow \alpha}\ (\text{we-right}).$$

The logical inference rules of ILT_ω *are of the form:*

$$\frac{\Gamma \Rightarrow \mathrm{X}^i\alpha \quad \mathrm{X}^i\beta, \Sigma \Rightarrow \Delta}{\mathrm{X}^i(\alpha{\to}\beta), \Gamma, \Sigma \Rightarrow \Delta}\ (\to\text{left}) \qquad \frac{\mathrm{X}^i\alpha, \Gamma \Rightarrow \mathrm{X}^i\beta}{\Gamma \Rightarrow \mathrm{X}^i(\alpha{\to}\beta)}\ (\to\text{right})$$

$$\frac{\mathrm{X}^i\alpha, \Gamma \Rightarrow \Delta}{\mathrm{X}^i(\alpha \wedge \beta), \Gamma \Rightarrow \Delta}\ (\wedge\text{left1}) \qquad \frac{\mathrm{X}^i\beta, \Gamma \Rightarrow \Delta}{\mathrm{X}^i(\alpha \wedge \beta), \Gamma \Rightarrow \Delta}\ (\wedge\text{left2})$$

$$\frac{\Gamma \Rightarrow \mathrm{X}^i\alpha \quad \Gamma \Rightarrow \mathrm{X}^i\beta}{\Gamma \Rightarrow \mathrm{X}^i(\alpha \wedge \beta)}\ (\wedge\text{right}) \qquad \frac{\mathrm{X}^i\alpha, \Gamma \Rightarrow \Delta \quad \mathrm{X}^i\beta, \Gamma \Rightarrow \Delta}{\mathrm{X}^i(\alpha \vee \beta), \Gamma \Rightarrow \Delta}\ (\vee\text{left})$$

$$\frac{\Gamma \Rightarrow \mathrm{X}^i\alpha}{\Gamma \Rightarrow \mathrm{X}^i(\alpha \vee \beta)}\ (\vee\text{right1}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i\beta}{\Gamma \Rightarrow \mathrm{X}^i(\alpha \vee \beta)}\ (\vee\text{right2})$$

$$\frac{\Gamma \Rightarrow \mathrm{X}^i\alpha}{\mathrm{X}^i\neg\alpha, \Gamma \Rightarrow}\ (\neg\text{left}) \qquad \frac{\mathrm{X}^i\alpha, \Gamma \Rightarrow}{\Gamma \Rightarrow \mathrm{X}^i\neg\alpha}\ (\neg\text{right})$$

$$\frac{\mathrm{X}^i\alpha(t), \Gamma \Rightarrow \Delta}{\mathrm{X}^i\forall x\alpha(x), \Gamma \Rightarrow \Delta}\ (\forall\text{left}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i\alpha(a)}{\Gamma \Rightarrow \mathrm{X}^i\forall x\alpha(x)}\ (\forall\text{right})$$

$$\frac{\mathrm{X}^i\alpha(a), \Gamma \Rightarrow \Delta}{\mathrm{X}^i\exists x\alpha(x), \Gamma \Rightarrow \Delta}\ (\exists\text{left}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i\alpha(t)}{\Gamma \Rightarrow \mathrm{X}^i\exists x\alpha(x)}\ (\exists\text{right})$$

$$\frac{\mathrm{X}^{i+k}\alpha, \Gamma \Rightarrow \Delta}{\mathrm{X}^i\mathrm{G}\alpha, \Gamma \Rightarrow \Delta}\ (\text{Gleft}) \qquad \frac{\{\ \Gamma \Rightarrow \mathrm{X}^{i+j}\alpha\ \}_{j\in\omega}}{\Gamma \Rightarrow \mathrm{X}^i\mathrm{G}\alpha}\ (\text{Gright})$$

$$\frac{\{\ \mathrm{X}^{i+j}\alpha, \Gamma \Rightarrow \Delta\ \}_{j\in\omega}}{\mathrm{X}^i\mathrm{F}\alpha, \Gamma \Rightarrow \Delta}\ (\text{Fleft}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^{i+k}\alpha}{\Gamma \Rightarrow \mathrm{X}^i\mathrm{F}\alpha}\ (\text{Fright})$$

where a in ($\forall$right) *and* ($\exists$left) *is a free variable which must not occur in the lower sequents, and t in* ($\forall$left) *and* ($\exists$right) *is an arbitrary term.*

An expression $\alpha \Leftrightarrow \beta$ means the sequents $\alpha \Rightarrow \beta$ and $\beta \Rightarrow \alpha$.

Proposition 3. *The following sequents are provable in* ILT_ω*: for any formulas α, β and any $i \in \omega$,*

1. $\mathrm{X}^i(\alpha \circ \beta) \Leftrightarrow \mathrm{X}^i\alpha \circ \mathrm{X}^i\beta$ *where* $\circ \in \{\rightarrow, \wedge, \vee\}$,
2. $\mathrm{X}^i\neg\alpha \Leftrightarrow \neg\mathrm{X}^i\alpha$,
3. $\mathrm{G}\alpha \Rightarrow \mathrm{X}\alpha$,
4. $\mathrm{G}\alpha \Rightarrow \mathrm{XG}\alpha$,
5. $\mathrm{G}\alpha \Rightarrow \mathrm{GG}\alpha$,
6. $\alpha, \mathrm{G}(\alpha \rightarrow \mathrm{X}\alpha) \Rightarrow \mathrm{G}\alpha$ *(temporal induction).*

A rule R of inference is said to be *admissible* in a sequent calculus L if the following condition is satisfied: for any instance

$$\frac{S_1 \cdots S_n}{S}$$

of R, if $L \vdash S_i$ for all i, then $L \vdash S$. Moreover, R is said to be *derivable* in L if there is a derivation from $S_1, ..., S_n$ to S in L. Note that derivability implies admissibility.

The following cut-elimination theorems for LT_ω and ILT_ω are known [6,5]: Let L be LT_ω or ILT_ω. The rule (cut) of L is admissible in cut-free L.

Proposition 4. *Let Δ be a single formula or an empty set. Then, the rules of the form:*

$$\frac{\Gamma \Rightarrow \Sigma}{\mathrm{X}\Gamma \Rightarrow \mathrm{X}\Sigma}\ (\mathrm{Xregu}) \qquad \frac{\Sigma \Rightarrow \Delta}{\mathrm{X}\Sigma \Rightarrow \mathrm{X}\Delta}\ (\mathrm{Xregu})$$

are admissible in cut-free LT_ω *and cut-free* ILT_ω*, respectively.*

Propositions 3 and 4 will be used for proving the embedding theorem.

3 Embedding Theorem

Definition 5. *A mapping g on the set of formulas of the underlying logics is defined by*

1. *for any atomic formula p, $g(p) := \neg\neg p$,*
2. *$g(\alpha \sharp \beta) := g(\alpha) \sharp g(\beta)$ where $\sharp \in \{\rightarrow, \wedge\}$,*
3. *$g(\alpha \vee \beta) := \neg(\neg g(\alpha) \wedge \neg g(\beta))$,*
4. *$g(\sharp\alpha) := \sharp g(\alpha)$ where $\sharp \in \{\neg, \forall x, \mathrm{X}, \mathrm{G}\}$,*
5. *$g(\exists x\alpha) := \neg\forall x \neg g(\alpha)$,*
6. *$g(\mathrm{F}\alpha) := \neg\mathrm{G}\neg g(\alpha)$.*

An expression $g(\Gamma)$ denotes the result of replacing every occurrence of a formula α in Γ by an occurrence of $g(\alpha)$.

Lemma 6. *For any formula α,*

$$\mathrm{LT}_\omega \vdash \alpha \Leftrightarrow g(\alpha).$$

Proof. By induction on α. We show some cases.
Case ($\alpha \equiv \mathrm{X}\beta$): We show only $\mathrm{LT}_\omega \vdash \mathrm{X}\beta \Rightarrow g(\mathrm{X}\beta)$. By induction hypothesis, we have: $\mathrm{LT}_\omega \vdash \beta \Rightarrow g(\beta)$, and hence obtain the required fact:

$$\frac{\begin{array}{c}\vdots\\ \beta \Rightarrow g(\beta)\end{array}}{\mathrm{X}\beta \Rightarrow \mathrm{X}g(\beta)}\ (\mathrm{Xregu})$$

where $\mathrm{X}g(\beta)$ coincides with $g(\mathrm{X}\beta)$ by the definition of g.

Case ($\alpha \equiv \mathrm{G}\beta$): We show only $\mathrm{LT}_\omega \vdash \mathrm{G}\beta \Rightarrow g(\mathrm{G}\beta)$. By induction hypothesis, we have: $\mathrm{LT}_\omega \vdash \beta \Rightarrow g(\beta)$, and hence obtain the required fact:

$$\frac{\dfrac{\begin{array}{c}\vdots\\ \beta \Rightarrow g(\beta)\\ \vdots\ (\mathrm{Xregu})\end{array}}{\{\ \mathrm{X}^j\beta \Rightarrow \mathrm{X}^j g(\beta)\ \}_{j\in\omega}}\ (\mathrm{Gleft})}{\{\ \mathrm{G}\beta \Rightarrow \mathrm{X}^j g(\beta)\ \}_{j\in\omega}}\ (\mathrm{Gright})$$

$$\mathrm{G}\beta \Rightarrow \mathrm{G}g(\beta)$$

where $\mathrm{G}g(\beta)$ coincides with $g(\mathrm{G}\beta)$ by the definition of g.

Case ($\alpha \equiv \mathrm{F}\beta$): We show $\mathrm{LT}_\omega \vdash \mathrm{F}\beta \Leftrightarrow g(\mathrm{F}\beta)$. By induction hypothesis, we have: $\mathrm{LT}_\omega \vdash \beta \Leftrightarrow g(\beta)$, and hence obtain the required facts:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\vdots}{\beta \Rightarrow g(\beta)}}{\beta, \neg g(\beta) \Rightarrow}}{\{\ \mathrm{X}^j\beta, \mathrm{X}^j\neg g(\beta) \Rightarrow\ \}_{j\in\omega}}\ (\mathrm{Xregu})}{\{\ \mathrm{X}^j\beta, \mathrm{G}\neg g(\beta) \Rightarrow\ \}_{j\in\omega}}\ (\mathrm{Gleft})}{\mathrm{F}\beta, \mathrm{G}\neg g(\beta) \Rightarrow}\ (\mathrm{Fleft})}{\mathrm{F}\beta \Rightarrow \neg\mathrm{G}\neg g(\beta)} \qquad \dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\vdots}{g(\beta) \Rightarrow \beta}}{\Rightarrow \beta, \neg g(\beta)}}{\{\ \Rightarrow \mathrm{X}^j\beta, \mathrm{X}^j\neg g(\beta)\ \}_{j\in\omega}}\ (\mathrm{Xregu})}{\{\ \Rightarrow \mathrm{F}\beta, \mathrm{X}^j\neg g(\beta)\ \}_{j\in\omega}}\ (\mathrm{Fright})}{\Rightarrow \mathrm{F}\beta, \mathrm{G}\neg g(\beta)}\ (\mathrm{Gright})}{\neg\mathrm{G}\neg g(\beta) \Rightarrow \mathrm{F}\beta}$$

where $\neg\mathrm{G}\neg g(\beta)$ coincides with $g(\mathrm{F}\beta)$ by the definition of g. ∎

Lemma 7. *For any formula α,*

$$\mathrm{ILT}_\omega \vdash \neg\neg g(\alpha) \Rightarrow g(\alpha).$$

Proof. By induction on α.
Case ($\alpha \equiv \mathrm{X}\beta$): By induction hypothesis, we have: $\mathrm{ILT}_\omega \vdash \neg\neg g(\beta) \Rightarrow g(\beta)$, and hence obtain the required fact:

$$\dfrac{\dfrac{\dfrac{\vdots\ \text{Prop. 3 (2)}}{\mathrm{X}\neg g(\beta) \Rightarrow \neg\mathrm{X}g(\beta)}}{\mathrm{X}\neg g(\beta), \neg\neg\mathrm{X}g(\beta) \Rightarrow} \qquad \dfrac{\vdots}{\neg\neg g(\beta) \Rightarrow g(\beta)}}{\neg\neg\mathrm{X}g(\beta) \Rightarrow \mathrm{X}\neg\neg g(\beta)} \qquad \dfrac{}{\mathrm{X}\neg\neg g(\beta) \Rightarrow \mathrm{X}g(\beta)}\ (\mathrm{Xregu})}{\neg\neg\mathrm{X}g(\beta) \Rightarrow \mathrm{X}g(\beta)}\ (\mathrm{cut})$$

where $\mathrm{X}g(\beta)$ coincides with $g(\mathrm{X}\beta)$ by the definition of g.

Case ($\alpha \equiv \mathrm{G}\beta$): By induction hypothesis, we have: $\mathrm{ILT}_\omega \vdash \neg\neg g(\beta) \Rightarrow g(\beta)$, and hence obtain the required fact:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\{\ \mathrm{X}^j g(\beta) \Rightarrow \mathrm{X}^j g(\beta)\ \}_{j\in\omega}}{\{\ \mathrm{G}g(\beta) \Rightarrow \mathrm{X}^j g(\beta)\ \}_{j\in\omega}}\ (\mathrm{Gleft})}{\{\ \mathrm{G}g(\beta), \mathrm{X}^j\neg g(\beta) \Rightarrow\ \}_{j\in\omega}}}{\{\ \mathrm{X}^j\neg g(\beta) \Rightarrow \neg\mathrm{G}g(\beta)\ \}_{j\in\omega}}}{\{\ \mathrm{X}^j\neg g(\beta), \neg\neg\mathrm{G}g(\beta) \Rightarrow\ \}_{j\in\omega}}}{\{\ \neg\neg\mathrm{G}g(\beta) \Rightarrow \mathrm{X}^j\neg\neg g(\beta)\ \}_{j\in\omega}}}{\neg\neg\mathrm{G}g(\beta) \Rightarrow \mathrm{G}\neg\neg g(\beta)}\ (\mathrm{Gright}) \qquad \dfrac{\dfrac{\dfrac{\vdots}{\neg\neg g(\beta) \Rightarrow g(\beta)}}{\{\ \mathrm{X}^j\neg\neg g(\beta) \Rightarrow \mathrm{X}^j g(\beta)\ \}_{j\in\omega}}\ (\mathrm{Xregu})}{\{\ \mathrm{G}\neg\neg g(\beta) \Rightarrow \mathrm{X}^j g(\beta)\ \}_{j\in\omega}}\ (\mathrm{Gleft})}{\mathrm{G}\neg\neg g(\beta) \Rightarrow \mathrm{G}g(\beta)}\ (\mathrm{Gright})}{\neg\neg\mathrm{G}g(\beta) \Rightarrow \mathrm{G}g(\beta)}\ (\mathrm{cut})$$

where $\neg\neg\mathrm{G}g(\beta) \Rightarrow \mathrm{G}g(\beta)$ coincides with $\neg\neg g(\mathrm{G}\beta) \Rightarrow g(\mathrm{G}\beta)$ by the definition of g.

Case ($\alpha \equiv \mathrm{F}\beta$): We obtain the required fact:

$$\dfrac{\dfrac{\dfrac{\dfrac{\mathrm{G}\neg g(\beta) \Rightarrow \mathrm{G}\neg g(\beta)}{\neg\mathrm{G}\neg g(\beta), \mathrm{G}\neg g(\beta) \Rightarrow}}{\mathrm{G}\neg g(\beta) \Rightarrow \neg\neg\mathrm{G}\neg g(\beta)}}{\mathrm{G}\neg g(\beta), \neg\neg\neg\mathrm{G}\neg g(\beta) \Rightarrow}}{\neg\neg\neg\mathrm{G}\neg g(\beta) \Rightarrow \neg\mathrm{G}\neg g(\beta)}$$

where $\neg\neg\neg\mathrm{G}\neg g(\beta) \Rightarrow \neg\mathrm{G}\neg g(\beta)$ coincides with $\neg\neg g(\mathrm{F}\beta) \Rightarrow g(\mathrm{F}\beta)$ by the definition of g. ∎

Lemma 8. *For any sequent $\Sigma \Rightarrow \Pi$,*

if $\mathrm{LT}_\omega \vdash \Sigma \Rightarrow \Pi$, *then* $\mathrm{ILT}_\omega \vdash g(\Sigma), \neg g(\Pi) \Rightarrow$.

Proof. By induction on the proofs P of $\Sigma \Rightarrow \Pi$ in LT_ω. We distinguish the cases according to the last inference of P. We show some cases.

Case (Fleft): The last inference of P is of the form:

$$\dfrac{\{\ \mathrm{X}^{i+j}\alpha, \Gamma \Rightarrow \Delta\ \}_{j\in\omega}}{\mathrm{X}^i\mathrm{F}\alpha, \Gamma \Rightarrow \Delta}\ (\mathrm{Fleft}).$$

By induction hypothesis, we have: for all $j \in \omega$, $\mathrm{ILT}_\omega \vdash g(\mathrm{X}^{i+j}\alpha), g(\Gamma), \neg g(\Delta) \Rightarrow$ where $g(\mathrm{X}^{i+j}\alpha)$ coincides with $\mathrm{X}^{i+j}g(\alpha)$ by the definition of g. We thus obtain the required fact:

$$\dfrac{\dfrac{\dfrac{\vdots}{\{\ \mathrm{X}^{i+j}g(\alpha), g(\Gamma), \neg g(\Delta) \Rightarrow\ \}_{j\in\omega}}}{\{\ g(\Gamma), \neg g(\Delta) \Rightarrow \mathrm{X}^{i+j}\neg g(\alpha)\ \}_{j\in\omega}}}{g(\Gamma), \neg g(\Delta) \Rightarrow \mathrm{X}^i\mathrm{G}\neg g(\alpha)}\ (\mathrm{Gright})$$
$$\mathrm{X}^i\neg\mathrm{G}\neg g(\alpha), g(\Gamma), \neg(\Delta) \Rightarrow$$

where $\mathrm{X}^i\neg\mathrm{G}\neg g(\alpha)$ coincides with $g(\mathrm{X}^i\mathrm{F}\alpha)$ by the definition of g.

Case (Fright): The last inference of P is of the form:

$$\dfrac{\Gamma \Rightarrow \Delta, \mathrm{X}^{i+k}\alpha}{\Gamma \Rightarrow \Delta, \mathrm{X}^i\mathrm{F}\alpha}\ (\mathrm{Fright}).$$

By induction hypothesis, we have: $\mathrm{ILT}_\omega \vdash g(\Gamma), \neg g(\Delta), \neg g(\mathrm{X}^{i+k}\alpha) \Rightarrow$ where $\neg g(\mathrm{X}^{i+k}\alpha)$ coincides with $\neg\mathrm{X}^{i+k}g(\alpha)$ by the definition of g. We thus obtain the required fact:

$$\dfrac{\dfrac{\dfrac{\dfrac{\mathrm{X}^{i+k}\neg g(\alpha) \Rightarrow \neg\mathrm{X}^{i+k}g(\alpha) \quad g(\Gamma), \neg g(\Delta), \neg\mathrm{X}^{i+k}g(\alpha) \Rightarrow}{g(\Gamma), \neg g(\Delta), \mathrm{X}^{i+k}\neg g(\alpha) \Rightarrow}\ (\mathrm{Gleft})}{g(\Gamma), \neg g(\Delta), \mathrm{X}^i\mathrm{G}\neg g(\alpha) \Rightarrow}}{g(\Gamma), \neg g(\Delta) \Rightarrow \mathrm{X}^i\neg\mathrm{G}\neg g(\alpha)}}{g(\Gamma), \neg g(\Delta), \neg\mathrm{X}^i\neg\mathrm{G}\neg g(\alpha) \Rightarrow}\ (\mathrm{cut})$$

$\vdots$ Prop.3 (2) $\qquad$ $\vdots$

where $\neg \mathrm{X}^i \neg \mathrm{G} \neg g(\alpha)$ coincides with $\neg g(\mathrm{F}\alpha)$ by the definition of g.

Case (Gright): The last inference of P is of the form:

$$\frac{\{\ \Gamma \Rightarrow \Delta, \mathrm{X}^{i+j}\alpha\ \}_{j\in\omega}}{\Gamma \Rightarrow \Delta, \mathrm{X}^i \mathrm{G}\alpha}\ (\mathrm{Gright}).$$

By induction hypothesis, we have: for all $j \in \omega$, $\mathrm{ILT}_\omega \vdash g(\Gamma), \neg g(\Delta), \neg g(\mathrm{X}^{i+j}\alpha) \Rightarrow$ where $\neg g(\mathrm{X}^{i+j}\alpha)$ coincides with $\neg g(\mathrm{X}^{i+j}\alpha)$ by the definition of g. By Lemma 7, we also have: $\mathrm{ILT}_\omega \vdash \neg\neg g(\mathrm{X}^{i+j}\alpha) \Rightarrow g(\mathrm{X}^{i+j}\alpha)$. Thus, we obtain: for all $j \in \omega$,

$$\dfrac{\dfrac{\vdots}{\dfrac{g(\Gamma), \neg g(\Delta), \neg g(\mathrm{X}^{i+j}\alpha) \Rightarrow}{g(\Gamma), \neg g(\Delta) \Rightarrow \neg\neg g(\mathrm{X}^{i+j}\alpha)}} \qquad \begin{array}{c}\vdots\\ \neg\neg g(\mathrm{X}^{i+j}\alpha) \Rightarrow g(\mathrm{X}^{i+j}\alpha)\end{array}}{g(\Gamma), \neg g(\Delta) \Rightarrow g(\mathrm{X}^{i+j}\alpha)}\ (\mathrm{cut})$$

where $g(\mathrm{X}^{i+j}\alpha)$ coincides with $\mathrm{X}^{i+j} g(\alpha)$ by the definition of g. We then obtain the required fact:

$$\dfrac{\dfrac{\dfrac{\vdots}{\{\ g(\Gamma), \neg g(\Delta) \Rightarrow \mathrm{X}^{i+j} g(\alpha)\ \}_{j\in\omega}}}{g(\Gamma), \neg g(\Delta) \Rightarrow \mathrm{X}^i \mathrm{G} g(\alpha)}}{g(\Gamma), \neg g(\Delta), \neg \mathrm{X}^i \mathrm{G} g(\alpha) \Rightarrow}$$

where $\neg \mathrm{X}^i \mathrm{G} g(\alpha)$ coincides with $\neg g(\mathrm{X}^i \mathrm{G}\alpha)$ by the definition of g.

Case (Gleft): The last inference of P is of the form:

$$\frac{\mathrm{X}^{i+k}\alpha, \Gamma \Rightarrow \Delta}{\mathrm{X}^i \mathrm{G}\alpha, \Gamma \Rightarrow \Delta}\ (\mathrm{Gleft}).$$

By induction hypothesis, we have: $\mathrm{ILT}_\omega \vdash g(\mathrm{X}^{i+k}\alpha), g(\Gamma), \neg g(\Delta) \Rightarrow$ where $g(\mathrm{X}^{i+k}\alpha)$ coincides with $\mathrm{X}^{i+k} g(\alpha)$ by the definition of g. We then obtain the required fact:

$$\frac{\mathrm{X}^{i+k} g(\alpha), g(\Gamma), \neg g(\Delta) \Rightarrow}{\mathrm{X}^i \mathrm{G} g(\alpha), g(\Gamma), \neg g(\Delta) \Rightarrow}\ (\mathrm{Gleft})$$

where $\mathrm{X}^i \mathrm{G} g(\alpha)$ coincides with $g(\mathrm{X}^i \mathrm{G}\alpha)$ by the definition of g.

Case ($\neg$right): The last inference of P is of the form:

$$\frac{\mathrm{X}^i \alpha, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta, \mathrm{X}^i \neg\alpha}\ (\neg\mathrm{right}).$$

By induction hypothesis, we have: $\mathrm{ILT}_\omega \vdash g(\mathrm{X}^i\alpha), g(\Gamma), \neg g(\Delta) \Rightarrow$, and obtain:

$$\dfrac{\dfrac{\vdots}{g(\mathrm{X}^i\alpha), g(\Gamma), \neg g(\Delta) \Rightarrow}}{g(\Gamma), \neg g(\Delta) \Rightarrow \neg g(\mathrm{X}^i\alpha)}}{g(\Gamma), \neg g(\Delta), \neg\neg g(\mathrm{X}^i\alpha) \Rightarrow}$$

where $\neg\neg g(\mathrm{X}^i\alpha)$ coincides with $\neg\neg\mathrm{X}^i g(\alpha)$ by the definition of g. Thus, we obtain the required fact:

$$\dfrac{\dfrac{\dfrac{\vdots\ \text{Prop.3 (2)}}{\neg\mathrm{X}^i g(\alpha) \Rightarrow \mathrm{X}^i\neg g(\alpha)}}{\neg\mathrm{X}^i g(\alpha), \neg\mathrm{X}^i\neg g(\alpha) \Rightarrow} \qquad \vdots}{\neg\mathrm{X}^i\neg g(\alpha) \Rightarrow \neg\neg\mathrm{X}^i g(\alpha) \qquad g(\Gamma), \neg g(\Delta), \neg\neg\mathrm{X}^i g(\alpha) \Rightarrow}\ (\mathrm{cut})$$
$$g(\Gamma), \neg g(\Delta), \neg\mathrm{X}^i\neg g(\alpha) \Rightarrow$$

where $\neg\mathrm{X}^i\neg g(\alpha)$ coincides with $\neg g(\mathrm{X}^i\neg\alpha)$ by the definition of g.

Case ($\vee$right1): The last inference of P is of the form:

$$\dfrac{\Gamma \Rightarrow \Delta, \mathrm{X}^i\alpha}{\Gamma \Rightarrow \Delta, \mathrm{X}^i(\alpha \vee \beta)}\ (\vee\text{right1}).$$

By induction hypothesis, we have: $\mathrm{ILT}_\omega \vdash g(\Gamma), \neg g(\Delta), \neg g(\mathrm{X}^i\alpha) \Rightarrow$ where $\neg g(\mathrm{X}^i\alpha)$ coincides with $\neg\mathrm{X}^i g(\alpha)$ by the definition of g. We then obtain the required fact:

$$\dfrac{\dfrac{\vdots\ A}{\neg\mathrm{X}^i\neg(\neg g(\alpha) \wedge \neg g(\beta)) \Rightarrow \neg\mathrm{X}^i g(\alpha) \wedge \neg\mathrm{X}^i g(\beta)} \qquad \dfrac{\vdots}{g(\Gamma), \neg g(\Delta), \neg\mathrm{X}^i g(\alpha) \Rightarrow}}{g(\Gamma), \neg g(\beta), \neg\mathrm{X}^i\neg(\neg g(\alpha) \wedge \neg g(\beta)) \Rightarrow}\ (\mathrm{cut})$$

where $\neg\mathrm{X}^i\neg(\neg g(\alpha) \wedge \neg g(\beta))$ coincides with $\neg g(\mathrm{X}^i(\alpha \vee \beta))$ by the definition of g, and A is:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\mathrm{X}^i g(\alpha) \Rightarrow \mathrm{X}^i g(\alpha)}{\mathrm{X}^i\neg g(\alpha), \mathrm{X}^i g(\alpha) \Rightarrow}}{\mathrm{X}^i(\neg g(\alpha) \wedge \neg g(\beta))), \mathrm{X}^i g(\alpha) \Rightarrow}}{\mathrm{X}^i g(\alpha) \Rightarrow \mathrm{X}^i\neg(\neg g(\alpha) \wedge \neg g(\beta))}}{\neg\mathrm{X}^i\neg(\neg g(\alpha) \wedge \neg g(\beta)), \mathrm{X}^i g(\alpha) \Rightarrow}}{\neg\mathrm{X}^i\neg(\neg g(\alpha) \wedge \neg g(\beta)) \Rightarrow \neg\mathrm{X}^i g(\alpha)} \qquad \dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\mathrm{X}^i g(\beta) \Rightarrow \mathrm{X}^i g(\beta)}{\mathrm{X}^i\neg g(\beta), \mathrm{X}^i g(\beta) \Rightarrow}}{\mathrm{X}^i(\neg g(\alpha) \wedge \neg g(\beta))), \mathrm{X}^i g(\beta) \Rightarrow}}{\mathrm{X}^i g(\beta) \Rightarrow \mathrm{X}^i\neg(\neg g(\alpha) \wedge \neg g(\beta))}}{\neg\mathrm{X}^i\neg(\neg g(\alpha) \wedge \neg g(\beta)), \mathrm{X}^i g(\beta) \Rightarrow}}{\neg\mathrm{X}^i\neg(\neg g(\alpha) \wedge \neg g(\beta)) \Rightarrow \neg\mathrm{X}^i g(\beta)}}{\neg\mathrm{X}^i\neg(\neg g(\alpha) \wedge \neg g(\beta)) \Rightarrow \neg\mathrm{X}^i g(\alpha) \wedge \neg\mathrm{X}^i g(\beta)}.$$

We then obtain the following embedding theorem.

Theorem 9 (Embedding). *For any formula α,*

$$\mathrm{LT}_\omega \vdash \Rightarrow \alpha \ \textit{iff}\ \ \mathrm{ILT}_\omega \vdash \Rightarrow g(\alpha).$$

Proof. ($\Longrightarrow$): By Lemma 8. ($\Longleftarrow$): Suppose $\mathrm{ILT}_\omega \vdash \Rightarrow g(\alpha)$. Then, $\mathrm{LT}_\omega \vdash \Rightarrow g(\alpha)$, and hence $\mathrm{LT}_\omega \vdash \Rightarrow \alpha$ by Lemma 6 with (cut). ∎

4 Concluding Remarks

In this paper, a theorem for embedding LT_ω into ILT_ω was proved based on Baratella and Masini's temporal extension of the Gödel-Gentzen negative translation. The Baratella-Masini translation was originally applied to natural deduction systems for logics of positions. We showed in this paper that this translation can also be applied to Gentzen-type sequent calculi for first-order (classical and intuitionistic) LTLs. The result of this paper gives a bridge between classical temporal reasoning, which is useful for verifying concurrent systems, and intuitionistic temporal reasoning, which is useful for formalizing computational systems.

It is finally remarked that the embedding result of this paper can also be applied to Baratella and Masini's *2-sequent calculus* $2\mathrm{S}\omega$ [2] and its intuitionistic counterpart, although such an intuitionistic counterpart was not introduced in [2]. The system $2\mathrm{S}\omega$ for LTL is a natural extension of the usual sequent calculus. The cut-elimination and completeness theorems for this calculus were proved based on an analogy between LTL and Peano arithmetic with ω-rule. A direct syntactical equivalence between LT_ω and $2\mathrm{S}\omega$ was shown by Kamide introducing the translations that preserve cut-free proofs of these calculi [4]. Thus, by using these translations, we can obtain a theorem for embedding $2\mathrm{S}\omega$ into its intuitionistic counterpart.

Acknowledgments. This research was partially supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Young Scientists (B) 20700015.

References

1. Baratella, S., Masini, A.: A proof-theoretic investigation of a logic of positions. Annals of Pure and Applied Logic 123, 135–162 (2003)
2. Baratella, S., Masini, A.: An approach to infinitary temporal proof theory. Archive for Mathematical Logic 43(8), 965–990 (2004)
3. Ishihara, H.: A note on the Gödel-Gentzen translation. Mathematical Logic Quarterly 46(1), 135–137 (2000)
4. Kamide, N.: An equivalence between sequent calculi for linear-time temporal logic. Bulletin of the Section of the Logic 35(4), 187–194 (2006)
5. Kamide, N., Wansing, H.: Combining linear-time temporal logic with constructiveness and parconsistency. Journal of Applied Logic 8, 33–61 (2010)
6. Kawai, H.: Sequential calculus for a first order infinitary temporal logic. Zeitschrift für Mathematische Logik und Grundlagen der Mathematik 33, 423–432 (1987)
7. Pnueli, A.: The temporal logic of programs. In: Proceedings of the 18th IEEE Symposium on Foundations of Computer Science, pp. 46–57 (1977)
8. Troelstra, A.S., van Dalen, D.: Constructivism in mathematics, vol. 1. North-Holland Publishing Company, Amsterdam (1998)

A Decidable First-Order Logic for Medical Reasoning

Norihiro Kamide

Waseda Institute for Advanced Study, Waseda University,
1-6-1 Nishi Waseda, Shinjuku-ku, Tokyo 169-8050, Japan
drnkamide08@kpd.biglobe.ne.jp

Abstract. This paper is intended to construct a decidable first-order logic for appropriately expressing medical reasoning which may require to express not only time-dependency, paraconsistency, constructiveness, resource-sensitivity, but also order-sensitivity. A first-order temporal paraconsistent non-commutative logic is introduced as a Gentzen-type sequent calculus. This logic has no structural rules and has some bounded temporal operators and a paraconsistent negation connective. This logic is shown to be decidable and cut-eliminable.

Keywords: Medical reasoning, decidability, first-order logic, temporal logic, non-commutative logic.

1 Introduction

Medical reasoning is of growing importance in knowledge representation in AI, since medical databases and ontologies require appropriate reasoning mechanisms. Expressive and decidable logical systems for such medical reasoning have been required. For the expressiveness of such logics, some first-order and temporal expressions are needed in general, and moreover, some paraconsistent (or inconsistency-tolerant) expressions are also desired.

In this paper, a *first-order temporal paraconsistent non-commutative logic*, TPN[l], is introduced as a Gentzen-type sequent calculus. The logic TPN[l] can appropriately express medical reasoning which may require to express not only time-dependency, paraconsistency (or inconsistency-tolerance), constructiveness, resource-sensitivity, but also order-sensitivity. TPN[l] has no structural rules (i.e., contraction, exchange and weakening) and has some l-bounded linear-time temporal operators X (next), G (globally) and F (eventually) and a paraconsistent negation connective $\sim$. TPN[l] is shown to be decidable and cut-eliminable. The decidability of TPN[l] is an advantage over other proposals, because some standard first-order (classical and temporal) logics are undecidable. To obtain a decidable extended first-order temporal logic has been required for medical reasoning.

We adopt two techniques to obtain a decidable first-order temporal logic. One is to restrict the time domain ω of natural numbers by the set $\omega_l := \{x \in \omega \mid x \leq l\}$ with a fixed positive integer l [2]. The other is to use a *substructural*

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 235–245, 2011.

logic without contraction rule (e.g., a non-commutative logic) as a base logic. It is known that non-modal first-order substructual logics without contraction rule are decidable in general. For example, it was shown by Komori [7] that some first-order intuitionistic substructural logics without contraction rule are decidable. For a review of the decision problems for substructural logics, see [4].

The proposed logic TPN[l] is regarded as a modified extension of both the *temporal non-commutative logic* by Kamide [5] and the *constructive sequential propositional logic* (COSPL) by Wansing [11]. TPN[l] is also regarded as a bounded-time, constructive, paraconsistent and non-commutative modification of Kawai's sequent calculus LT_ω [6] for *linear-time temporal logic* (LTL). Although a number of *non-commutative logics*, which are substructural logics without exchange rule, have been proposed and studied by many researchers, the original and basic non-commutative logic is *Lambek calculus* [8], which has no structural rules. An enrichment FL of the Lambek calculus by full set of connectives has also been studied by many logicians. COSPL is a paraconsistent extension of FL.

By the virtue of paraconsistency and constructiveness in TPN[l], we can suitably express some vague and incomplete concepts such as symptoms and diseases which frequently appear in medical reasoning. By the virtue of first-order, time-dependent, resource-sensitive, and order-sensitive expressions of TPN[l], we can also suitably express the following situation: $\forall x\mathrm{G}(\mathit{meal}(x) * m(x) \rightarrow \mathrm{F}\ \mathit{recover}(x))$ which means "if a person x eats a meal and takes a medicine m in this order, then x will eventually make a recovery from the disease." In this expression, $*$ and $\rightarrow$ are an order-sensitive conjunction and a resource-sensitive implication, respectively. For a detailed explanation, see Section 3.

2 Logic

The following list of symbols is used for the language of the underlying logics: free variables $a_0, a_1, ...$, bound variables $x_0, x_1, ...$, functions $f_0, f_1, ...$, predicates $p_0, p_1, ...$, $\sim$ (paraconsistent negation), $\rightarrow$ (right implication), $\leftarrow$ (left implication), $\wedge$ (conjunction), $\vee$ (disjunction), $*$ (fusion), $\forall$ (any), $\exists$ (exists), X (next), G (globally) and F (eventually). The numbers of free and bound variables are assumed to be countable, and the numbers of functions and predicates are also assumed to be countable. It is also assumed that there is at least one predicate. A 0-ary function is an individual constant, and a 0-ary predicate is a propositional variable. Greek lower-case letters $\alpha, \beta, ...$ are used for formulas, and Greek capital letters $\Gamma, \Delta, ...$ are used for finite (possibly empty) sequences of formulas. For any $\sharp \in \{\sim, \mathrm{X}, \mathrm{G}, \mathrm{F}\}$, an expression $\sharp\Gamma$ is used to denote the sequence $\langle \sharp\gamma \mid \gamma \in \Gamma \rangle$. The symbol ω is used to represent the set of natural numbers. Let l be a fixed positive integer. Then, the symbol ω_l is used to represent the set $\{i \in \omega \mid i \leq l\}$. An expression $\mathrm{X}^i\alpha$ for any $i \in \omega$ is defined inductively by $(\mathrm{X}^0\alpha \equiv \alpha)$ and $(\mathrm{X}^{n+1}\alpha \equiv \mathrm{X}\mathrm{X}^n\alpha)$. Lower-case letters i and j are used to denote any natural numbers. An expression of the form $\Gamma \Rightarrow \gamma$ where γ is a single formula is called a *sequent*. If a sequent S is provable in a sequent calculus L, then such a fact is denoted as $L \vdash S$ or $\vdash S$. A rule R of inference is said to be *admissible* in a sequent calculus L if the following condition is satisfied: for any instance

$$\frac{S_1 \cdots S_n}{S}$$

of R, if $L \vdash S_i$ for all i, then $L \vdash S$.

A first-order temporal paraconsistent non-commutative logic, TPN[l], is introduced below.

Definition 1 (TPN[l]). *Let l be a fixed positive integer (called a* time bound*). The initial sequents of* TPN[l] *are of the form: for any atomic formula p,*

$$\mathrm{X}^i p \Rightarrow \mathrm{X}^i p \qquad\qquad \mathrm{X}^i{\sim}p \Rightarrow \mathrm{X}^i{\sim}p.$$

The cut rule of TPN[l] *is of the form:*

$$\frac{\Gamma \Rightarrow \alpha \quad \Sigma, \alpha, \Delta \Rightarrow \gamma}{\Sigma, \Gamma, \Delta \Rightarrow \gamma}\ (\mathrm{cut}).$$

The logical inference rules of TPN[l] *are of the form: for any $k \in \omega_l$ and any positive integer m,*

$$\frac{\Gamma \Rightarrow \mathrm{X}^i\alpha \quad \Sigma, \mathrm{X}^i\beta, \Delta \Rightarrow \gamma}{\Sigma, \mathrm{X}^i(\alpha{\rightarrow}\beta), \Gamma, \Delta \Rightarrow \gamma}\ (\rightarrow\mathrm{left}) \qquad \frac{\Gamma, \mathrm{X}^i\alpha \Rightarrow \mathrm{X}^i\beta}{\Gamma \Rightarrow \mathrm{X}^i(\alpha{\rightarrow}\beta)}\ (\rightarrow\mathrm{right})$$

$$\frac{\Gamma \Rightarrow \mathrm{X}^i\alpha \quad \Sigma, \mathrm{X}^i\beta, \Delta \Rightarrow \gamma}{\Sigma, \Gamma, \mathrm{X}^i(\alpha{\leftarrow}\beta), \Delta \Rightarrow \gamma}\ (\leftarrow\mathrm{left}) \qquad \frac{\mathrm{X}^i\alpha, \Gamma \Rightarrow \mathrm{X}^i\beta}{\Gamma \Rightarrow \mathrm{X}^i(\alpha{\leftarrow}\beta)}\ (\leftarrow\mathrm{right})$$

$$\frac{\Gamma, \mathrm{X}^i\alpha, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i(\alpha \wedge \beta), \Delta \Rightarrow \gamma}\ (\wedge\mathrm{left1}) \qquad \frac{\Gamma, \mathrm{X}^i\beta, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i(\alpha \wedge \beta), \Delta \Rightarrow \gamma}\ (\wedge\mathrm{left2})$$

$$\frac{\Gamma \Rightarrow \mathrm{X}^i\alpha \quad \Gamma \Rightarrow \mathrm{X}^i\beta}{\Gamma \Rightarrow \mathrm{X}^i(\alpha \wedge \beta)}\ (\wedge\mathrm{right}) \qquad \frac{\Gamma, \mathrm{X}^i\alpha, \Delta \Rightarrow \gamma \quad \Gamma, \mathrm{X}^i\beta, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i(\alpha \vee \beta), \Delta \Rightarrow \gamma}\ (\vee\mathrm{left})$$

$$\frac{\Gamma \Rightarrow \mathrm{X}^i\alpha}{\Gamma \Rightarrow \mathrm{X}^i(\alpha \vee \beta)}\ (\vee\mathrm{right1}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i\beta}{\Gamma \Rightarrow \mathrm{X}^i(\alpha \vee \beta)}\ (\vee\mathrm{right2})$$

$$\frac{\Gamma, \mathrm{X}^i\alpha, \mathrm{X}^i\beta, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i(\alpha * \beta), \Delta \Rightarrow \gamma}\ (*\mathrm{left}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i\alpha \quad \Delta \Rightarrow \mathrm{X}^i\beta}{\Gamma, \Delta \Rightarrow \mathrm{X}^i(\alpha * \beta)}\ (*\mathrm{right})$$

$$\frac{\Gamma, \mathrm{X}^i\alpha(t), \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i\forall x\alpha(x), \Delta \Rightarrow \gamma}\ (\forall\mathrm{left}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i\alpha(a)}{\Gamma \Rightarrow \mathrm{X}^i\forall x\alpha(x)}\ (\forall\mathrm{right})$$

$$\frac{\Gamma, \mathrm{X}^i\alpha(a), \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i\exists x\alpha(x), \Delta \Rightarrow \gamma}\ (\exists\mathrm{left}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i\alpha(t)}{\Gamma \Rightarrow \mathrm{X}^i\exists x\alpha(x)}\ (\exists\mathrm{right})$$

$$\frac{\Gamma, \mathrm{X}^l\alpha, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^{l+m}\alpha, \Delta \Rightarrow \gamma}\ (\mathrm{Xleft}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^l\alpha}{\Gamma \Rightarrow \mathrm{X}^{l+m}\alpha}\ (\mathrm{Xright})$$

$$\frac{\Gamma, \mathrm{X}^{i+k}\alpha, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i\mathrm{G}\alpha, \Delta \Rightarrow \gamma}\ (\mathrm{Gleft}) \qquad \frac{\{\ \Gamma \Rightarrow \mathrm{X}^{i+j}\alpha\ \}_{j\in\omega_l}}{\Gamma \Rightarrow \mathrm{X}^i\mathrm{G}\alpha}\ (\mathrm{Gright})$$

$$\frac{\{\ \Gamma, \mathrm{X}^{i+j}\alpha, \Delta \Rightarrow \gamma\ \}_{j\in\omega_l}}{\Gamma, \mathrm{X}^i\mathrm{F}\alpha, \Delta \Rightarrow \gamma}\ (\mathrm{Fleft}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^{i+k}\alpha}{\Gamma \Rightarrow \mathrm{X}^i\mathrm{F}\alpha}\ (\mathrm{Fright})$$

$$\frac{\Gamma, \mathrm{X}^i\alpha, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i{\sim}{\sim}\alpha, \Delta \Rightarrow \gamma}\ (\sim\sim\text{left}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i\alpha}{\Gamma \Rightarrow \mathrm{X}^i{\sim}{\sim}\alpha}\ (\sim\sim\text{right})$$

$$\frac{\Gamma, \mathrm{X}^i\alpha, \mathrm{X}^i{\sim}\beta, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i{\sim}(\alpha{\rightarrow}\beta), \Delta \Rightarrow \gamma}\ (\sim\rightarrow\text{left}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i\alpha \quad \Delta \Rightarrow \mathrm{X}^i{\sim}\beta}{\Gamma, \Delta \Rightarrow \mathrm{X}^i{\sim}(\alpha{\rightarrow}\beta)}\ (\sim\rightarrow\text{right})$$

$$\frac{\Gamma, \mathrm{X}^i{\sim}\beta, \mathrm{X}^i\alpha, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i{\sim}(\alpha{\leftarrow}\beta), \Delta \Rightarrow \gamma}\ (\sim\leftarrow\text{left}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i{\sim}\beta \quad \Delta \Rightarrow \mathrm{X}^i\alpha}{\Gamma, \Delta \Rightarrow \mathrm{X}^i{\sim}(\alpha{\leftarrow}\beta)}\ (\sim\leftarrow\text{right})$$

$$\frac{\Gamma, \mathrm{X}^i{\sim}\alpha, \Delta \Rightarrow \gamma \quad \Gamma, \mathrm{X}^i{\sim}\beta, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i{\sim}(\alpha \wedge \beta), \Delta \Rightarrow \gamma}\ (\sim \wedge \text{left})$$

$$\frac{\Gamma \Rightarrow \mathrm{X}^i{\sim}\alpha}{\Gamma \Rightarrow \mathrm{X}^i{\sim}(\alpha \wedge \beta)}\ (\sim \wedge \text{right1}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i{\sim}\beta}{\Gamma \Rightarrow \mathrm{X}^i{\sim}(\alpha \wedge \beta)}\ (\sim \wedge \text{right2})$$

$$\frac{\Gamma, \mathrm{X}^i{\sim}\alpha, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i{\sim}(\alpha \vee \beta), \Delta \Rightarrow \gamma}\ (\sim \vee \text{left1}) \qquad \frac{\Gamma, \mathrm{X}^i{\sim}\beta, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i{\sim}(\alpha \vee \beta), \Delta \Rightarrow \gamma}\ (\sim \vee \text{left2})$$

$$\frac{\Gamma \Rightarrow \mathrm{X}^i{\sim}\alpha \quad \Gamma \Rightarrow \mathrm{X}^i{\sim}\beta}{\Gamma \Rightarrow \mathrm{X}^i{\sim}(\alpha \vee \beta)}\ (\sim \vee \text{right})$$

$$\frac{\Gamma, \mathrm{X}^i{\sim}\alpha, \mathrm{X}^i{\sim}\beta, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i{\sim}(\alpha * \beta), \Delta \Rightarrow \gamma}\ (\sim * \text{left}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i{\sim}\alpha \quad \Delta \Rightarrow \mathrm{X}^i{\sim}\beta}{\Gamma, \Delta \Rightarrow \mathrm{X}^i{\sim}(\alpha * \beta)}\ (\sim * \text{right})$$

$$\frac{\Gamma, \mathrm{X}^i{\sim}\alpha(a), \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i{\sim}\forall x\alpha(x), \Delta \Rightarrow \gamma}\ (\sim\forall\text{left}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i{\sim}\alpha(t)}{\Gamma \Rightarrow \mathrm{X}^i{\sim}\forall x\alpha(x)}\ (\sim\forall\text{right})$$

$$\frac{\Gamma, \mathrm{X}^i{\sim}\alpha(t), \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i{\sim}\exists x\alpha(x), \Delta \Rightarrow \gamma}\ (\sim\exists\text{left}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i{\sim}\alpha(a)}{\Gamma \Rightarrow \mathrm{X}^i{\sim}\exists x\alpha(x)}\ (\sim\exists\text{right})$$

$$\frac{\Gamma, \mathrm{X}^i{\sim}\alpha, \Delta \Rightarrow \gamma}{\Gamma, {\sim}\mathrm{X}^i\alpha, \Delta \Rightarrow \gamma}\ (\sim\text{Xleft}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^i{\sim}\alpha}{\Gamma \Rightarrow {\sim}\mathrm{X}^i\alpha}\ (\sim\text{Xright})$$

$$\frac{\{\ \Gamma, \mathrm{X}^{i+j}{\sim}\alpha, \Delta \Rightarrow \gamma\ \}_{j\in\omega_l}}{\Gamma, \mathrm{X}^i{\sim}\mathrm{G}\alpha, \Delta \Rightarrow \gamma}\ (\sim\text{Gleft}) \qquad \frac{\Gamma \Rightarrow \mathrm{X}^{i+k}{\sim}\alpha}{\Gamma \Rightarrow \mathrm{X}^i{\sim}\mathrm{G}\alpha}\ (\sim\text{Gright})$$

$$\frac{\Gamma, \mathrm{X}^{i+k}{\sim}\alpha, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i{\sim}\mathrm{F}\alpha, \Delta \Rightarrow \gamma}\ (\sim\text{Fleft}) \qquad \frac{\{\ \Gamma \Rightarrow \mathrm{X}^{i+j}{\sim}\alpha\ \}_{j\in\omega_l}}{\Gamma \Rightarrow \mathrm{X}^i{\sim}\mathrm{F}\alpha}\ (\sim\text{Fright})$$

where a in ($\forall$right), ($\exists$left), ($\sim\exists$right) *and* ($\sim\forall$left) *is a free variable which must not occur in the lower sequents of the rules, and t in* ($\forall$left), ($\exists$right), ($\sim\exists$left) *and* ($\sim\forall$right) *is an arbitrary term.*

Remark that TPN[l] is just a logic parameterized by a fixed concrete positive integer l. Thus, before the detailed discussion, we have to fix TPN[l] as a concrete logic such as TPN[5]. Indeed, for example, TPN[2] is different from TPN[1]: $p \wedge \mathrm{X}p \Rightarrow \mathrm{G}p$ is provable in TPN[1], but it is not provable in TPN[2].

Let TPN[ω] be the logic obtained from TPN[l] by deleting {(Xleft), (Xright)} and replacing ω_l with ω. Then, TPN[ω] is regarded as a non-commutative, paraconsistent and intuitionistic version of Kawai's LT_ω [6]. Since the treatment of the infinite premises rules of TPN[ω] is somewhat difficult, we do not know whether TPN[ω] is decidable or not. Such a problem is remained as an open question.

An expression $\alpha \Leftrightarrow \beta$ means the sequents $\alpha \Rightarrow \beta$ and $\beta \Rightarrow \alpha$.

Proposition 2. *The following sequents are provable in* TPN[l]*: for any formulas* α, β*, any* $i \in \omega$ *and any positive integer* m*,*

1. $\mathrm{X}^i\alpha \Rightarrow \mathrm{X}^i\alpha$,
2. $\mathrm{X}^i(\alpha \circ \beta) \Leftrightarrow \mathrm{X}^i\alpha \circ \mathrm{X}^i\beta$ *where* $\circ \in \{\rightarrow, \leftarrow, \wedge, \vee, *\}$,
3. $\mathrm{X}^i Q\alpha(x) \Leftrightarrow Q\mathrm{X}^i\alpha(x)$ *where* $Q \in \{\forall x, \exists x\}$,
4. $\mathrm{X}^i T\alpha \Leftrightarrow T\mathrm{X}^i\alpha$ *where* $T \in \{\mathrm{G}, \mathrm{F}\}$,
5. $\mathrm{G}\alpha \Rightarrow \mathrm{F}\alpha$,
6. $\mathrm{G}\alpha \Rightarrow \mathrm{X}\alpha$,
7. $\mathrm{G}\alpha \Rightarrow \mathrm{XG}\alpha$,
8. $\mathrm{G}\alpha \Rightarrow \mathrm{GG}\alpha$,
9. $\mathrm{X}^{l+m}\alpha \Leftrightarrow \mathrm{X}^l\alpha$,
10. $\mathrm{G}\alpha \Leftrightarrow (\alpha \wedge \mathrm{X}\alpha \wedge \mathrm{X}^2\alpha \wedge \cdots \wedge \mathrm{X}^l\alpha)$,
11. $\mathrm{F}\alpha \Leftrightarrow (\alpha \vee \mathrm{X}\alpha \vee \mathrm{X}^2\alpha \vee \cdots \vee \mathrm{X}^l\alpha)$,
12. $\sim\sim\alpha \Leftrightarrow \alpha$,
13. $\sim(\alpha\rightarrow\beta) \Leftrightarrow \alpha * \sim\beta$,
14. $\sim(\alpha\leftarrow\beta) \Leftrightarrow \sim\beta * \alpha$,
15. $\sim(\alpha \wedge \beta) \Leftrightarrow \sim\alpha \vee \sim\beta$,
16. $\sim(\alpha \vee \beta) \Leftrightarrow \sim\alpha \wedge \sim\beta$,
17. $\sim(\alpha * \beta) \Leftrightarrow \sim\alpha * \sim\beta$,
18. $\sim\forall x\alpha(x) \Leftrightarrow \exists x\sim\alpha(x)$,
19. $\sim\exists x\alpha(x) \Leftrightarrow \forall x\sim\alpha(x)$,
20. $\sim\mathrm{X}\alpha \Leftrightarrow \mathrm{X}\sim\alpha$,
21. $\sim\mathrm{G}\alpha \Leftrightarrow \mathrm{F}\sim\alpha$,
22. $\sim\mathrm{F}\alpha \Leftrightarrow \mathrm{G}\sim\alpha$.

In order to obtain a translation of TPN[l] into its non-temporal part, we introduce the logic PN (*first-order paraconsistent non-commutative logic*) which is a first-order extension of the constant-free fragment of COSPL [11]. PN enjoys cut-elimination and is decidable.

Definition 3 (PN). *The logic* PN *is obtained from* TPN[l] *by deleting* (Xleft), (Xright), (Gleft), (Gright), (Fleft), (Fright), (∼Xleft), (∼Xright), (∼Gleft), (∼Gright), (∼Fleft), (∼Fright) *and replacing* X^i *with* X^0 *(i.e., all the occurrences of* X *in the inference rules are deleted). The modified inference rules for* PN *by replacing* i *with* 0 *are denoted by labeling "PN" in superscript position, e.g.,* $(\rightarrow\mathrm{left}^{PN})$.

3 Medical Reasoning

Paraconsistency. It is known that logics with paraconsistency can deal with inconsistency-tolerant reasoning more appropriately. An example using paraconsistency is briefly explained below. Assume a large medical knowledge-base MKB of symptoms and diseases, such as an expert system based on TPN[l]. It can also be assumed that MKB is inconsistent in the sense that there is a symptom predicate $s(x)$ such that $\sim s(x), s(x) \in MKB$, where $\sim s(x)$ means "a

person x does not have a symptom s." This assumption is very realistic, because symptom is a vague concept, which is difficult to determine by any diagnosis. Then, MKB does not derive arbitrary disease $d(x)$, which means "a person x suffers form a disease d", since paraconsistency ensures the fact that for some formulas α and β, both the sequents $\sim\alpha, \alpha \Rightarrow \beta$ and $\alpha, \sim\alpha \Rightarrow \beta$ are not provable. The paraconsistent TPN[l]-based MKB is thus inconsistency-tolerant. In the classical and intuitionistic logics, the sequent $\sim s(x), s(x) \Rightarrow d(x)$ is provable for any disease d, and hence the non-paraconsistent formulation based on the logics are regarded as inappropriate to the application of medical knowledge base.

Constructiveness. It is known that the following property of *constructible falsity* guarantees the constructiveness of the underlying negation connective [9,11]: If $\Rightarrow \sim(\alpha \wedge \beta)$ is provable, then either $\Rightarrow \sim\alpha$ or $\Rightarrow \sim\beta$ is provable. The disjunction connective $\vee$ of the intuitionistic logic is known to be constructive, since it has the disjunction property: If $\Rightarrow \alpha \vee \beta$ is provable, then either $\Rightarrow \alpha$ or $\Rightarrow \beta$ is provable. It is remark that both the constructible falsity and the disjunction property hold for TPN[l] and that these properties does not hold for classical logic. Both the properties for TPN[l] are derived from the cut-elimination theorem for TPN[l]. The constructible falsity, which does not hold for the intuitionistic logic, is regarded as the dual notion of the disjunction property. It is also known that logics with this property can allow to express *inexact predicates*. An inexact predicate is an incomplete predicate in an empirical domain. An example of an inexact predicate is a disease or symptom predicate such as $melancholia(x)$, which means "a person x suffers from the first-stage melancholia." This predicate is incomplete in the sense that we can not determine exactly that the formula $\sim melancholia(x) \vee melancholia(x)$ is true. For more detailed discussions and examples, see e.g., [10].

Resource-Sensitivity. It is known that logics without the contraction rule:

$$\frac{\Gamma, \alpha, \alpha, \Delta \Rightarrow \gamma}{\Gamma, \alpha, \Delta \Rightarrow \gamma}\ (\mathrm{co})$$

can elegantly represent the concept of "resource consumption" [1,3]. For example, we consider a sequent: $coin, coin \Rightarrow coffee$, which means "if we expend two coins, then we can take a cup of coffee." Then, if assuming the classical or intuitionistic logic, this sequent is logically equivalent to the sequent: $coin \Rightarrow coffee$, because of the presence of the contraction rule. On the other hand, we desire to distinguish such two sequents in the sense of the "resource-sensitivity", i.e., one coin and two coins have the different effect as resources. It is noted that TPN[l] is one of such resource-sensitive logics, since it has no contraction rule.

An appropriate resource consumption example is medicine consumption in medical reasoning. Consider a medicine m as a resource. An expression $m(x) \Rightarrow recover(x)$ means "if a person x uses a medicine m to recover from a disease, then x makes a recovery from the disease with the medicine." In this case, $m(x), m(x) \Rightarrow recover(x)$ and $m(x) \Rightarrow recover(x)$ have the completely

different meaning in the real world, because two medicines and one medicine have the different effect in general.

Order-Sensitivity. In the case of medicine consumption discussed above, it may not be sufficent to consider the effects of medicines. For example, if we consider two distinct medicines m_1 and m_2, then the meanings of the following two expressions are ragarded as different: $m_1(x), m_2(x) \Rightarrow recover(x)$ and $m_2(x), m_1(x) \Rightarrow recover(x)$, because the order of using medicines change the effect of the medicines. In other words, the *time priority* of using medicines is more important in general. A more detailed example is expressed as follows. An expression $meal(x)$ means "a person x have a meal." Then, $m(x)$,$meal(x) \Rightarrow recover(x)$ and $meal(x), m(x) \Rightarrow recover(x)$ have the different meaning, i.e., the effect of the medicine m is different whether the medicine is used after or before the meal.

To express such fine-grained medical reasoning, we have to use a *non-commutative logic*, such as TPN[l], because, for example, logics with the exchange rule:

$$\frac{\Gamma, \beta, \alpha, \Delta \Rightarrow \gamma}{\Gamma, \alpha, \beta\Delta \Rightarrow \gamma} \text{ (ex)}$$

can not express the priority of the use of medicines. It can be known that in a sequent expression $\gamma_1, \gamma_2, ..., \gamma_n \Rightarrow \beta$ in TPN[l], the antecedent $(\gamma_1, \gamma_2, ..., \gamma_n)$ can express the time priority of consumming the resources $\gamma_1, \gamma_2, ..., \gamma_n$, in fact, $(\gamma_1, \gamma_2, ..., \gamma_n)$ is a sequence of formulas in TPN[l], since TPN[l] has no exchange rule. It is remarked that two sequents $\gamma_1, \gamma_2, ..., \gamma_n \Rightarrow \beta$ and $\gamma_1 * \gamma_2 * \cdots * \gamma_n \Rightarrow \beta$ are logically equivalent in TPN[l], and hence an expression $\gamma_1 * \gamma_2$ means "first γ_1 is consumed, next so is γ_2." It is also noted that in two expressions $\alpha{\rightarrow}\beta$ and $\alpha{\leftarrow}\beta$, the implications $\rightarrow$ and $\leftarrow$ represent resource consumption with priority, e.g., $\rightarrow$ means the consumption of (subscription) *descending order priority*, and $\leftarrow$ means the consumption of *ascending order priority*.

The following realistic order-sensitive, time-dependent and first-order expression, which represents a *liveness property*, can be obtained in TPN[l]:

$$\forall x \mathrm{G}(meal(x) * m(x) {\rightarrow} \mathrm{F}\ recover(x))$$

which means "if a person x eats a meal and takes a medicine m in this order, then x will eventually make a recovery from the disease."

4 Decidability

An expression like $\bigwedge\{\alpha_i \mid i \in \omega_l\}$ (or $\bigvee\{\alpha_i \mid i \in \omega_l\}$) where $\{\alpha_i \mid i \in \omega_l\}$ is a multiset means $\alpha_0 \wedge \alpha_1 \wedge \cdots \wedge \alpha_l$ (or $\alpha_0 \vee \alpha_1 \vee \cdots \vee \alpha_l$, respectively). For example, $\bigwedge\{\alpha, \alpha, \beta\}$ means $\alpha \wedge \alpha \wedge \beta$.

Definition 4. *We fix a countable non-empty set Φ of atomic formulas, and define the sets $\Phi_i := \{p_i \mid p \in \Phi\}$ $(i \in \omega)$ of atomic formulas with $p_0 := p$, i.e., $\Phi_0 := \Phi$. The language $\mathcal{L}_{\mathrm{TPN}[l]}$ (or the set of formulas) of* TPN[l] *is defined by*

using Φ, $\sim$, $\rightarrow, \leftarrow, \wedge, \vee, *$, $\forall, \exists$, X, G *and* F. *The language* $\mathcal{L}_{\mathrm{PN}}$ *of* PN *is defined by using* $\bigcup_{i\in\omega}\Phi_i$, $\sim$, $\rightarrow, \leftarrow$, $\wedge, \vee$, $*$, $\forall$ *and* $\exists$.

A mapping f *from* $\mathcal{L}_{\mathrm{TPN}[l]}$ *to* $\mathcal{L}_{\mathrm{PN}}$ *is defined by: for any* $i \in \omega$ *and any positive integer* m,

1. $f(\mathrm{X}^i p) := p_i \in \Phi_i$ *for any* $p \in \Phi$ *(especially,* $f(p) := p \in \Phi_0$*),*
2. $f(\mathrm{X}^i{\sim}\alpha) := {\sim}f(\mathrm{X}^i\alpha)$,
3. $f(\mathrm{X}^i(\alpha \circ \beta)) := f(\mathrm{X}^i\alpha) \circ f(\mathrm{X}^i\beta)$ *where* $\circ \in \{\rightarrow, \leftarrow, \wedge, \vee, *\}$,
4. $f(\mathrm{X}^i Q\alpha(x)) := Qf(\mathrm{X}^i\alpha(x))$ *where* $Q \in \{\forall x, \exists x\}$,
5. $f(\mathrm{X}^{l+m}\alpha) := f(\mathrm{X}^l\alpha)$,
6. $f(\mathrm{X}^i\mathrm{G}\alpha) := \bigwedge\{f(\mathrm{X}^{i+j}\alpha) \mid j \in \omega_l\}$,
7. $f(\mathrm{X}^i\mathrm{F}\alpha) := \bigvee\{f(\mathrm{X}^{i+j}\alpha) \mid j \in \omega_l\}$.

An expression $f(\Gamma)$ denotes the result of replacing every occurrence of a formula α in Γ by an occurrence of $f(\alpha)$.

Strictly speaking, the mapping f should be denoted as f_l since f is strongly dependent on the time bound l. Indeed, $f_3(\mathrm{G}p)$ and $f_5(\mathrm{G}p)$ are different. But, for the sake of brevity, a simple expression f will be used in the following.

Theorem 5 (Embedding). *Let* Γ *be a sequence of formulas in* $\mathcal{L}_{\mathrm{TPN}[l]}$, γ *be a formula in* $\mathcal{L}_{\mathrm{TPN}[l]}$, *and* f *be the mapping defined in Definition 4. Then:*

1. $\mathrm{TPN}[l] \vdash \Gamma \Rightarrow \gamma$ *iff* $\mathrm{PN} \vdash f(\Gamma) \Rightarrow f(\gamma)$.
2. $\mathrm{TPN}[l] - (\mathrm{cut}) \vdash \Gamma \Rightarrow \gamma$ *iff* $\mathrm{PN} - (\mathrm{cut}) \vdash f(\Gamma) \Rightarrow f(\gamma)$.

Proof. Since the case (2) can be obtained as a subproof of the case (1), we show only (1).

• ($\Longrightarrow$): By induction on the proofs P of $\Gamma \Rightarrow \gamma$ in $\mathrm{TPN}[l]$. We distinguish the cases according to the last inference of P, and show some cases.

Case ($\mathrm{X}^i{\sim}p \Rightarrow \mathrm{X}^i{\sim}p$): The last inference of P is of the form: $\mathrm{X}^i{\sim}p \Rightarrow \mathrm{X}^i{\sim}p$. In this case, we obtain $\mathrm{PN} \vdash f(\mathrm{X}^i{\sim}p) \Rightarrow f(\mathrm{X}^i{\sim}p)$, since $f(\mathrm{X}^i{\sim}p)$ coincides with ${\sim}p_i$ by the definition of f.

Case ($\rightarrow$left). The last inference of P is of the form:

$$\frac{\Gamma \Rightarrow \mathrm{X}^i\alpha \quad \Sigma, \mathrm{X}^i\beta, \Delta \Rightarrow \gamma}{\Sigma, \mathrm{X}^i(\alpha{\rightarrow}\beta), \Gamma, \Delta \Rightarrow \gamma}\ (\rightarrow\mathrm{left}).$$

By induction hypothesis, we have $\mathrm{PN} \vdash f(\Gamma) \Rightarrow f(\mathrm{X}^i\alpha)$ and $\mathrm{PN} \vdash f(\Sigma), f(\mathrm{X}^i\beta), f(\Delta) \Rightarrow f(\gamma)$. Then we obtain the required fact:

$$\frac{\begin{array}{cc}\vdots & \vdots \\ f(\Gamma) \Rightarrow f(\mathrm{X}^i\alpha) & f(\Sigma), f(\mathrm{X}^i\beta), f(\Delta) \Rightarrow f(\gamma)\end{array}}{f(\Sigma), f(\mathrm{X}^i\alpha){\rightarrow}f(\mathrm{X}^i\beta), f(\Gamma), f(\Delta) \Rightarrow f(\gamma)}\ (\rightarrow\mathrm{left}^{PN})$$

where $f(\mathrm{X}^i\alpha){\rightarrow}f(\mathrm{X}^i\beta)$ coincides with $f(\mathrm{X}^i(\alpha{\rightarrow}\beta))$ by the definition of f.

Case (Gleft). The last inference of P is of the form:

$$\frac{\Gamma, \mathrm{X}^{i+k}\alpha, \Delta \Rightarrow \gamma}{\Gamma, \mathrm{X}^i\mathrm{G}\alpha, \Delta \Rightarrow \gamma}\ (\mathrm{Gleft}).$$

By induction hypothesis, we have $\mathrm{PN} \vdash f(\Gamma), f(\mathrm{X}^{i+k}\alpha), f(\Delta) \Rightarrow f(\gamma)$, and hence obtain:

$$\dfrac{\begin{array}{c}\vdots\\ f(\Gamma), f(\mathrm{X}^{i+k}\alpha), f(\Delta) \Rightarrow f(\gamma)\\ \vdots\ (\wedge\mathrm{left1}^{PN})\ \textit{and}\ (\wedge\mathrm{left2}^{PN})\end{array}}{f(\Gamma), \bigwedge\{f(\mathrm{X}^{i+j}\alpha) \mid j \in \omega_l\}, f(\Delta) \Rightarrow f(\gamma)}$$

where $\bigwedge\{f(\mathrm{X}^{i+j}\alpha) \mid j \in \omega_l\}$ coincides with $f(\mathrm{X}^i\mathrm{G}\alpha)$ by the definition of f, and $f(\mathrm{X}^{i+k}\alpha) \in \{f(\mathrm{X}^{i+j}\alpha) \mid j \in \omega_l\}$. Remark that the case $i > l$ is also included in this proof. In such a case, $f(\mathrm{X}^{i+k}\alpha)$ and $\bigwedge\{f(\mathrm{X}^{i+j}\alpha) \mid j \in \omega_l\}$ mean $f(\mathrm{X}^l\alpha)$ and $\overbrace{f(\mathrm{X}^l\alpha) \wedge f(\mathrm{X}^l\alpha) \wedge \cdots \wedge f(\mathrm{X}^l\alpha)}^{l}$, respectively.

Case (Gright). The last inference of P is of the form:

$$\dfrac{\{\ \Gamma \Rightarrow \mathrm{X}^{i+j}\alpha\ \}_{j\in\omega_l}}{\Gamma \Rightarrow \mathrm{X}^i\mathrm{G}\alpha}\ (\mathrm{Gright}).$$

By induction hypothesis, we have $\mathrm{PN} \vdash f(\Gamma) \Rightarrow f(\mathrm{X}^{i+j}\alpha)$ for all $j \in \omega_l$. Let Φ be the multiset $\{f(\mathrm{X}^{i+j}\alpha) \mid j \in \omega_l\}$. We obtain

$$\dfrac{\begin{array}{c}\vdots\\ \{\ f(\Gamma) \Rightarrow f(\mathrm{X}^{i+j}\alpha)\ \}_{f(\mathrm{X}^{i+j}\alpha)\in\Phi}\\ \vdots\ (\wedge\mathrm{right}^{PN})\end{array}}{f(\Gamma) \Rightarrow \bigwedge \Phi}$$

where $\bigwedge\Phi$ coincides with $f(\mathrm{X}^i\mathrm{G}\alpha)$ by the definition of f.

Case ($\sim$Xleft). The last inference of P is of the form:

$$\dfrac{\Gamma, \mathrm{X}^i{\sim}\alpha, \Delta \Rightarrow \gamma}{\Gamma, {\sim}\mathrm{X}^i\alpha, \Delta \Rightarrow \gamma}\ (\sim\mathrm{Xleft}).$$

By induction hypothesis, we have $\mathrm{PN} \vdash f(\Gamma), f(\mathrm{X}^i{\sim}\alpha), f(\Delta) \Rightarrow f(\gamma)$, and hence obtain the required fact $\mathrm{PN} \vdash f(\Gamma), f({\sim}\mathrm{X}^i\alpha), f(\Delta) \Rightarrow f(\gamma)$ since $f({\sim}\mathrm{X}^i\alpha)$ coincides with $f(\mathrm{X}^i{\sim}\alpha)$ by the definition of f.

• ($\Longleftarrow$): By induction on the proofs Q of $f(\Gamma) \Rightarrow f(\gamma)$ in PN. We distinguish the cases according to the last inference of Q, and show only the following case.

Case (cut). The last inference of Q is of the form:

$$\dfrac{f(\Gamma) \Rightarrow \beta \quad f(\Delta_1), \beta, f(\Delta_2) \Rightarrow f(\gamma)}{f(\Delta_1), f(\Gamma), f(\Delta_2) \Rightarrow f(\gamma)}\ (\mathrm{cut}).$$

Since β is in $\mathcal{L}_{\mathrm{PN}}$, we have the fact $\beta = f(\beta)$. This fact can be shown by induction on β. Then, by induction hypothesis, we have: $\mathrm{TPN}[l] \vdash \Gamma \Rightarrow \beta$ and $\mathrm{TPN}[l] \vdash \Delta_1, \beta, \Delta_2 \Rightarrow \gamma$. We then obtain the required fact: $\mathrm{TPN}[l] \vdash \Delta_1, \Gamma, \Delta_2 \Rightarrow \gamma$ by using (cut) in $\mathrm{TPN}[l]$. ∎

Theorem 6 (Cut-elimination). *The rule* (cut) *is admissible in cut-free* TPN[l].

Proof. Suppose TPN[l] $\vdash \Gamma \Rightarrow \gamma$. Then, we have PN $\vdash f(\Gamma) \Rightarrow f(\gamma)$ by Theorem 5 (1), and hence PN $-$ (cut) $\vdash f(\Gamma) \Rightarrow f(\gamma)$ by the cut-elimination theorem for PN. By Theorem 5 (2), we obtain TPN[l] $-$ (cut) $\vdash \Gamma \Rightarrow \gamma$. ∎

Theorem 7 (Decidability). TPN[l] *is decidable.*

Proof. By decidability of PN, for each α, it is possible to decide if $f(\alpha)$ is PN-provable. Then, by Theorem 5, TPN[l] is decidable. ∎

5 Concluding Remarks

In this paper, the logic TPN[l] was introduced as a Gentzen type sequent calculus. TPN[l] was shown to be decidable and cut-eliminable. It was shown that TPN[l] is useful for medical reasoning because it can appropriately express paraconsistency and constructiveness with respect to symptoms and diseases, and time-, resource- and order-sensitivities for medicine consumptions. An advantage of TPN[l] over other proposals is its decidability. This decidability was obtained based on the time-boundedness and non-commutativity of TPN[l].

It is remarked that we can obtain the decidability and cut-elimination results for the logics which are obtained from TPN[l] by adding any combinations of the exchange rule (ex) and the weakening rule:

$$\frac{\Gamma, \Delta \Rightarrow \gamma}{\Gamma, \alpha, \Delta \Rightarrow \gamma}\ \text{(we)}.$$

It is also remarked that an extension of TPN[l] + (ex) + (co), which is neither resource-sensitive nor order-sensitive, is undecidable.

A future work on this topic may be a semantical analysis of TPN[l]. As presented in [5], some phase-space semantics can naturally be obtained for some fragments of TPN[l], and as presented in [3], some Kripke-type semantics can be given for some expressive temporal substructural logics.

Acknowledgments. This research was partially supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Young Scientists (B) 20700015.

References

1. Girard, J.-Y.: Linear logic. Theoretical Computer Science 50, 1–102 (1987)
2. Hodkinson, I., Wolter, F., Zakharyaschev, M.: Decidable fragments of first-order temporal logics. Annals of Pure and Applied Logic 106, 85–134 (2000)
3. Kamide, N.: Linear and affine logics with temporal, spatial and epistemic operators. Theoretical Computer Science 353(1–3), 165–207 (2006)

4. Kamide, N.: Linear exponentials as resource operators: A decidable first-order linear logic with bounded exponentials. In: Hölldobler, S., Lutz, C., Wansing, H. (eds.) JELIA 2008. LNCS (LNAI), vol. 5293, pp. 245–257. Springer, Heidelberg (2008)
5. Kamide, N.: Temporal non-commutative logic: expressing time, resource, order and hierarchy. Logic and Logical Philosophy 18, 97–126 (2009)
6. Kawai, H.: Sequential calculus for a first order infinitary temporal logic. Zeitschrift für Mathematische Logik und Grundlagen der Mathematik 33, 423–432 (1987)
7. Komori, Y.: Predicate logics without the structure rules. Studia Logica 45, 393–404 (1986)
8. Lambek, J.: The mathematics of sentence structure. The American Mathematical Monthly 65, 154–170 (1958)
9. Nelson, D.: Constructible falsity. Journal of Symbolic Logic 14, 16–26 (1949)
10. Wagner, G.: Logic programming with strong negation and inexact predicates. Journal of Logic and Computation 1(6), 835–859 (1991)
11. Wansing, H.: The Logic of Information Structures. LNCS (LNAI), vol. 681, pp. 1–163. Springer, Heidelberg (1993)

Interpolation Theorems for Some Extended Description Logics

Norihiro Kamide

Waseda Institute for Advanced Study, Waseda University,
1-6-1 Nishi Waseda, Shinjuku-ku, Tokyo 169-8050, Japan
drnkamide08@kpd.biglobe.ne.jp

Abstract. Description logics have been studied as a logical foundation of web ontology languages. Interpolation theorems for description logics are known to be useful for extracting modular ontologies. In this paper, the interpolation theorems for two extended paraconsistent and temporal description logics are proved using some theorems for embedding these logics into a standard description logic.

1 Introduction

Description logics [4] including a standard description logic $\mathcal{ALC}$ [26] have been well-studied for a logical foundation of web ontology languages. *Paraconsistent (or inconsistency-tolerant) description logics*, which are extensions of some standard description logics, have been studied by several researchers [15–18, 20–22, 27, 32, 33, 10]. *Temporal description logics*, which are extensions of some standard description logics, have also been studied by many researchers (see e.g., [2, 14] for surveys and [5, 3, 31, 11] for some recent results).

Craig interpolation theorem for classical logic was originally introduced and proved in [7], and this theorem and its variations have been studied by many researchers for a number of non-classical logics. A strong version of the Craig interpolation theorem (called *uniform interpolation theorem*) for $\mathcal{ALC}$ was proved in [6]. The uniform interpolation theorem for $\mathcal{ALC}$ is known to be useful for extracting modular ontologies from a given large-scale ontology [12]. To show the Craig interpolation theorems for some extended description logics has been required.

In this paper, the Craig interpolation theorems for two paraconsistent and temporal description logics $\mathcal{PALC}$ and $\mathcal{XALC}$ are proved using some theorems for embedding these logics into $\mathcal{ALC}$. The logic $\mathcal{PALC}$, which was introduced in [10], is based on *Nelson's paraconsistent four-valued logic N4* [1, 19, 29]. The logic $\mathcal{XALC}$, which was introduced in [11], is based on *Prior's tomorrow tense logic* [23, 24]. Some theorems for embedding $\mathcal{PALC}$ and $\mathcal{XALC}$ into $\mathcal{ALC}$ were also proved in [10, 11]. In the present paper, we re-use these embedding theorems to show the Craig interpolation theorems for $\mathcal{PALC}$ and $\mathcal{XALC}$. Moreover, we need some critical lemmas for $\mathcal{PALC}$ and $\mathcal{XALC}$ to show the theorems. This paper is thus focused on proving such critical lemmas.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 246–255, 2011.

The contents of this paper are then summarized as follows. In Section 2, some preliminaries are given: The definition of $\mathcal{ALC}$ and the statement of the Craig interpolation theorem for $\mathcal{ALC}$ are addressed. In Section 3, $\mathcal{PALC}$ is introduced, the embedding theorem for $\mathcal{PALC}$ is reviewed, and the interpolation theorem for $\mathcal{PALC}$ is then proved by showing some critical lemmas. In Section 4, $\mathcal{XALC}$ is introduced, the embedding theorem for $\mathcal{XALC}$ is reviewed, and the interpolation theorem for $\mathcal{XALC}$ is then proved by showing some critical lemmas.

2 Preliminaries

In the following, we present the base logic $\mathcal{ALC}$. The $\mathcal{ALC}$-concepts are constructed from atomic concepts, roles, $\top$ (truth constant), $\bot$ (falsity constant), $\sqcap$ (intersection), $\sqcup$ (union), $\neg$ (classical negation or complement), $\forall R$ (universal concept quantification) and $\exists R$ (existential concept quantification). We use the letters A and A^j for atomic concepts, the letter R for roles, and the letters C and D for concepts.

Definition 1. Concepts C *are defined by the following grammar:*

$$C ::= \top \mid \bot \mid A \mid \neg C \mid C \sqcap C \mid C \sqcup C \mid \forall R.C \mid \exists R.C$$

Definition 2. *An* interpretation $\mathcal{I}$ *is a pair* $\langle \Delta^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$ *where*

1. $\Delta^{\mathcal{I}}$ *is a non-empty set,*
2. $\cdot^{\mathcal{I}}$ *is an interpretation function which assigns to every atomic concept* A *a set* $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ *and to every role* R *a binary relation* $R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$.

The interpretation function is extended to concepts by the following inductive definitions:

1. $\top^{\mathcal{I}} := \Delta^{\mathcal{I}}$,
2. $\bot^{\mathcal{I}} := \emptyset$,
3. $(\neg C)^{\mathcal{I}} := \Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$,
4. $(C \sqcap D)^{\mathcal{I}} := C^{\mathcal{I}} \cap D^{\mathcal{I}}$,
5. $(C \sqcup D)^{\mathcal{I}} := C^{\mathcal{I}} \cup D^{\mathcal{I}}$,
6. $(\forall R.C)^{\mathcal{I}} := \{a \in \Delta^{\mathcal{I}} \mid \forall b\ [(a,b) \in R^{\mathcal{I}} \Rightarrow b \in C^{\mathcal{I}}]\}$,
7. $(\exists R.C)^{\mathcal{I}} := \{a \in \Delta^{\mathcal{I}} \mid \exists b\ [(a,b) \in R^{\mathcal{I}} \wedge b \in C^{\mathcal{I}}]\}$.

An expression of the form $C \sqsubseteq D$ *for any concepts* C *and* D*, is called a* TBox statement. *An interpretation* $\mathcal{I} := \langle \Delta^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$ *is called a* model *of* $C \sqsubseteq D$ *(denoted as* $\mathcal{I} \models C \sqsubseteq D$*) if* $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$. *A TBox statement* $C \sqsubseteq D$ *is called* valid *in* $\mathcal{ALC}$ *(denoted as* $\models_{\mathcal{ALC}} C \sqsubseteq D$*) if* $\mathcal{I} \models C \sqsubseteq D$ *for any interpretations* $\mathcal{I}$.

An expression $V(C)$ denotes the set of all atomic concept symbols in a concept C.

The following theorem is a weak form of the uniform interpolation theorem for $\mathcal{ALC}$, which was proved in [6].

Theorem 3 (Interpolation). *For any concepts* C *and* D*, if* $\models_{\mathcal{ALC}} C \sqsubseteq D$*, then there exists a concept* E *such that*

1. $\models_{\mathcal{ALC}} C \sqsubseteq E$ *and* $\models_{\mathcal{ALC}} E \sqsubseteq D$,
2. $V(E) \subseteq V(C) \cap V(D)$.

3 Interpolation for $\mathcal{PALC}$

Similar notions and terminologies for $\mathcal{ALC}$ are also used for $\mathcal{PALC}$. The $\mathcal{PALC}$-concepts are constructed from the $\mathcal{ALC}$-concepts by adding $\sim$ (paraconsistent negation).

Definition 4. *A* paraconsistent interpretation $\mathcal{PI}$ *is a structure* $\langle \Delta^{\mathcal{PI}}, \cdot^{\mathcal{I}^+}, \cdot^{\mathcal{I}^-} \rangle$ *where*

1. $\Delta^{\mathcal{PI}}$ *is a non-empty set,*
2. $\cdot^{\mathcal{I}^+}$ *is an interpretation function which assigns to every atomic concept* A *a set* $A^{\mathcal{I}^+} \subseteq \Delta^{\mathcal{PI}}$ *and to every role* R *a binary relation* $R^{\mathcal{I}^+} \subseteq \Delta^{\mathcal{PI}} \times \Delta^{\mathcal{PI}}$,
3. $\cdot^{\mathcal{I}^-}$ *is an interpretation function which assigns to every atomic concept* A *a set* $A^{\mathcal{I}^-} \subseteq \Delta^{\mathcal{PI}}$ *and to every role* R *a binary relation* $R^{\mathcal{I}^-} \subseteq \Delta^{\mathcal{PI}} \times \Delta^{\mathcal{PI}}$,
4. *for any role* R, $R^{\mathcal{I}^+} = R^{\mathcal{I}^-}$.

The interpretation functions are extended to concepts by the following inductive definitions:

1. $\top^{\mathcal{I}^+} := \Delta^{\mathcal{PI}}$,
2. $\bot^{\mathcal{I}^+} := \emptyset$,
3. $(\sim C)^{\mathcal{I}^+} := C^{\mathcal{I}^-}$,
4. $(\neg C)^{\mathcal{I}^+} := \Delta^{\mathcal{PI}} \setminus C^{\mathcal{I}^+}$,
5. $(C \sqcap D)^{\mathcal{I}^+} := C^{\mathcal{I}^+} \cap D^{\mathcal{I}^+}$,
6. $(C \sqcup D)^{\mathcal{I}^+} := C^{\mathcal{I}^+} \cup D^{\mathcal{I}^+}$,
7. $(\forall R.C)^{\mathcal{I}^+} := \{a \in \Delta^{\mathcal{PI}} \mid \forall b\, [(a,b) \in R^{\mathcal{I}^+} \Rightarrow b \in C^{\mathcal{I}^+}]\}$,
8. $(\exists R.C)^{\mathcal{I}^+} := \{a \in \Delta^{\mathcal{PI}} \mid \exists b\, [(a,b) \in R^{\mathcal{I}^+} \wedge b \in C^{\mathcal{I}^+}]\}$,
9. $\top^{\mathcal{I}^-} := \emptyset$,
10. $\bot^{\mathcal{I}^-} := \Delta^{\mathcal{PI}}$,
11. $(\sim C)^{\mathcal{I}^-} := C^{\mathcal{I}^+}$,
12. $(\neg C)^{\mathcal{I}^-} := \Delta^{\mathcal{PI}} \setminus C^{\mathcal{I}^-}$,
13. $(C \sqcap D)^{\mathcal{I}^-} := C^{\mathcal{I}^-} \cup D^{\mathcal{I}^-}$,
14. $(C \sqcup D)^{\mathcal{I}^-} := C^{\mathcal{I}^-} \cap D^{\mathcal{I}^-}$,
15. $(\forall R.C)^{\mathcal{I}^-} := \{a \in \Delta^{\mathcal{PI}} \mid \exists b\, [(a,b) \in R^{\mathcal{I}^-} \wedge b \in C^{\mathcal{I}^-}]\}$,
16. $(\exists R.C)^{\mathcal{I}^-} := \{a \in \Delta^{\mathcal{PI}} \mid \forall b\, [(a,b) \in R^{\mathcal{I}^-} \Rightarrow b \in C^{\mathcal{I}^-}]\}$.

A paraconsistent interpretation $\mathcal{PI} := \langle \Delta^{\mathcal{PI}}, \cdot^{\mathcal{I}^+}, \cdot^{\mathcal{I}^-} \rangle$ *is called a* model *of* $C \sqsubseteq D$ *(denoted as* $\mathcal{PI} \models C \sqsubseteq D$*) if* $C^{\mathcal{I}^+} \subseteq D^{\mathcal{I}^+}$. *A TBox statement* $C \sqsubseteq D$ *is called* valid *in* $\mathcal{PALC}$ *(denoted as* $\models_{\mathcal{PALC}} C \sqsubseteq D$*) if* $\mathcal{PI} \models C \sqsubseteq D$ *for any paraconsistent interpretations* $\mathcal{PI}$.

The interpretation functions $\cdot^{\mathcal{I}^+}$ and $\cdot^{\mathcal{I}^-}$ are intended to represent "verification" (or "support of truth") and "falsification" (or "support of falsity"), respectively.

In the following, we introduce a translation of $\mathcal{PALC}$ into $\mathcal{ALC}$. This translation is almost the same as the translation presented in [10]. The translation is also a slight modification of the translation originally introduced by Ma et al. [15] to embed $\mathcal{ALC}4$ into $\mathcal{ALC}$. Some similar translations have been used by

Gurevich [9] and Rautenberg [25] to embed Nelson's constructive three-valued logic [1, 19] into intuitionistic logic, and have also been used by Wansing [30] to embed Nelson's paraconsistent four-valued logic [1, 19] into a modal logic.

Definition 5. *Let N_C be a non-empty set of atomic concepts, N'_C be the set $\{A' \mid A \in N_C\}$ of atomic concepts, and N_R be a non-empty set of roles. The language $\mathcal{L}^{\sim}$ of $\mathcal{PALC}$ is defined using N_C, N_R, $\top$, $\bot$, $\sim$, $\neg, \sqcap, \sqcup$, $\forall R$ and $\exists R$. The language $\mathcal{L}$ of $\mathcal{ALC}$ is obtained from $\mathcal{L}^{\sim}$ by adding N'_C and deleting $\sim$.*

A mapping f from $\mathcal{L}^{\sim}$ to $\mathcal{L}$ is defined inductively by

1. *for any $R \in N_R$, $f(R) := R$,*
2. *for any $A \in N_C$, $f(A) := A$ and $f(\sim A) := A' \in N'_C$,*
3. *$f(\natural) := \natural$ where $\natural \in \{\top, \bot\}$,*
4. *$f(\neg C) := \neg f(C)$,*
5. *$f(C \sharp D) := f(C) \sharp f(D)$ where $\sharp \in \{\sqcap, \sqcup\}$,*
6. *$f(\forall R.C) := \forall R.f(C)$,*
7. *$f(\exists R.C) := \exists R.f(C)$,*
8. *$f(\sim\top) := \bot$,*
9. *$f(\sim\bot) := \top$,*
10. *$f(\sim\sim C) := f(C)$,*
11. *$f(\sim\neg C) := \neg f(\sim C)$,*
12. *$f(\sim(C \sqcap D)) := f(\sim C) \sqcup f(\sim D)$,*
13. *$f(\sim(C \sqcup D)) := f(\sim C) \sqcap f(\sim D)$,*
14. *$f(\sim\forall R.C) := \exists R.f(\sim C)$,*
15. *$f(\sim\exists R.C) := \forall R.f(\sim C)$.*

Using Definition 5, we show Theorem 6, which is a slight modification of the semantical embedding theorem in [10], and the proof of Theorem 6 can thus be obtained similarly as in [10].

Theorem 6 (Embedding). *Let f be the mapping defined in Definition 5. For any TBox statement $C \sqsubseteq D$,*

$$\models_{\mathcal{PALC}} C \sqsubseteq D \quad \textit{iff} \quad \models_{\mathcal{ALC}} f(C) \sqsubseteq f(D).$$

We now start to prove the interpolation theorem for $\mathcal{PALC}$.

Lemma 7. *Let f be the mapping defined in Definition 5. For any atomic concept A in $\mathcal{L}^{\sim}$ and any concept C in $\mathcal{L}^{\sim}$,*

1. *$A \in V(C)$ iff $B \in V(f(C))$ for some $B \in \{A, A'\}$,*
2. *$A \in V(\sim C)$ iff $B \in V(f(\sim C))$ for some $B \in \{A, A'\}$.*

Proof. By (simultaneous) induction on C. We show some cases.

• Case ($C \equiv A$). For 1, we have: $A \in V(A)$ and $A = f(A) \in V(f(A))$ by the definition of f. For 2, we have: $A \in V(\sim A)$ and $A' = f(\sim A) \in V(f(\sim A))$ by the definition of f.

• Case ($C \equiv \sim D$). For 1, we obtain: $A \in V(\sim D)$ iff $B \in V(f(\sim D))$ for some $B \in \{A, A'\}$ (by induction hypothesis for 2). For 2, we obtain:

$A \in V(\sim\sim D)$
iff $A \in V(D)$
iff $B \in V(f(D))$ for some $B \in \{A, A'\}$ (by induction hypothesis for 1)
iff $B \in V(f(\sim\sim D))$ for some $B \in \{A, A'\}$ (by the definition of f).

- Case ($C \equiv \neg D$). For 1, we obtain:

$A \in V(\neg D)$
iff $A \in V(D)$
iff $B \in V(f(D))$ for some $B \in \{A, A'\}$ (by induction hypothesis for 1)
iff $B \in V(\neg f(D))$ for some $B \in \{A, A'\}$
iff $B \in V(f(\neg D))$ for some $B \in \{A, A'\}$ (by the definition of f).

For 2, we obtain:

$A \in V(\sim\neg D)$
iff $A \in V(\sim D)$
iff $B \in V(f(\sim D))$ for some $B \in \{A, A'\}$ (by induction hypothesis for 2)
iff $B \in V(\neg f(\sim D))$ for some $B \in \{A, A'\}$
iff $B \in V(f(\sim\neg D))$ for some $B \in \{A, A'\}$ (by the definition of f).

- Case ($C \equiv \forall R.D$). For 1, we obtain:

$A \in V(\forall R.D)$
iff $A \in V(D)$
iff $B \in V(f(D))$ for some $B \in \{A, A'\}$ (by induction hypothesis for 1)
iff $B \in V(\forall R.f(D))$ for some $B \in \{A, A'\}$
iff $B \in V(f(\forall R.D))$ for some $B \in \{A, A'\}$ (by the definition of f).

For 2, we obtain:

$A \in V(\sim\forall R.D)$
iff $A \in V(\sim D)$
iff $B \in V(f(\sim D))$ for some $B \in \{A, A'\}$ (by induction hypothesis for 2)
iff $B \in V(\exists R.f(\sim D))$ for some $B \in \{A, A'\}$
iff $B \in V(f(\sim\forall R.D))$ for some $B \in \{A, A'\}$ (by the definition of f). ∎

Lemma 8. *Let f be the mapping defined in Definition 5. For any concepts C and D in $\mathcal{L}^{\sim}$, if $V(f(C)) \subseteq V(f(D))$, then $V(C) \subseteq V(D)$.*

Proof. Suppose $A \in V(C)$. Then, we obtain $B \in V(f(C))$ for some $B \in \{A, A'\}$ by Lemma 7. By the assumption, we obtain $B \in V(f(D))$ for some $B \in \{A, A'\}$, and hence obtain $A \in V(D)$ by Lemma 7. ∎

Theorem 9 (Interpolation). *For any concepts C and D, if $\models_{\mathcal{PALC}} C \sqsubseteq D$, then there exists a concept E such that*

1. $\models_{\mathcal{PALC}} C \sqsubseteq E$ *and* $\models_{\mathcal{PALC}} E \sqsubseteq D$,
2. $V(E) \subseteq V(C) \cap V(D)$.

Proof. Suppose $\models_{\mathcal{PALC}} C \sqsubseteq D$. Then, we have $\models_{\mathcal{ALC}} f(C) \sqsubseteq f(D)$ by Theorem 6. By Theorem 3, we have the following: there exists a concept E in $\mathcal{L}$ such that

1. $\models_{\mathcal{ALC}} f(C) \sqsubseteq E$ and $\models_{\mathcal{ALC}} E \sqsubseteq f(D)$,
2. $V(E) \subseteq V(f(C)) \cap V(f(D))$.

Since E is in $\mathcal{L}$, we have the fact $E = f(E)$. This fact can be shown by induction on E. By Theorem 6, we thus obtain the following: there exists a concept E such that

1. $\models_{\mathcal{PALC}} C \sqsubseteq E$ and $\models_{\mathcal{PALC}} E \sqsubseteq D$,
2. $V(f(E)) \subseteq V(f(C)) \cap V(f(D))$.

Now it is sufficient to show that $V(f(E)) \subseteq V(f(C)) \cap V(f(D))$ implies $V(E) \subseteq V(C) \cap V(D)$. This is shown by Lemma 8. ∎

4 Interpolation for $\mathcal{XALC}$

Similar notions and terminologies for $\mathcal{ALC}$ are also used for $\mathcal{XALC}$. The symbol ω is used to represent the set of natural numbers. The $\mathcal{XALC}$-concepts are constructed from the $\mathcal{ALC}$-concepts by adding X (next-time operator). An expression $\mathrm{X}^n C$ is inductively defined by $\mathrm{X}^0 C := C$ and $\mathrm{X}^{n+1} C := \mathrm{X}\mathrm{X}^n C$.

Definition 10. *A* temporal interpretation $\mathcal{TI}$ *is a structure* $\langle \Delta^{\mathcal{TI}}, \{\cdot^{\mathcal{I}^i}\}_{i\in\omega}\rangle$ *where*

1. $\Delta^{\mathcal{TI}}$ *is a non-empty set,*
2. *each* $\cdot^{\mathcal{I}^i}$ $(i \in \omega)$ *is an interpretation function which assigns to every atomic concept* A *a set* $A^{\mathcal{I}^i} \subseteq \Delta^{\mathcal{TI}}$ *and to every role* R *a binary relation* $R^{\mathcal{I}^i} \subseteq \Delta^{\mathcal{TI}} \times \Delta^{\mathcal{TI}}$,
3. *for any role* R *and any* $i, j \in \omega$, $R^{\mathcal{I}^i} = R^{\mathcal{I}^j}$.

The interpretation function is extended to concepts by the following inductive definitions:

1. $\top^{\mathcal{I}^i} := \Delta^{\mathcal{TI}}$,
2. $\bot^{\mathcal{I}^i} := \emptyset$,
3. $(\mathrm{X}C)^{\mathcal{I}^i} := C^{\mathcal{I}^{i+1}}$,
4. $(\neg C)^{\mathcal{I}^i} := \Delta^{\mathcal{TI}} \setminus C^{\mathcal{I}^i}$,
5. $(C \sqcap D)^{\mathcal{I}^i} := C^{\mathcal{I}^i} \cap D^{\mathcal{I}^i}$,
6. $(C \sqcup D)^{\mathcal{I}^i} := C^{\mathcal{I}^i} \cup D^{\mathcal{I}^i}$,
7. $(\forall R.C)^{\mathcal{I}^i} := \{a \in \Delta^{\mathcal{TI}} \mid \forall b\ [(a,b) \in R^{\mathcal{I}^i} \Rightarrow b \in C^{\mathcal{I}^i}]\}$,
8. $(\exists R.C)^{\mathcal{I}^i} := \{a \in \Delta^{\mathcal{TI}} \mid \exists b\ [(a,b) \in R^{\mathcal{I}^i} \wedge b \in C^{\mathcal{I}^i}]\}$.

A temporal interpretation $\mathcal{TI} := \langle \Delta^{\mathcal{TI}}, \{\cdot^{\mathcal{I}^i}\}_{i\in\omega}\rangle$ *is called a* model *of* $C \sqsubseteq D$ *(denoted as* $\mathcal{TI} \models C \sqsubseteq D$*) if* $C^{\mathcal{I}^0} \subseteq D^{\mathcal{I}^0}$. *A TBox statement* $C \sqsubseteq D$ *is called* valid *in* $\mathcal{XALC}$ *(denoted as* $\models_{\mathcal{XALC}} C \sqsubseteq D$*) if* $\mathcal{TI} \models C \sqsubseteq D$ *for any temporal interpretations* $\mathcal{TI}$.

The interpretation functions $\cdot^{\mathcal{I}^i}$ are intended to represent "verification at a time point i".

Definition 11 and Theorem 12 presented below are slight modifications of those in [11].

Definition 11. *Let N_R be a non-empty set of roles, N_C be a non-empty set of atomic concepts, and N_C^i be the set $\{A^i \mid A \in N_C\}$ of atomic concepts where $A^0 = A$, i.e., $N_C^0 = N_C$. The language $\mathcal{L}^x$ of $\mathcal{XALC}$ is defined using N_C, N_R, $\top$, $\bot$, X, $\neg, \sqcap, \sqcup$, $\forall R$ and $\exists R$. The language $\mathcal{L}$ of $\mathcal{ALC}$ is obtained from $\mathcal{L}^x$ by adding $\bigcup_{i\in\omega} N_C^i$ and deleting* X.

A mapping f from $\mathcal{L}^x$ to $\mathcal{L}$ is defined inductively by

1. *for any $R \in N_R$, $f(R) := R$,*
2. *for any $A \in N_C$, $f(\mathrm{X}^i A) := A^i \in N_C^i$, esp. $f(A) := A$,*
3. *$f(\mathrm{X}^i \top) := \Delta^{\mathcal{TI}}$,*
4. *$f(\mathrm{X}^i \bot) := \emptyset$,*
5. *$f(\mathrm{X}^i \neg C) := \neg f(\mathrm{X}^i C)$,*
6. *$f(\mathrm{X}^i (C \sharp D)) := f(\mathrm{X}^i C) \sharp f(\mathrm{X}^i D)$ where $\sharp \in \{\sqcap, \sqcup\}$,*
7. *$f(\mathrm{X}^i \forall R.C) := \forall R.f(\mathrm{X}^i C)$,*
8. *$f(\mathrm{X}^i \exists R.C) := \exists R.f(\mathrm{X}^i C)$.*

Theorem 12 (Embedding). *Let f be the mapping defined in Definition 11. For any TBox statements $C \sqsubseteq D$,*

$$\models_{\mathcal{XALC}} C \sqsubseteq D \quad \textit{iff} \quad \models_{\mathcal{ALC}} f(C) \sqsubseteq f(D).$$

Lemma 13. *Let f be the mapping defined in Definition 11. For any $i \in \omega$, any atomic concept A in $\mathcal{L}^x$ and any concept C in $\mathcal{L}^x$,*

$$A \in V(\mathrm{X}^i C) \textit{ iff } A^j \in V(f(\mathrm{X}^i C)) \textit{ for some } j \in \omega.$$

Proof. By induction on C. We show some cases.

• Case ($C \equiv \mathrm{X}D$). By induction hypothesis, we have the required fact: $A \in V(\mathrm{X}^{i+1} D)$ iff $A^j \in V(f(\mathrm{X}^{i+1} D))$ for some $j \in \omega$.

• Case ($C \equiv D \sqcap E$). We obtain:

$A \in V(\mathrm{X}^i (D \sqcap E))$
iff $A \in V(\mathrm{X}^i D)$ or $A \in V(\mathrm{X}^i E)$
iff $[A^j \in V(f(\mathrm{X}^i D))$ for some $j \in \omega]$ or $[A^k \in V(f(\mathrm{X}^i E))$ for some $k \in \omega]$ (by induction hypothesis)
iff $A^l \in V(f(\mathrm{X}^i D) \wedge f(\mathrm{X}^i E))$ with $l \in \{j, k\}$
iff $A^l \in V(f(\mathrm{X}^i (D \sqcap E)))$ for some $l \in \omega$ (by the definition of f).

• Case ($C \equiv \forall R.D$). We obtain:

$A \in V(\mathrm{X}^i \forall R.D)$
iff $A \in V(\mathrm{X}^i D)$
iff $A^j \in V(f(\mathrm{X}^i D))$ for some $j \in \omega$ (by induction hypothesis)
iff $A^j \in V(\forall R.f(\mathrm{X}^i D))$ for some $j \in \omega$
iff $A^j \in V(f(\mathrm{X}^i \forall R.D))$ for some $j \in \omega$ (by the definition of f). ∎

Lemma 14. *Let f be the mapping defined in Definition 11. For any concepts C and D in $\mathcal{L}^x$, if $V(f(C)) \subseteq V(f(D))$, then $V(C) \subseteq V(D)$.*

Proof. Suppose $A \in V(C)$. Then, we obtain $A^j \in V(f(C))$ for some $j \in \omega$ by Lemma 13 taking 0 for i in X^i. By the assumption, we obtain $A^j \in V(f(D))$ for some $j \in \omega$, and hence obtain $A \in V(D)$ by Lemma 13 taking 0 for i in X^i. ∎

Theorem 15 (Interpolation). *For any concepts C and D, if $\models_{\mathcal{XALC}} C \sqsubseteq D$, then there exists a concept E such that*

1. *$\models_{\mathcal{XALC}} C \sqsubseteq E$ and $\models_{\mathcal{XALC}} E \sqsubseteq D$,*
2. *$V(E) \subseteq V(C) \cap V(D)$.*

Proof. Suppose $\models_{\mathcal{XALC}} C \sqsubseteq D$. Then, we have $\models_{\mathcal{ALC}} f(C) \sqsubseteq f(D)$ by Lemma 12. By Theorem 3, we have the following: there exists a concept E in $\mathcal{L}$ such that

1. $\models_{\mathcal{ALC}} f(C) \sqsubseteq E$ and $\models_{\mathcal{ALC}} E \sqsubseteq f(D)$,
2. $V(E) \subseteq V(f(C)) \cap V(f(D))$.

Since E is in $\mathcal{L}$, we have the fact $E = f(E)$. This fact can be shown by induction on E. By Lemma 12, we thus obtain the following: there exists a concept E such that

1. $\models_{\mathcal{XALC}} C \sqsubseteq E$ and $\models_{\mathcal{XALC}} E \sqsubseteq D$,
2. $V(f(E)) \subseteq V(f(C)) \cap V(f(D))$.

Now it is sufficient to show that $V(f(E)) \subseteq V(f(C)) \cap V(f(D))$ implies $V(E) \subseteq V(C) \cap V(D)$. This is shown by Lemma 14. ∎

5 Concluding Remarks

In this paper, the Craig interpolation theorems for $\mathcal{PALC}$ and $\mathcal{XALC}$ were proved using some critical lemmas which are based on the previously established embedding theorems into $\mathcal{ALC}$.

As mentioned in Section 1, uniform interpolation (or equivalently forgetting), which is a version of Craig interpolation, is known to be useful for extracting modular ontologies. Although the Craig interpolation theorems for $\mathcal{PALC}$ and $\mathcal{XALC}$ are weaker than the uniform interpolation theorems for them, the results for $\mathcal{PALC}$ and $\mathcal{XALC}$ may be regarded as a first step to showing their uniform interpolation theorems.

The rest of this section is devoted to give a brief survey of the studies of uniform interpolation. Uniform interpolation is a strengthening of Craig interpolation in that the interpolant can be obtained in some strong conditions. Uniform interpolation has been studied by many AI researchers as *semantic forgetting* (see e.g., [8] for a recent result). Semantic forgetting means that we can forget some redundant concepts in a given large ontology, preserving some relevant conditions on reasoning. Uniform interpolation (or forgetting) has been studied for some DLs such as DL-Lite and some extended $\mathcal{EL}$ (see e.g., [28, 13, 12]).

Acknowledgments. This research was partially supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Young Scientists (B) 20700015.

References

1. Almukdad, A., Nelson, D.: Constructible falsity and inexact predicates. Journal of Symbolic Logic 49, 231–233 (1984)
2. Artale, A., Franconi, E.: A survey of temporal extensions of description logics. Annals of Mathematics and Artificial Intelligence 30, 171–210 (2000)
3. Baader, F., Bauer, A., Lippmann, M.: Runtime verification using a temporal description logic. In: Ghilardi, S., Sebastiani, R. (eds.) FroCoS 2009. LNCS, vol. 5749, pp. 149–164. Springer, Heidelberg (2009)
4. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.F. (eds.): The description logic handbook: Theory, implementation and applications. Cambridge University Press, Cambridge (2003)
5. Baader, F., Ghilardi, S., Lutz, C.: LTL over description logic axioms. In: Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning (KR 2008), pp. 684–694. AAAI Press, Menlo Park (2008)
6. ten Cate, B., Conradie, W., Marx, M., Venema, Y.: Definitorially complete description logics. In: Proceedings of the 10th International Conference on Principles of Knowledge Representation and Reasoning (KR 2006), pp. 79–89. AAAI Press, Menlo Park (2006)
7. Craig, W.: Three uses of the Herbrand-Gentzen theorems in relating model theory and proof theory. Journal of Symbolic Logic 22(3), 269–285 (1957)
8. Eiter, T., Wang, K.: Semantic forgetting in answer set programming. Artificial Intelligence 172(14), 1644–1672 (2008)
9. Gurevich, Y.: Intuitionistic logic with strong negation. Studia Logica 36, 49–59 (1977)
10. Kamide, N.: Paraconsistent description logics revisited. In: Proceedings of the 23rd International Workshop on Description Logics (DL 2010), CEUR Workshop Proceedings, vol. 573 (2010)
11. Kamide, N.: A compatible approach to temporal description logics. In: Proceedings of the 23rd International Workshop on Description Logics (DL 2010), CEUR Workshop Proceedings, vol. 573 (2010)
12. Knov, B., Walther, D., Wolter, F.: Forgetting and uniform interpolation in large-scale description logic terminologies. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence, pp. 830–835 (2009)
13. Kontchakov, R., Wolter, F., Zakharyaschev, M.: Can you tell the difference between DL-Lite ontologies? In: Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning (KR 2008), pp. 285–295 (2008)
14. Lutz, C., Wolter, F., Zakharyashev, M.: Temporal description logics: A survey. In: Proceedings of the 15th International Symposium on Temporal Representation and Reasoning (TIME 2008), pp. 3–14. IEEE Computer Society, Los Alamitos (2008)
15. Ma, Y., Hitzler, P., Lin, Z.: Algorithms for paraconsistent reasoning with OWL. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 399–413. Springer, Heidelberg (2007)

16. Ma, Y., Hitzler, P., Lin, Z.: Paraconsistent reasoning for expressive and tractable description logics. In: Proceedings of the 21st International Workshop on Description Logic (DL 2008), CEUR Workshop Proceedings, vol. 353 (2008)
17. Meghini, C., Straccia, U.: A relevance terminological logic for information retrieval. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 197–205 (1996)
18. Meghini, C., Sebastiani, F., Straccia, U.: Mirlog: A logic for multimedia information retrieval. In: Uncertainty and Logics: Advanced Models for the Representation and Retrieval of Information, pp. 151–185. Kluwer Academic Publishing, Dordrecht (1998)
19. Nelson, D.: Constructible falsity. Journal of Symbolic Logic 14, 16–26 (1949)
20. Odintsov, S.P., Wansing, H.: Inconsistency-tolerant description logic: Motivation and basic systems. In: Hendricks, V.F., Malinowski, J. (eds.) Trends in Logic: 50 Years of Studia Logica, pp. 301–335. Kluwer Academic Publishers, Dordrecht (2003)
21. Odintsov, S.P., Wansing, H.: Inconsistency-tolerant Description Logic. Part II: Tableau Algorithms. Journal of Applied Logic 6, 343–360 (2008)
22. Patel-Schneider, P.F.: A four-valued semantics for terminological logics. Artificial Intelligence 38, 319–351 (1989)
23. Prior, A.N.: Time and modality. Clarendon Press, Oxford (1957)
24. Prior, A.N.: Past, present and future. Clarendon Press, Oxford (1967)
25. Rautenberg, W.: Klassische und nicht-klassische Aussagenlogik, Vieweg, Braunschweig (1979)
26. Schmidt-Schauss, M., Smolka, G.: Attributive concept descriptions with complements. Artificial Intelligence 48, 1–26 (1991)
27. Straccia, U.: A sequent calculus for reasoning in four-valued description logics. In: Galmiche, D. (ed.) TABLEAUX 1997. LNCS, vol. 1227, pp. 343–357. Springer, Heidelberg (1997)
28. Wang, Z., Wang, K., Topor, R.W., Pan, J.Z.: Forgetting concepts in DL-Lite. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 245–257. Springer, Heidelberg (2008)
29. Wansing, H.: The Logic of Information Structures. LNCS (LNAI), vol. 681, pp. 1–163. Springer, Heidelberg (1993)
30. Wansing, H.: Displaying The Modal Logic of Consistency. Journal of Symbolic Logic 64(4), 1573–1590 (1999)
31. Wolter, F., Zakharyashev, M.: Temporalizing description logic. In: Gabbay, D., de Rijke, M. (eds.) Frontiers of Combining Systems, pp. 379–402. Studies Press/Wiley (1999)
32. Zhang, X., Lin, Z.: Paraconsistent reasoning with quasi-classical semantics in $\mathcal{ALC}$. In: Calvanese, D., Lausen, G. (eds.) RR 2008. LNCS, vol. 5341, pp. 222–229. Springer, Heidelberg (2008)
33. Zhang, X., Qi, G., Ma, Y., Lin, Z.: Quasi-classical semantics for expressive description logics. In: Proceedings of the 22nd International Workshop on Description Logic (DL 2009), CEUR Workshop Proceedings, vol. 477 (2009)

Investigating Attachment Behavior of Nodes during Evolution of a Complex Social Network:

A Case of a Scientific Collaboration Network

Alireza Abbasi and Liaquat Hossain

Centre for Complex Systems Research,
Faculty of Engineering and IT, University of Sydney
Sydney, NSW 2006, Australia
{alireza.abbasi,liaquat.hossain}@sydney.edu.au

Abstract. Complex networks (systems) as a phenomenon can be observed by a wide range of networks in nature and society. There is a growing interest to study complex networks from the evolutionary and behavior perspective. Studies on evolving dynamical networks have been resulted in a class of models to explain their evolving dynamic behavior that indicate a new node attaches preferentially to some old nodes in the network based on their number of links. In this study, we aim to explore if there are any other characteristics of the old nodes which affect on the preferential attachment of new nodes. We explore the evolution of a co-authorship network over time and find that while the association between number of new attached nodes to an existing node and all its main centrality measures (i.e., degree, closeness and betweenness) is almost positive and significant but betweenness centrality correlation coefficient is always higher and increasing as network evolved over time. Identifying the attachment behavior of nodes in complex networks (e.g., traders, disease propagation and emergency management) help policy and decision makers to focus on the nodes (actors) in order to control the resources distribution, information dissemination, disease propagation and so on due to type of the network.

Keywords: Network evolution, preferential attachment, selection, social network analysis, centrality measures, co-authorship network.

1 Introduction

Complex networks (systems) surround our human life by a wide range of networks in nature and society. "Studying an evolving complex system and drawing some conclusions from it is an integral part of nature-inspired computing; being a part of that complex system, some insight can also be gained from our knowledge of it" [1]. By the increasing evidence that real networks obey unexpected scaling laws [2, 3], interpreted as signatures of deviation from randomness [4], there have been efforts resulted in a class of models that view networks as evolving dynamical systems, rather than static graphs. Those approaches, look for universalities in the dynamics governing their evolution [4].

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 256–264, 2011.

Most evolving network models are based on two ingredients [3]: growth and preferential attachment. The growth suggests that networks continuously expand by the addition of new nodes and links between the nodes, while the preferential attachment states that new nodes attach preferentially to existing (old) nodes that are already well connected or in other words, a new node is connected to some old node in the network based on its number of links [5]. This model indicates that "the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems" [3]. The preferential attachment ("rich get richer") principle was originally proposed by Yule [6] and was also elaborated by Simon [7].

Recent research into the development and evolution of networks suggest that evolution of networks not only consist of the growth of network (adding more nodes and links to the network) but also the preferential attachment behavior of the new nodes. Preferential attachment can be seen from how new nodes attach preferentially to existing nodes that are already well connected. Barabasi and Albert [3] proposed that the new nodes prefer to attach to nodes who are connected to more node (higher degree centrality) and they called this phenomena as "preferential attachment" which is based on the "the rich get richer" principal.

In the language of social network analysis language, based on preferential attachment rule, new nodes prefer to attach to the old nodes with higher degree centrality. On the other hand, each node's centrality measures (i.e., degree, closeness and betweenness) reflect specific characteristics of the node. In this paper, in order to find the characteristics of the old nodes which affect on the selection of them by new nodes (for attachment), we explore a study of collaboration (co-authorship) network and investigate how the nodes (authors) behave during network evolution. We use social network analysis centrality concepts and measures to answer following questions: How do nodes behave during co-authorship network evolution? How does positioning of existing nodes (and role) in the network impact differently on their growth during network evolution?

We hypothesize that nodes attach preferentially to the actors who are already well connected (high degree centrality) or that are already well close to all others (high closeness centrality) or that are already well intermediate to connect others to each other (high betweenness centrality). We calculate all actors' centrality measures in each period of time and the number of new individuals who has attached to them in the following time frame. Using correlation coefficient between actors' centrality measure and the number of attached nodes, we find that while almost all three measures shows a significantly positive association but the correlation between actors' betweenness centrality and the number of attachments is always positive and the coefficient is much higher than other measures.

The attachment behavior of nodes during network evolution is important for predicting and controlling the evolutionary behavior of whole network. For instance, identifying the nodes to which new nodes prefer to connect is very important to control spread of a disease in an infectious disease network or very useful for advertising purpose in marketing and business networks.

Social network analysis has produced many results concerning social influence, social groupings, inequality, disease propagation, communication of information, and indeed almost every topic that has interested 20th century sociology [8]. Social networks operate on many levels, from families up to the level of nations. They play a

critical role in determining the way problems are solved, organizations are run, markets evolve, and the degree to which individuals succeed in achieving their goals [9]. Social networks have been analyzed to identify areas of strengths and weaknesses within and among research organizations, businesses, and nations as well as to direct scientific development and funding policies [10, 11].

A method used to understand networks and their participants is to evaluate the location of nodes in the network. These measures help determine the importance of a node in the network. Bavelas [12] was the pioneer who initially investigates formal properties of centrality and proposed several centrality concepts. Later, Freeman [13] found that centrality has an important structural factor influencing leadership, satisfaction, and efficiency. The important node centrality measures are *Degree Centrality* which measures node centrality by number of directly connected links; *Closeness Centrality* which measures node centrality using the shortest distance of the node to other nodes in the network; *Betweenness Centrality* which measures node centrality using the number of shortest paths (between all pairs of nodes) that pass through a given node.

In the following section, data source and collection method have been explained in addition to the method and measures used in this study. Section 3 shows the results of our analysis to find if the characteristic of a node in a network is associated more with its attachment behavior during the evolution of network. The paper ends with conclusions and talking about research limitations and implications of this study.

2 Data and Methods

2.1 Data Sample

Scopus (www.scopus.com) is one of the main sources which present bibliometric data on their website. To construct our database for this study, publications were extracted using "steel structure" phrase in their titles or keywords or abstracts in top 17 specified journals of the field (shortlisted by an expert) and restricting to just English language publications. After extracting the publications meta-data, we used an application program for extracting publication information (i.e., title, publication date, author names, affiliations, publisher, number of citations, etc). It also extracts relationships (e.g., co-authorships) between researchers and stores the data in the format of tables in its local database.

After the cleansing of the publication data, the resulting database contained 2235 papers (which 9 of the papers had no author information) reflecting the contributions of 5251 authors (117 of authors had no affiliation data) from 1324 organizations (i.e., universities and private companies) of 83 countries.

2.2 Methodology: Social Network Analysis

Measuring the actors' location in a network is about determining the centrality of an actor. These measures help determine the importance of a node in the network. To quantify the importance of an actor in a social network, various centrality measures have been proposed over the years [14]. Freeman [13] defined centrality in terms of

point (degree), betweenness, and closeness, each having important implications on social outcome and processes. Freeman [13] found that centrality has an important structural factor influencing leadership, satisfaction, and efficiency.

A social network is a set of individuals or groups each of which has connections of some kind to some or all of the others [15]. In the language of social network analysis, the people or groups are called ''actors'' or "nodes" and the connections ''ties'' or "links". Both actors and ties can be defined in different ways depending on the questions of interest. An actor might be a single person, a team, or a company. A tie might be a friendship between two people, collaboration or common member between two teams, or a business relationship between companies [8]. In scientific collaboration network actors are authors and ties (links) are co-authorship relations among them. A tie exists between each two actor if two scholars have at least a co-authored paper.

2.3 Centrality Measures

- **Degree Centrality**

The degree is simply the number of other points connected directly to a point. Necessarily, a central point is not physically in the centre of the network. As degree of a point is calculated in terms of the number of its adjacent points, the degree can be regarded as a measure of local centrality [14]. Thus, degree centrality of point k (P_k) is given by:

$$C_D(p_k) = \sum_{i=1}^{n} a(p_i, p_k)$$

Where n is the number of points (nodes) in the network and $a(p_i, p_k) = 1$ if and only if point i [16] and k (p_k) are connected and $a(p_i, p_k) = 0$ otherwise.

Thus, a person (point) in a position with having high degree centrality can influence the group by withholding or distorting information in transmission [13, 17]. So, degree centrality is an indicator of an actor's communication activity or popularity. Having relative or normalized centrality measure, we can compare points in different networks but similar types of relations as the number of connection between points (nodes) vary based on network relation types (e.g., friendship, financial, knowledge exchange, etc).

- **Closeness Centrality**

Freeman [13, 18] proposed closeness in terms of the distance among various points. Sabidussi [19] used the same concept in his work as 'sum distance', the sum of the 'geodesic' distances (the shortest path between any particular pair of points in a network) to all other points in the network. A node is globally central if it lies at the shortest distance from many other nodes which means it is 'close' to many of the other points in the network. Simply by calculating the sum of distances of a point to others we will have 'farness', how far the point is from other points and then we need to use the inverse of farness as a measure of closeness. In unconnected networks as every point is at an infinite distance from at least one point, closeness centrality of all points would be 0. To solve this problem, Freeman proposed another way for calculating closeness of a point by *"sum of reciprocal distance"* of that point to any other nodes. So, closeness centrally of node k (p_k) is given by:

$$C_C(p_k) = \sum_{i=1}^{n} d(p_i, p_k)^{-1}$$

A node in the nearest position on average, to all others, can most efficiently obtain information. So, closeness is a measure for the cost of time and efficiency for communicating with other nodes (actors) in the network.

- **Betweenness Centrality**

Freeman [13] yet proposed another concept of centrality which measures the number of times a particular node lies 'between' the various other points in the network (graph). Betweenness centrality is defined more precisely as "the number of shortest paths (between all pairs of nodes) that pass through a given node" [20]. So, the betweenness of node k (p_k) is given by:

$$C_B(p_k) = \sum_{i<j}^{n} \sum_{j}^{n} \frac{g_{ij}(p_k)}{g_{ij}} \quad , i \neq j \neq k$$

Where g_{ij} is the number of geodesic (shortest paths) linking p_i and p_j and $g_{ij}(p_k)$ is the number of geodesic linking p_i and p_j that contains p_k.

Betweenness is an indicator of the potential of an actor (or node) which plays the part of a 'broker' or 'gatekeeper' which can most frequently control information flow (communication) in the network.

3 Analysis and Results

Figure 1 highlights the growth of the network by showing the cumulative number of actors and links among them during network evolution between 1998 and 2009. While naturally the number of actors and links among them is increasing rapidly but the average number of links per actor remains almost steady between 1.14 (in 1998) and 1.52 (in 2009) with very small fluctuations in between.

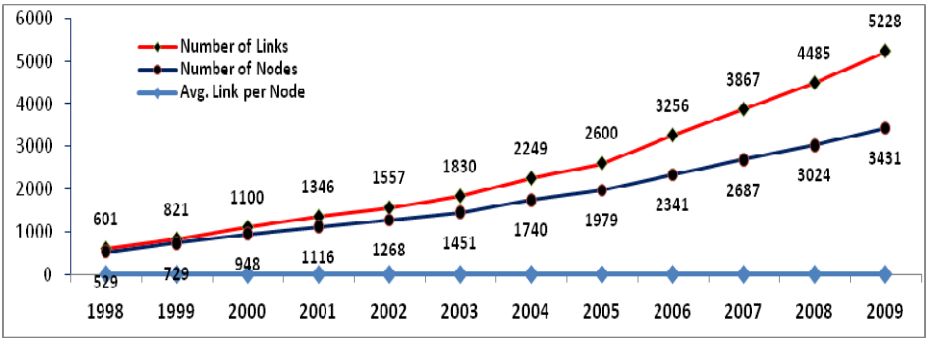


Fig. 1. Number of nodes and links during network evolution between 1998 and 2009

After calculating degree, closeness and betweenness centrality measures of each actor (author) at time t (for instance, 1998) and also numbers of attached actors (authors) to each node in time t+1 (for instance, 1999), using the Spearman correlation test, we find association between each actors' centrality measures (e.g., degree, closeness and betweenness) for each time t and numbers of attached actors to each node in time t+1 (time scale is years between 1998 and 2008). To test the association between network centrality measures (i.e., degree, closeness and betweenness centralities) and the number of attached actors to the existing ones, we use the Spearman correlation test. As Table 1 shows, each actor's centrality measures positively correlated with the numbers of actors attached to him/her except for degree centrality in 1998 and 2000 and closeness in 1998.

The results of correlation test, not only supports the preferential attachment rule that actors with high number of connected nodes (high degree centrality) receive more actors in following year but also indicate that also the actors who are closest to all other nodes in the network (high closeness centrality) and the ones who are more frequently in the path between every other pair of actors (high betweenness centrality) have received more attachments in following year. It may be due to the preference of new actors to attach the old (existing) actors that have higher degree, closeness and betweenness.

Table 1. Spearman correlation coefficients between nodes' centrality measures and their attachment frequency per time frame

Nodes Centrality Measures		Number of new added nodes in										
		1998	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
Degree	p	-.027	**.109***	.051	**.156***	**.100***	**.115***	**.163***	**.149***	**.131***	**.139***	**.138***
Centrality	N	529	729	948	1116	1268	1451	1740	1979	2341	2687	3024
Closeness	p	.003	**.100***	.073	**.175***	**.115***	**.118***	**.121***	**.116***	**.102***	**.108***	**.115***
Centrality	N	529	729	948	1116	1268	1451	1740	1979	2341	2687	3024
Betweenness	p	**.278***	**.279***	**.182***	**.297***	**.235***	**.207***	**.245***	**.227***	**.279***	**.255***	**.249***
Centrality	N	529	729	948	1116	1268	1451	1740	1979	2341	2687	3024

*. Significant at the 0.01 level

Furthermore, the result shows the correlation coefficient between betweenness centrality of actors and their attachment frequency in following year is much higher than degree and closeness centralities for during all the years the network evolves. Thus, as betweenness is an indicator of the potential of an actor which plays the part of a 'broker' or 'gatekeeper' which can most frequently control information flow (communication) in the network; we can infer that new actors (authors) prefer to attach themselves to the ones who have higher control of communication in the network.

3.1 New Nodes' Attachment Behavior

Looking at Table 2, we can find the statistics about total number of actors, links and sum of links per year and also number of new added actors (authors) and also the

number (and percent) of new nodes in each year which has attached to at least one old actor (author).

As the results shows, a few percent of the new actors, which has added to network during its evolution, has attached to old (existing) actors and most of them has just attached to other new actors.

Table 2. Statistics about new actors' attachment frequency to old nodes and other new nodes

Year	# of actors	# of new actors	# of new actors which attach to at least one old actor
1998	529	-	-
1999	729	200	36 (18%)
2000	948	219	27 (12%)
2001	1116	168	49 (29%)
2002	1268	152	33 (22%)
2003	1451	183	38 (21%)
2004	1740	289	49 (17%)
2005	1979	239	46 (19%)
2006	2341	362	84 (23%)
2007	2687	346	77 (22%)
2008	3024	337	99 (29%)
2009	3431	407	111 (27%)

Table 3. Statistics about old actors' attachment frequency to new actors and other old actors

Year	# of actors	# of old actors which attach to			
		at least a actor (any)	at least a new actor	another old actor by a new link	another old connected
1998	529	29 (5%)	23 (4%)	2 (0%)	13 (2%)
1999	729	68 (9%)	45 (6%)	7 (1%)	42 (6%)
2000	948	65 (7%)	45 (5%)	6 (1%)	38 (4%)
2001	1116	64 (6%)	42 (4%)	6 (1%)	45 (4%)
2002	1268	66 (5%)	40 (3%)	12 (1%)	36 (3%)
2003	1451	79 (5%)	61 (4%)	18 (1%)	39 (3%)
2004	1740	79 (5%)	60 (3%)	18 (1%)	40 (2%)
2005	1979	112 (6%)	79 (4%)	27 (1%)	55 (3%)
2006	2341	136 (6%)	102 (4%)	24 (1%)	87 (4%)
2007	2687	147 (5%)	101 (4%)	36 (1%)	95 (4%)
2008	3024	133 (4%)	94 (3%)	39 (1%)	76 (3%)
2009	3431	-	-	-	-

3.2 Old Nodes' Attachment Behavior

Table 3 shows the statistics about total number of actors per year and also number of old (existing) actors in the network which has attached (during network growth) to: (i) at least one actor; (ii) at least one new actor; (iii) another old actors which has not any link with him before; (iv) another old connected actors.

As the results indicate, in general a few of old (existing) actors has attached to others (no matter old or new actors) but among those few, old actors has been attached more to new actors and next having another links with other old actors which were already connected to them rather than having new links to the old actors in the network which were not connected to them before.

4 Conclusion

In order to test and extend the preferential attachment rule during the evolution of complex networks, we explore the evolution of a sample co-authorship network during 10 years to find the best characteristic of the old nodes which affect on the selection of them by new nodes. In general, we find that position of an author in a co-authorship network has impact on the evolution of the collaboration (co-authorship) network and how that author attracts more co-authors to attach to him/her. In particular, the result show that while the association between number of new attached nodes to an existing node and all its main centrality measures (i.e., degree, closeness and betweenness) is almost positive and significant but betweenness centrality correlation coefficient is more stronger and increasing over time rather than degree and closeness centrality.

During network evolution, existing actors which are more frequent during the path that connects any other pair of actors (have higher betweenness centrality), which are having the power of controlling the communication and information flow, receive more new co-authors attached to them rather than the actors who had already more connected (higher degree) and the ones who were more close to all others (higher closeness centrality). We further reveal that authors rarely make collaboration with other authors who have already published works and prefer to have collaboration with new authors or having another collaboration with their co-authors (the ones the already have a link).

"Science as done by humans is a complex system and, as such, is composed of individual agents that take informed decisions" [1] and these decisions show patterns that, unsurprisingly, are so similar to the patterns created in other complex systems (e.g., predator-prey interactions or metabolic networks) [21]. Thus, we can expect that the attachment behavior of nodes we find in this co-authorship may be observed in other similar complex networks too. However, to validate and extend our findings towards proposing a model describing behavior of nodes in evolutionary complex networks, future work will include applying this method to other complex networks (e.g., traders, disease propagation and emergency management). Identifying the attachment behavior of nodes in complex networks help policy and decision makers to focus on the nodes (actors) in order to control the resources distribution, information dissemination, disease propagation and so on due to type of the network.

References

1. Cotta, C., Merelo, J.J.: Where is evolutionary computation going? A temporal analysis of the EC community. Genetic Programming and Evolvable Machines 8(3), 239–253 (2007)
2. Albert, R., Jeong, H., Barabási, A.L.: Internet: Diameter of the world-wide web. Nature 401(6749), 130–131 (1999)
3. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509 (1999)
4. Jeong, H., Néda, Z., Barabási, A.: Measuring preferential attachment in evolving networks. EPL (Europhysics Letters) 61, 567 (2003)
5. Sanyal, S.: Effect of citation patterns on network structure. Arxiv preprint physics/0611139 (2006)
6. Yule, G.U.: A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character 213, 21–87 (1925)
7. Simon, H.A.: On a class of skew distribution functions. Biometrika 42(3-4), 425 (1955)
8. Newman, M.E.J.: Scientific collaboration networks. I. Network construction and fundamental results. Physical review E 64(1), 16131 (2001)
9. Abbasi, A., Altmann, J., Hwang, J.: Evaluating scholars based on their academic collaboration activities: two indices, the RC-index and the CC-index, for quantifying collaboration activities of researchers and scientific communities. Scientometrics 83(1), 1–13 (2010)
10. Owen-Smith, J., et al.: A comparison of US and European university-industry relations in the life sciences. Management Science 48(1), 24–43 (2002)
11. Sonnenwald, D.: Scientific collaboration: a synthesis of challenges and strategies. Annual Review of Information Science and Technology 41, 643–681 (2007)
12. Bavelas, A.: Communication patterns in task-oriented groups. Journal of the Acoustical Society of America 22, 725–730 (1950)
13. Freeman, L.C.: Centrality in social networks conceptual clarification. Social Networks 1(3), 215–239 (1979)
14. Scott, J.: Social network analysis: a handbook. Sage, Thousand Oaks (1991)
15. Abbasi, A., et al.: Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. Journal of Informetrics (2011) doi: 10.1016/j.joi.2011.05.007
16. Batista, P., Campiteli, M., Kinouchi, O.: Is it possible to compare researchers with different scientific interests? Scientometrics 68(1), 179–189 (2006)
17. Bavelas, A.: A mathematical model for group structures. Human organization 7(3), 16–30 (1947)
18. Freeman, L.C.: The gatekeeper, pair-dependency and structural centrality. Quality and Quantity 14(4), 585–592 (1980)
19. Sabidussi, G.: The centrality index of a graph. Psychometrika 31(4), 581–603 (1966)
20. Borgatti, S.: Centrality and AIDS. Connections 18(1), 112–114 (1995)
21. Newman, M.E.J.: The structure and function of complex networks. SIAM review 45(2), 167–256 (2003)

Activity Recognition in Healthcare Monitoring Systems Using Wireless Sensor Networks

Rodica-Elena Doran and Ioana-Iuliana Farkas

Technical University of Cluj-Napoca, Str. C. Daicoviciu Nr. 15
400020 Cluj-Napoca, Romania
Elena.DORAN@bel.utcluj.ro

Abstract. Advanced researches using sensor networks for monitoring home care environments have become a significant concern over the past years. Tracking posture presents many challenges due to the large number of degrees of freedom of the human body. Using the system developed by Brusey [1], in this paper was added an additional hardware component and implemented a new system for posture analysis. An algorithm with multiple threshold methods for detecting and classifying the postures of the subjects in free living live was used. Our new design, which targets a Verdex XM4-bt Gumstix at 400 Hz with a netmicroSD-vx board, is able to collect data and perform posture analysis. Four healthy subjects carry out a set of postures/movements and seven basic positions have been identified by this system: standing, kneeling, sitting, crawling, walking, lying up, and lying on one side.

Keywords: accelerations sensor boards, Gumstix Verdex device, netmicroSD-vx board, posture analysis.

1 Introduction

Advances in wireless sensor networking have opened up new opportunities providing health-care systems for applications such as monitoring home care patients/environments. Because those suffering from diseases such as epilepsy, Alzheimer or old age will increase, most researches are directed towards the use of sensor networks for promoting healthy behavior, and early disease detection. From these detections the treatment can be improved providing necessary medical care for patients. Examples of areas in which future medical systems can benefit the most from wireless sensor networks are in-home assistance, smart nursing homes, and clinical trial. With the increasing number of elderly replying on homecare, it is crucial to monitoring and analysis health systems; in this way, the quality of life for elderly will improve. These systems are not intended to replace health clinics, hospitals; the systems are designed to focus on daily activities as part of the health-care.

This paper brings contributions improving the system developed by Brusey [1] by adding a new hardware component, netmicroSD-vx board. A new version of hardware and software was designed in order to assess data capacity, to evaluate and extend battery life of more than two hours available for Brusey's system, to expand the

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 265–274, 2011.

memory, and record the human postures/movements. This contribution is important because of the small memory available on the Gumstix device (4 MB) which is insufficient to save data from accelerometers for eight hours required for our experiments.

For detecting and classifying the postures/movements, an algorithm with multiple threshold methods has been implemented and used [2]. Our goal is to build a new sensing system that can monitor and assess daily activities in free living live and to use this system for home care patients using body accelerometers. The system is able to perform postures detection at a rate of 10 frames per second. Accelerometers are very useful because they are low cost and very small instruments that provide measurements most widely used in applications which respond to both acceleration due to gravity and acceleration due to body movement. This makes them suitable for measuring postural orientations as well as body movements.

Four healthy subjects carry out a set of postures and seven positions have been identified by the algorithm: standing, kneeling, sitting, crawling, walking, lying up, and lying on one side.

2 Related Work

The work related to the system presented in this paper approaches different monitoring system designs for acquiring the data from sensors for long periods of time without the need to replace the batteries. There are many systems which can monitor and measure human movements. Most of them use a combination of wireless and miniaturized sensors technologies for monitoring a human body.

Chen et al. [3] presented a prototype of a wearable wireless sensor system allowing acquisition of data activities of a person in need of medical care. The system consists of two main components: a wearable sensor unit and a data logger unit. The wearable sensor unit is used for measuring the person's movements. The data logger unit received the samples data via an IEEE 802.15.4 wireless transceiver. The measured data are downloaded to a PC through a cable connection for analysis.

Perolle at al. [4] developed a system composed by a mobile module worn by the user, performing the user localization, automatic fall detection and activity monitoring, and a call centre for data reception from the mobile module for analyzing and saving the information. Their system is wearable and chargeable batteries powered. A battery charge indicator is telling to the user when to change the system's battery.

Milenkovic et al. [5] proposed a WWBAN (Wearable Wireless Body/Personal Area Network) design using for health monitoring. Their prototype consists of several motion sensors that monitor the user's activity and an ECG sensor for monitoring heart activity. They development this prototype system to satisfy the following requirements: small size, low power consumption, and secure communication. For extending the lifetime of nodes, the power dissipation was reduced as much as possible. Therefore, this system allowed designers to use smaller batteries.

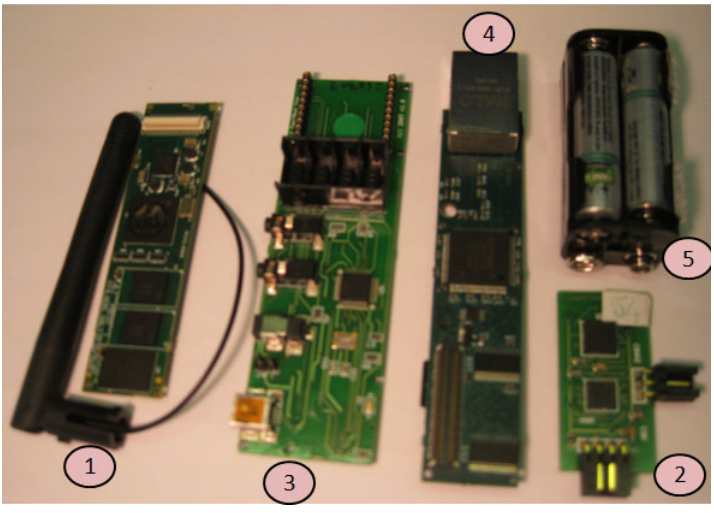


Fig. 1. System components (1) Gumstix Verdex, (2) Sensor board, (3) Expansion board, (4) NetmicroSD-vx board, (5) Batteries

3 Hardware Characterization

Using the system developed at Cogent Computing Applied Research Centre, Coventry University [1], which consists of two Gumstix devices, two expansions boards, 9 sensors (five sensors was used for the upper body and connected to one node, whilst the four sensors are fitted on the lower body and connected to the second node), and two sets of 4 batteries, in this paper we designed a new version of hardware and software system for collecting the data and recorded the human postures. In their paper was used the Verdex XM4-bt Gumstix device as the main processing platform, measuring 80 x 20 x 6.3mm, weighting 8 g and offering 64 MB of RAM, 16 MB of flash memory, and a Bluetooth communication facility. Each sensor board consisted of a microcontroller, a temperature sensor, an accelerometer which is capable of measuring acceleration over a bandwidth of 640 Hz for all axes, and an I^2C bus extender. The expansion board has been configured to interface with temperature sensors and accelerometers.

In our system, we added an additional hardware component, netmicroSD-vx board, and reduced the number of Gumstix platforms, expansions boards, sensors and the sets of batteries. This new system consists only of a Gumstix embedded computer, a custom expansion board, two acceleration sensor boards, a netmicroSD-vx board, and one set of 4 batteries, all shown in figure 1.

The netmicroSD-vx board makes the Verdex into a highly functional, small form factor computer that can be network attached with large amounts of data stored onboard. This daughterboard snaps onto the Verdex motherboard and adds a 10/100 Ethernet port and a MicroSD card slot [6]. This board is connected to the Gumstix via the Molex connectors.

The battery type has a direct impact on the lifetime of the system and the criteria that should be carefully considered before selecting a battery type are: capacity, voltage, peak current, rechargeable, price, material, size, self discharge current, availability, and environmentally friendly. Four rechargeable NiMH 1000mAh AAA batteries are used in our system.

4 System Analysis

Depending on various configurations options in the software, a series of experiments were performed in order to assess data capacity, to evaluate and extend battery life, to expand the memory for recording the human postures. The Bluetooth communication mode was considered "on"/"off", on the Gumstix platform, and the system was tested with one, two or three sensors, taking data at different frequencies 20 Hz and 40 Hz. Six experiments were performed by measuring the acceleration, collecting and receiving data from sensors, storage the data on the memory (USB flash memory or mini SD card) continuously and transmitting or not data continuously through Bluetooth to a monitoring point, in our case a computer (PC). The purpose was caused by the small memory on the Gumstix device, the total memory on it is 16 MB flash memory, and the memory remaining after configuring the device (installing the operation system and the programs necessary) is 4 MB and is insufficient to be able to save data from the accelerometers.

Following the analysis of experiments, it was concluded that the best solution to extend the memory capacity and to have a longer battery lifetime is to switch-off all the functions related to Bluetooth. Because of the small memory remaining on Gumstix system, the netmicroSD-vx board was assumed to be the best solution for expanding the memory. Therefore, the acquired data from sensors have been stored on the mini SD card. Table1 shown the power consumption and battery lifetime with and without Bluetooth, with data saved on mini SD card (The third line shows the power an battery life when data is sent via Bluetooth to PC while at the same time the data is stored continuously on the mini SD card, the fourth line show the power and battery life when data is stored continuously on the mini SD card without sending via Bluetooth to PC, and the last line show the power and battery life when Bluetooth is "off" and data is saved on mini SD card). The assembled system for collecting and receiving data from accelerometers, saving them on mini SD card, is shown in figure 2.

Because the chosen algorithm for classifying the human postures is working better at the sampling frequency of 10 Hz, and our experiments were performed at 20 Hz and 40 Hz, we proposed in [7] a power algorithm for Gumstix device analyzing the energy usage and battery lifetime under the six experiments explained above. We focused on estimating power consumption and battery life using regression analysis. To record the data at 10 Hz, the algorithm predicts the battery life for 4 hours and 50 minutes when data is saved on mini SD card and Bluetooth is "off".

In addition, when the batteries ran out of energy and need to be replaced with new ones, it wasn't required to reconfigure the Gumstix system in order to start a new run,

Table 1. Experimental resutls

Number of sensors	1 Sensor		2 Sensors		3 Sensors	
Frequency[Hz]	**20**	**40**	**20**	**40**	**20**	**40**
Power [W]	1.177	1.280	1.280	1.331	1.331	1.382
Battery life	03:20:22	03:16:32	02:46:23	02:35:30	02:36:47	02:34:23
Power [W]	1.105	1.126	1.152	1.177	1.177	1.280
Battery life	03:19:33	03:19:01	03:19:10	02:57:14	02:57:44	02:52:31
Power [W]	0.716	0.742	0.768	0.819	0.819	0.921
Battery life	05:18:02	04:59:04	04:46:32	04:10:17	03:52:19	03:47:23

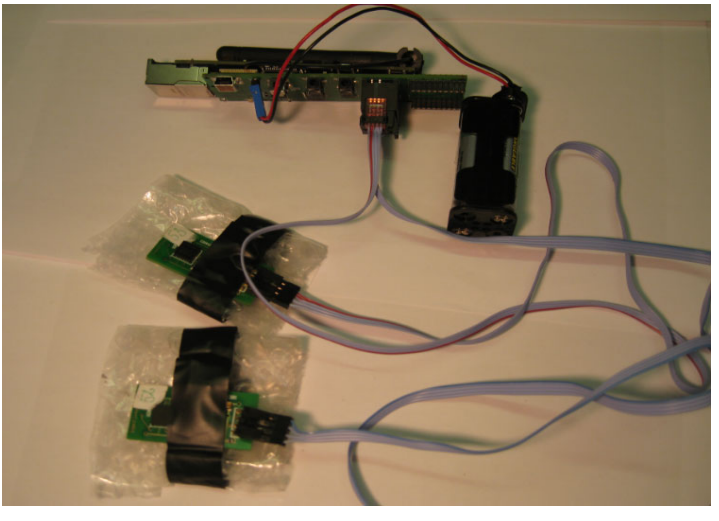


Fig. 2. The assembled system for receiving data from two sensors and save them on mini SD card

the system has been configured to start running again. When the embedded LED on the netmicroSD-vx board is going to flash continuously, the subjects know the batteries are ending and must be replaced with new ones/loaded.

The compact design of this hardware system provides power with reduced physical limitations. The dimension and weight played an important role because was required a minimal size and controllable processing capability. The system is a low cost- for recognizing the postures is enough to use only two tri-axial accelerometers, good accuracy-it can be applied on the clothes, easy to wear-it can be putted in a pocket and in this way it don't bother the person who is wearing for doing the daily movements.

This system has the capability to detect and classify postures with data collected from only two tri-axial accelerometers, is lightweight, has the ability to interconnect with other systems (using Bluetooth "on"/"off" and storing the data on mini SD card/USB flash drive). Based on data collected, decision systems can be ordered. Can

easily added new features like sensors, accelerometers (accelerometers board has incorporated a temperature sensor which can be used to measure the body temperature), and so on.

The system created by Brusey, where the data has been transmitted via Bluetooth, it was limited to the range of the Bluetooth instead our system has no limit because the data are collected and saved on mini SD card continuously without transmitting the data to a monitoring point (without using Bluetooth communication). Therefore, the subjects can do what they want; they are not restrained, so they can go for a walk outside, working around the house, and so on. Also the new system proposed has been configured to last for a long period of time; the subject should change the batteries in 8 hours only twice. If we supposed to use the Brusey's system, batteries must be replaced four times.

5 Data Analysis

Tri-axial accelerometers have been employed to monitor human movements in a variety of circumstances [2]. This classification was structured around a binary decision tree which means that all decision nodes of the tree have exactly two branches making the tree to be easily to read and in this way no logic paths has been inadvertently omitted. The movements were ordered hierarchically from the most general to the most specific in order to be placed in binary decision tree for classification. This was achieved by identifying relationships between categories and structuring these into the tree. All independent movements resulted from the same node were placed at the same hierarchical level. The algorithm for each decision node was based on a fixed threshold method. Changing the algorithm at the upper levels in the tree can affect the algorithms that occur in the lower levels of the tree. Using activity detection classifier it was possible to distinguish first movements. Periods of activity and periods of rest were identified in the signal obtained from the tri-axial accelerometers with this algorithm. Static postures were than classified in upright postures and lying postures.

First we applied a high pass filter (HPF) with the cut off frequency of 0.25Hz (as low as possible) to remove gravitational acceleration component from the data collected with tri-axial accelerometer. After gravitational component was removed from the signal, the high frequency noise spike has to be removed. For that we applied a low pass filter (LPF) with the cut off frequency of 0.25Hz, length n=19 samples, and a non-overlapping moving window of width w=0.1seconds. We removed the gravitational component and high frequency noise spikes in order to calculate signal magnitude area (SMA-magnitude x time). This one includes both effects of magnitude and duration to be able to distinguish between periods of activity and periods of rest. We applied for this purpose the preset threshold of $2.3ms^{-2}$. SMA was defined as the sum of the signal integrals, normalized to the length of the signal and calculated by the areas defined on each of the three axes:

$$SMA = \frac{1}{t} \times \left(\int_t \left| a_1(t) \right| dt + \int_t \left| a_2(t) \right| dt + \int_t \left| a_3(t) \right| dt \right). \quad (1)$$

where a_1, a_2 and a_3 are the acceleration signals from the tri-axial accelerometer.

First step of binary decision tree was made by separating in the first branch periods of activities and periods of rest. Posture of rest was separated in upright and lying positions by applying the threshold method at the vertical axis of the tri-axial accelerometer mounted on the hip. If the measured value of the tri-axial accelerometer, with the condition to be in a resting state, exceeds the preset threshold, the classifier reveals that the subject is engaged in an upright position; otherwise the subject is engaged in a lying posture.

After classifying the periods of rest in periods of upright and lying, it is possible to classify all the other posture taking into account the conditions of activity or rest state and for rest state the condition of upright-lying. First we applied to the original data a median filter with an overlap moving window of 30 samples. The threshold method was applied to different axes.

Because of the high accuracy obtained with this proposed algorithm, we used it for classifying the human postures using our system for collecting the data.

6 Experimental Procedure

Four subjects wore the system for 8 hours each. During this period, the subjects have cooked, ate, went to work, have activities around the house, have arranged things, went shopping and follow their free living live routine. All the subjects were performing the experiments by wearing a tri-axial accelerometer mounted on the right part of the hip and another tri-axial accelerometer mounted on the lower part of the right leg. Each subject carried out seven different activities at the sampling frequency of 10 Hz. The activities were sitting, standing, kneeling, crawling, walking, lying on one side and lying with the face up, all shown in figure 3.

Data collected with the accelerometer mounted on the right part of the hip was sufficient to classify the postures sitting, standing, lying with the face up, crawling and walking. One side and kneeling postures were classified using data collected with the tri-axial accelerometer mounted on the lower part of the right leg.

7 Experimental Results

4 healthy subjects, 2 female and 2 male, 19-32 years old, weight 49-82 kg and height 1.59-1.91m, were wearing the system in a free living live for a period of 8 hours each. The system was set to collect data at 10Hz sampling rate. Each subject noted at the end of the experiment what he did in that period of time.

Subject one note that in eight hours he was working at the computer few minutes, walking in the room, went outside to smoke, back to work at the computer and after that he went for a walk visiting a friend stopping from time to time to a shop to smoke a cigarette and to sit on a bench. He spoke with his friend while they were sitting on a couch and then he went back home. Here he was working in the house, went outside to smoke from time to time, make some order in the room and in some boxes on the floor while he was kneeling and crawling. When he finish ordering, he stay on the

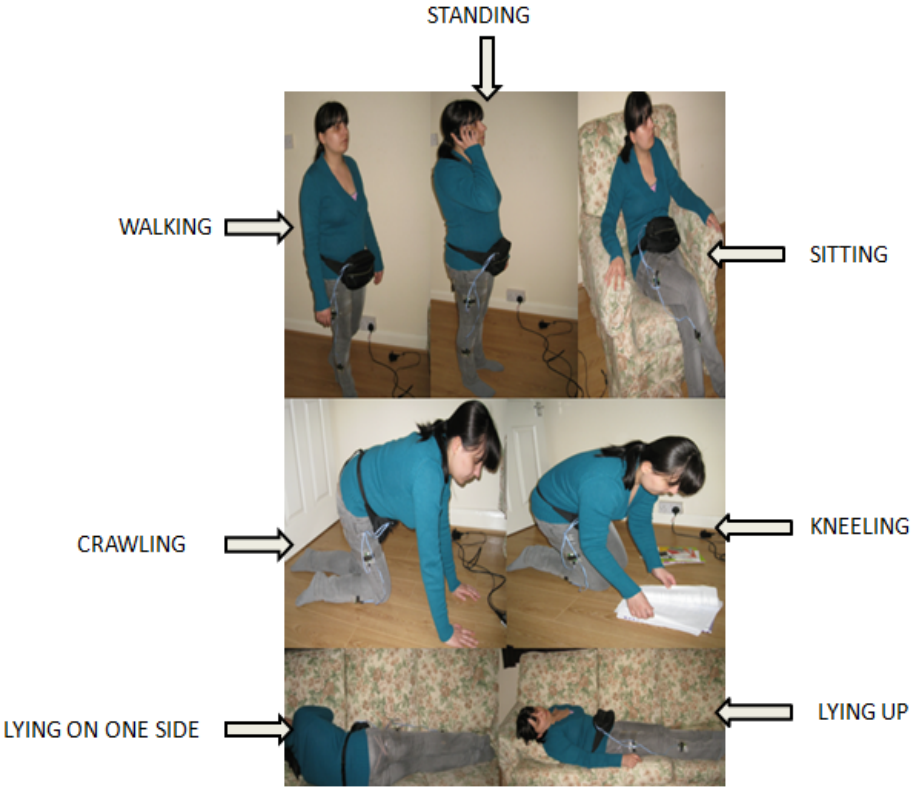


Fig. 3. The seven positions recognized by the algorithm: standing, kneeling, sitting, crawling, walking, lying up, and lying on one side

couch around half hour, walk into apartment, check his blood sugar, and went back in the room on the couch, lie on the bed with the face up for half hour, sit for few moments and after that he went in the kitchen and he began to cook while he was sitting, standing and walking. For this period of time we were able to detect the periods of sitting with an accuracy of 99.39%, standing with 98.18%, walking with 98.09%, lying up with 99.18%, kneeling with 96.44% and crawling with 96.44%. A total accuracy for all the movements was calculated as 97.95%.

Second subject note that his routine was: walking to his job, stopping from time to time to cross the road, sitting at the office, going outside to smoke a cigarette from time to time, back in the office and working at the computer and sometimes he went to the printer to print some papers, going to another floor to heat the lunch, coming back in the office to ate and after that working again at the computer, making order under the table in some boxes while he was kneeling and crawling, taking his staff and going home. On the way home subject stopped some times to make some shopping. When he arrived home he rested on the bed for few minutes, went outside to smoke sitting on a chair, come back in bad lying on one side and after that lying

with the face up. For this experiment we were able to detect and classify periods of sitting with the accuracy of 98.40%, standing with 99.46%, walking with 99.46%, lying with 97.51%, lie down on one side with 99.5%, kneeling with 99.01% and crawling with 99.01%. A total average for accuracy was found 98.94%.

Table 2. A total accuracy for all seven posture detected

Postures	ACC S1 [%]	ACC S2 [%]	ACC S3 [%]	ACC S4 [%]	Average ACC [%]
Sitting	99.39	98.40	99.28	99.74	99.20
Standing	98.18	99.64	99.79	99.28	99.22
Kneeling	96.44	99.01	-	99.14	98.19
Crawling	96.44	99.01	-	100	98.48
Walking	98.09	99.68	99.43	99.51	99.17
Lying up	99.18	97.51	-	99.68	98.79
One Side	-	99.57	-	100	99.78

Subject number three was working at the office for all 8 hours and in this period of time he was working at the computer and some time was walking in the office or standing doing other things. For this experiment we obtained the periods of sitting with the accuracy of 99.28%, standing with 99.79% and walking with 99.43%. A total accuracy for all the postures/movements of the subject was 99.5%.

The last subject was standing on the couch working at the computer, walking in the house and making order in some boxes while he was kneeling and crawling, speaking on the phone while he was standing and walking, back to the work at the computer, lying with the face up, lying on one side on the sofa, walking in the house , back to work at the computer, making a break while he was walking and standing in the room, back to work at the computer, standing for a while and walking, lying with the face up, walking and back to work at the computer. For this experiment we were able to detect periods of sitting with the accuracy of 99.74%, standing with 99.28%, walking with 99.51%, lying with the face up with 99.68%, lying on one side with 100%, kneeling with 99.14% and crawling with 100%. For this experiment we obtained a total accuracy of 99.62%.

As we can observe from Table2, a total accuracy for all the subjects and all the postures was 99.20% for sitting, 99.22% for standing, 99.17% for walking, 98.19% for kneeling, 98.48% for crawling, 98.79% for lying with the face up, 99.78% for lying on one side and a total average of 98.96%. We denoted ACC S1, ACC S2, ACC S3, and ACC S4 the average for subject one, two, three and respectively four, and Average ACC represent the total average of accuracy for all subjects.

8 Conclusions

Our purpose in this paper was to improve the Brusey's system by adding a new hardware component, netmicroSD-vx board. We designed a new version of hardware

and software system for collecting the data, extending battery life, expanding the memory and recording the human postures. These contributions are important because the system has been configured to last for long period of time; the sensing data from accelerometers are stored on mini SD card not in the memory device.

The purpose of building this new sensing system consists of monitoring human movements in free living live for long period of time. The position where the two accelerometers are placed on the body, on the right part of the hip and another on the lower part of the right leg, is very important in the measurements of body movements. The output of body worn accelerometers depends on the position where they are placed, their orientations relative to the subject, the posture of the subject and the activity being performed by the subject.

We managed to obtain better results than in previously work [2] where we was able to classify postures with the binary decision tree of an accuracy of 98.07% for short periods (min 1 minute - max 5minutes for each posture) of movements in a controlled environment comparing with the results obtained in free living live for long periods of time with an accuracy of 98.96%. In this way we were able to prove that this algorithm works better for long periods of time than for short periods of time.

The system is a low cost, has a good accuracy, portable and wearable (can be applied on the clothes). This system is practical for monitoring subjects in free-living environments.

Acknowledgement. Paper published in the Project development studies Ph. D. in advanced technologies “PRODOC” POSDRU/6/1.5/S/5 ID7676.

References

1. Brusey, J., Rednic, R., Gaura, E.I., Kemp, J., Poole, P.: Postural activity monitoring for increasing safety in bomb disposal missions. Measurement Science and Technology 20(7), 075204 (11pp) (2009)
2. Farkas, I.I., Doran, R.E.: Classification of a number of postures/activities based on unsupervised learning algorithms. In: Proceedings of the 1st International Conference on Quality and Innovation in Engineering and Management, pp. 277–281 (March 17-19, 2011)
3. Chen, C., Pomalaza-Raez, C.: Monitoring Human Movements at Home Using Wearable Wireless Sensors. In: Proceedings of the Third International Symposium on Medical Information and Communication Technology, Montreal, Canada (February 2009)
4. Perolle, G., Fraisse, P., Mavros, M., Etxeberria, I.: Automatic Fall Detection and Activity Monitoring for Elderly. In: Proceedings of MEDETEL (2006)
5. Milenkovic, A., Otto, C., Jovanov, E.: Wireless sensor networks for personal health monitoring: Issues and an implementation. Computer Communications 29(13-14), 2521–2533 (2006)
6. Hardware Products for Gumstix Motherboards `http://www.gumstix.com/store/catalog/product_info.php?cPath=31_39&products_id=207`
7. Doran, R.E., Farkas, I.I.: Prediction of power consumption in sensor network device. In: Proceedings of the 1st International Conference on Quality and Innovation in Engineering and Management (March 17-19, 2011)

Defining Events as a Foundation of an Event Notification Middleware for the Cloud Ecosystem

Rolf Sauter[1], Alex Stratz[1], Stella Gatziu Grivas[1], Marc Schaaf[1], and Arne Koschel[2]

[1] University of Applied Sciences Northwestern Switzerland
rolf.sauter@sunrise.net,
{stella.gatziugrivas,marc.schaaf}@fhnw.ch
[2] University of Applied Sciences and Arts Hannover, Germany
akoschel@acm.org

Abstract. This paper discusses the foundations for the implementation of an event processing framework for cloud ecosystems. The idea is to provide standard functionalities which allow users to build cloud solutions capable to respond autonomously to upcoming events, altered environmental conditions or changing needs. Therefore we propose an event-driven-service-oriented architecture. It consists of a set of web services which implement publish and subscribe based communication, complex event processing, rule based activity execution and event routing mechanisms for a cloud ecosystem.

Keywords: Cloud Computing, Event processing SOA, EDA, CEP, ECA, Event Handling, Event Routing.

1 Introduction

Cloud Computing became very popular over the last years, promising agility, flexibility and extensive scalability on demand. The main concept behind cloud computing is that computing power can be obtained similar to consumables like water or electricity. This pay-per-use model and the abolition of infrastructure and facility management make public clouds very attractive for organisations seeking to reduce their capital expenditures and boosting their agility [1] and [2].

To make best use of the advantages of this new technology, mechanisms are needed to manage and to adapt the cloud to changing requirements such as automatic creation, cloning and termination of virtual machines (VM) or service instances corresponding to the variation of the load. Standardisation, contemporary monitoring utilities and comprehensive management functions of the cloud services are key issues. Furthermore, from the application side, it is required to build applications based on scalable architectures to take these new opportunities into account.

Computing clouds are well suited for supporting distributed applications with heavily changing requirements, e.g., variable load. The provided infrastructure is capable of scaling upwards and downwards rapidly to adjust to the changing requirements. One key issue is the proposal of architectural models used for the implementation of scalable architectures. We have analyzed several applications areas in this direction like:

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 275–284, 2011.

- Opinion mining where massive amounts of content is gathered and processed from various heterogeneous, distributed sources, such as online discussion forums, blogs, news sites or social network sites to identify citizen's opinions and to run simulations [8].
- Disaster management where the input data has to be gathered from various distributed sources (e.g. data from scientific sensors) and is processed with the goal to (1) to identify critical situations and (2) react on disaster situations by e.g. efficiently locate persons and equipment.
- Online business development where the click steams of website visitors are processed as events to identify the interests of the visitor, we see as another application area.

A widely accepted interaction and control mechanism for loosely coupled and heterogeneous distributed information systems is the communication through events and the use of event processing technologies. Originating from early discrete event simulation systems and computer networks, the idea of using events as a substantial component of system interaction has been adapted for higher level application and business process architecture paradigms during the last two decades. Event Driven Architectures (EDA) [12], publish and subscribe-style middleware and Complex Event Processing (CEP) [11] are promising concepts in this area.

That the paradigms of event generation, observation and notification could also provide essential benefits for the construction of large 'internet-scale' applications was already assessed in the 90ies in the area of Active Database Management Systems (ADBMS) [6] or in [3] and [5] long before the current cloud computing phenomenon. These early works, however, set their focus mainly on the task of event distribution and notification over a scalable, publish-subscribe-style middleware.

Towards the development of scalable, reusable cloud applications based on event processing, we decouple the event observation and its associated active/triggering functionalities from the individual applications, by moving them into a standardized, shared component, the so-called *Activity Service* as part of the OM4SPACE (Open Mediation4 SOA and P2P Based Active Cloud Components) Project [17] and [18]. Cloud applications get the ability to communicate with other applications and resources in the cloud through the means of a standardized, publish-subscribe-based event notification mechanism. In OM4SPACE we merge the complementary concepts of SOA and EDA to provide an *open event notification middleware* for the decoupled communication between applications of all layers in the cloud. Our approach for the open event notification middleware is described in section 2.

Furthermore, similar to the approach of ADBMS and along the lines of Complex Event Processing (CEP), the applications should also be able to externalize often used 'active functionalities', to respond to and to act upon the occurrence of patterns of events in the cloud.

While the concepts of event processing are already well known in areas like ADBMS, or CEP; a feasible platform for the event processing within the highly volatile environment of the cloud is required. We identified that a foundation on the event semantics that allows the applications to benefit from the agility of the cloud is currently missing.

In this paper we also describe the *notion of events in the cloud*. We evaluate the key characteristics of events in the context of cloud computing and derive a proposal for an extensible event format. The overall contribution of this paper is thus the definition of the semantics of event processing in cloud environments to lay the foundation for the definition of the clearly defined semantics of the Activity Service. To round up the notion of events we briefly describe the architecture and the design of the Activity Service based on proven principles, patterns, and technologies from service oriented architectures (SOA). Details about the architecture and the prototypical implementation of the Activity Service have been discussed in [18].

The rest of the paper is structured as follows: Chapter 2 motivates the use of EDA in combination with SOA as basis for the Activity Service. Chapter 3 discusses an event model for cloud environments to define the semantic on which the Activity Service is based. It is followed by Chapter 4, which brings together the results by discussing the architecture of the Activity Service. Eventually we conclude with a summary and an outlook on the next steps for the whole project.

2 Open Event Notification Middleware

Different architectures like EDA and SOA were developed in the past to build distributed software solutions. These architectures usually focus only on specific requirements or application areas which limits their benefit. They are sometimes even seen as disparate and incompatible software design approaches [13]. Our concept is intended to merge the complementary concepts of SOA and EDA to provide an open event notification middleware for the decoupled communication between applications of all layers in the cloud [4] and [12].

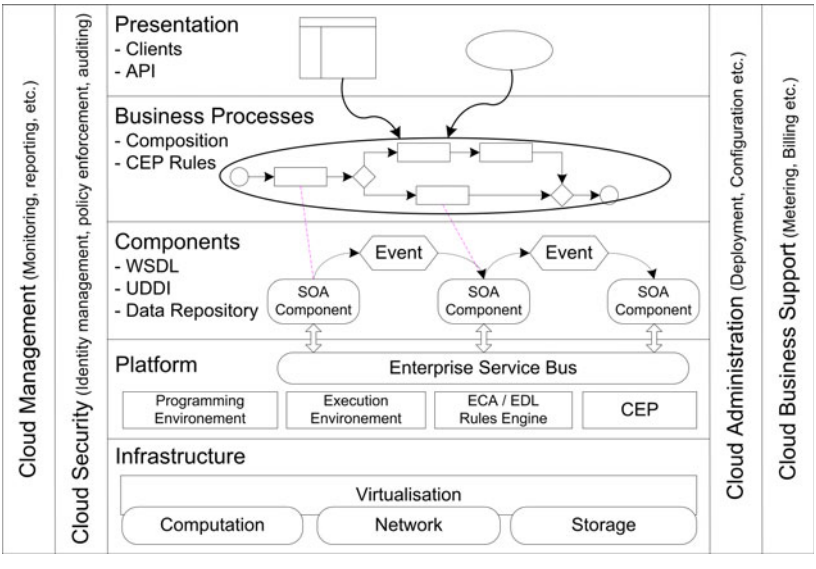


Fig. 1. Event-driven-SOA Cloud Stack (as an extension from [1])

Cloud Computing, SOA and EDA are all service oriented technologies, offering their various services on the network. SOA services can be easily combined to change or to build new applications. But they are limited to sequential and linear processes, because the behaviour is statically defined at design time. EDA conquers this limitation. It allows to build applications to handle unpredictable processes. The process flow of an EDA application is determined at runtime, triggered by events. Thus they are able to respond on casual events occurring in random order. While SOA is related to linear work flows, EDA fits for asynchronous long-running processes and to handle deterministic, unpredictable workflows.

Both concepts, SOA and EDA reduce the complexity of applications by dividing them into modular, reusable, encapsulated and lose coupled components. Both approaches are message-based, sharing common semantic only, and they are well suited to implement distributed applications across multiple platforms and enterprise boundaries. Hence they are the right architectures for cloud applications [10] and [19]. Moreover cloud computing profits from SOA's broad set of mature standards.

Using EDA as an implementation concept for the SOA the advantages of both worlds can be combined. The resulting Event-Driven-Service-Oriented-Architecture (EDSOA) will enrich cloud ecosystems with advanced cloud management capabilities, by adapting the cloud to changing needs delivering what's ordered according to predefined SLA [12].

The EDSOA described above leads to a new contemporary cloud stack consisting of layers for infrastructure, platform, components, business processes and presentation, which would look as outlined in Figure 1. The proposed Cloud Stack is a modular architecture using asynchronous messaging to interconnect its components. The components are fully decoupled, reusable and encapsulated building blocks. The stack provides the components the capabilities to handle unpredictable workflows and to respond in real-time to random events or event-streams. Communication becomes self-organized; single components do not need to know anymore if and how information they send is processed. Cloud applications become more reliable and flexible because their workflows are determined during runtime.

3 The Notion of Events in the Cloud Ecosystem

An event is a notable thing that happens inside or outside your business, e.g. an incoming order, arising errors, thresholds reached or just the change of a fact, value or state. In a generalized view all control and data flows in an EDSOA must be understood as events [7] and [12]. Events in an IT system are represented by messages, which are also called events. The message should contain all relevant information about the real event, like event-id, event origin, occurrence time, situation description etc. Events are usually posted without knowing if and how the events are handled by other components. Each party decides by itself to which type of events it will subscribe or to which it will respond. Thus EDA components act autonomous and are fully decoupled. This makes such architectures quite fault resistant. However, it is important that the format and the semantic of the events is understood by all components. The definition of format and semantic can therefore be seen as a formal contract.

3.1 Events in the Cloud

In our understanding 'the cloud' as an abstract term comprises the whole digital world of interconnected networks. In this sense, anything is either represented by information inside the cloud or exists only outside the cloud and is therefore not known. It embraces the infrastructures of different cloud providers as well as interconnected enterprise system environments. The cloud environment is heterogeneous and combines arbitrary system architectures, platforms and technologies.

With this view of the cloud, events become a very general meaning. An event must be understood as any kind of information sent as a notification from one component to another. Examples are the value of a temperature sensor, an error trap sent by a server or a service invocation sent to a web service endpoint. In spite of this general vision of events, the traditional publish-subscribe-paradigm remains. Events are still posted to a middleware without knowing if and how they are processed.

Nevertheless the cloud can be structured into different layers, e.g. infrastructure, platform or software layer and, more important, the cloud is divided into different administrative or user domains, also known as zones. Such domains are spanned from the user's PC throughout the enterprise IT environment and across the internet. The domains possess boundaries which delimit them securely from each other. This has some major impact on the event flow. While events shall be exchanged without any limitation within a single domain, event flows across domains should be limited by well-defined rules.

Moreover, the exchange of the contract that describes the format and the semantic of the events between different domains is another challenge. Without this information, events cannot be processed or even understood.

3.2 Characteristics of 'Events in the Cloud'

An event is a happening of interest at a certain point in time in a certain location. Due to the heterogeneous nature of the cloud ecosystem, many different types of events will occur. To allow an effective and efficient subscription mechanism, the various types of events must be well characterized. Figure 2 below outlines some major characteristics of event types. This model can be extended with additional characteristic elements as needed like time or location aspects.

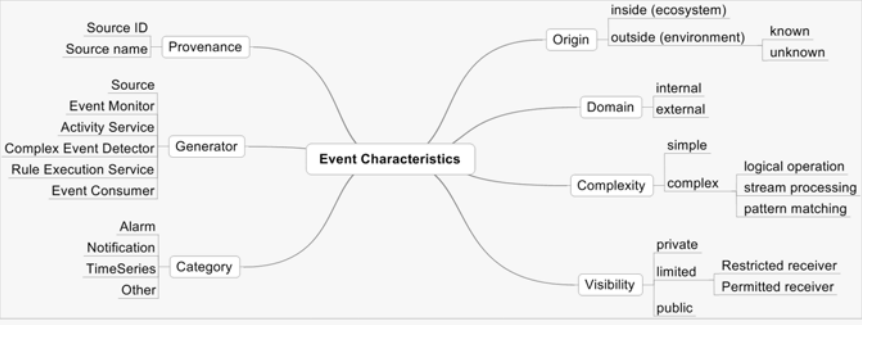


Fig. 2. Characteristics of event types

Events can be related to a phenomenon inside (ecosystem) or outside the cloud (environment). This means they reflect the system itself or they describe something that happens outside the system, e.g., business events. They can origin within the administrative or user domain (internal), or they come from another domain (external). The origin of an event is important for the purpose of message routing and security, e.g. to clarify the question if an event is trustworthy or to apply filters and access control lists. They can occur as single events, representing a status change like exceeding a threshold or the start or the end of a process. They can transport error messages and warnings which might be related to each other, crop up in parallel or in dedicated order (event clusters/clouds). Or they transmit the continuous values of time series (event streams).An event can be generated by a single phenomenon (simple events) or they are the result of a logical operation or a more complex analysis (event patterns, compounds or complex events).

Another characteristic to describe the event is its source. For example, was the event generated by the source it-self or by an event monitor which retrieves the information from a passive component? For complex events it would also be interesting to know the original provenance, the source which has triggered the event process. The event characteristic and source information can be understood as a fingerprint of the event which can be used e.g. in the subscription process to select the events of interest.

Take as an example a monitoring system. It might be interested to get all notifications related to service invocations and service completion from components which belong to a certain business application. The events of interest can be selected based on the SourceID of the application components and the Category of the event type. Knowing the event category and the URI of the interesting components, the monitoring system can subscribe for the required event types. Well-structured URIs and wildcards make it easy to select whole groups of components with a single request:

```
Subscribe (SourceIDstarts_with 'as.example.com/application'
and Category equal_to 'notification')
```

3.3 Cloud Events: Format and Semantics

In practice the format chosen to represent events plays a core role. It should be open and flexible to capture all possible events, going along with low overhead and readability for all participants. E.g., Common Base Event (CBE) promises to provide all of them. CBE is a unified, XML based consistent event format developed by IBM [14]. Since CBE was intentionally designed for system management it has a specific and complex data structure leading to high message overhead. Therefore we defined an own simple and adaptive lightweight event format for the cloud ecosystem, which is inspired by CBE.

In our solution the event type and the event message type are separated. Both are structured sets of data in a machine readable format. The event message type describes the event message format and belongs to a dedicated event type. The event message itself is divided into a header and the situation description. The header contains some metadata about the event, like source and occurrence time. The situation description contains a set of structured data elements, which allow an adaption of the format in a flexible way to specific types of situations or events.

The heterogeneity of the cloud ecosystem requires platform independent data formats and transport protocols. It is therefore an obvious choice to use XML and SOAP to exchange events. Both standards are easy to use and widespread in the cloud. They provide platform independence and allow seamless processing in multiple programming environments. Furthermore, XML can be adapted to model arbitrary event format and is easy to read even for human readers. While we propose to use XML and SOAP for a first implementation, the concept of the Activity Service is not limited to these standards.

A simple example of an event type description in XML format is shown below:

```
<EventType>
<etypeID>0000000000000033</etypeID>
<etName>MyEventType</etName>
  <characteristic>
    <category>alarm</category>
    <visibility>private</visibility>
    <domain>internal</domain>
    <origin>ecosystem</origin>
  </characteristic>
  <source>
    <sourceID>example.ch\insurance_comparison</sourceID>
    <monitorID>example.ch\insurance_comparison</monitorID>
  </source>
<MessageType>
  <mtypeID>0000000000000033:0001</mtypeID>
  <mtypeName>performance_threshold_reached</mtypeName>
  <mtypeFomat>es.example.ch/ptr0001/XMLScheme</mtypeFormat>
</MessageType>
</EventType>
```

The format defines how events are represented within the system, while the semantic defines how the content of the event messages must be understood. It determines the meaning of an event. Beside the transmission of known events, the determination or creation of events is one of the fundamental functions. First of all, the system must be able to detect the phenomenon of interest e.g. the invocation of a service. Secondly, the phenomenon must be translated into the internal event format. Events and the related conditions and constraints can be formally expressed in an event description language [11].

The event type description can be seen as the contract between event generator and event consumer. It contains two parts, the data format definition of the related event messages and as well the format and characteristics of the event type itself.

4 An Architecture for Event Processing

Current cloud solutions are not designed to support the special requirements of event processing. A dedicated architecture is required to fulfil these demands. The result is what we call *'an Activity Service for the Cloud'*. On one hand the solution must be able to detect, distribute and process events, on the other hand it should be possible to trigger activities in the cloud using common methods and well known standards.Our solution is based on the EDSOA approach outlined in Section 2. The design is fully open and extensible and allows to adapt the solution to new needs such as new types

of events. There active building blocks use notifications only for their communication. Reusability allows cloning of the components to use their functionality in other user domains, or to scale up if more performance is needed.

As illustrated in Figure 3 we consider associated building blocks implemented as loosely coupled web services. The core building block is the Event Service, which provides the functionality of message queuing and exchange in publish and subscribe manner. This function is similar to the one of an event aware middleware in conventional EDA environments, e.g. MOM or ESB. The Event Service contains as well a registry for event types, related message formats and subscribers. New event types and related message formats can be registered by any event source or event monitor. Event consumers may discover the event registry. If they find an event type of interest they can subscribe to the event type or to individual message types. The Event Service will further on forward all event messages of these types to the subscriber.

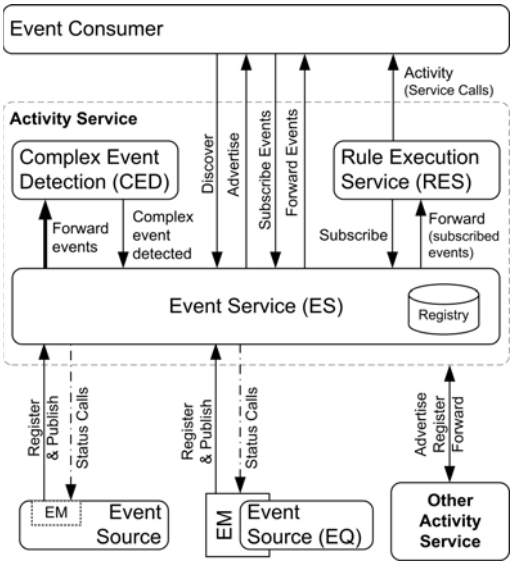


Fig. 3. Architecture of the Activity Service

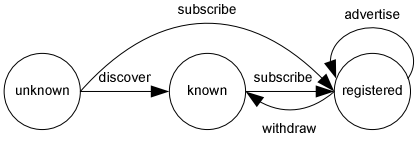


Fig. 4. Event Service – Subscriber State Diagrams

If an event type is changed by its source, the Event Service sends an advertisement to all subscribers before it continues forwarding the events in the revised format. A subscription remains until the event consumer withdraws it. The different registration states of events and subscribers are shown in Figure 4.Event monitors are responsible

to observe the event sources and to detect events of interest. The event monitor must be able to recognise and to transform an event into the internal format without any relevant impact on the source. The event monitor must be aware of the internal event format and the semantic to generate events according to event determination rules like '*send the response times of each service invocation'* or '*send all response times of service invocation which are above X ms.'*

All communication between the components is done by sending XML notifications using SOAP over HTTP or HTTPS. This provides two major advantages. First it enables seamless and secure communication across various networks and firewalls. Second it allows exchanging structured information between different platforms, in a flexible, machine readable format.

Messages sent from and to the Activity Service are exchanged in publish/subscribe style, which means that only the operation type *notification* is used. Within the Activity Service other operation types are valid also, but *notification* is the preferred one. The registration in a UDDI directory allows to discover the web services offered by the Activity Service. URIs are used to ensure the identification of the components across different domains.

5 Conclusion

In this paper, we outlined our concept of an event aware cloud stack by mergeing the two complementary concepts of SOA and EDA. The new stack provides an open event notification middleware for the decoupled communication between components of all layers in the cloud. The extension with CEP and the possibility to externalize common active functionality into a scalable, central component creates new means for developing scalable, reusable and self-adaptable cloud services.

The defined concept is used as basis for the first OM4SPACE prototype currently developed and for the usage in the several application domains. This work provides the foundation for our further research in the area of common semantics for cloud event processing like cloud rule execution semantics. Moreover it allows later on for the extension of the concept into multiple Activity Services in different cloud environments, e.g. communication across the border of a single cloud. Our work is a first step towards a cloud ecosystem which supports distributed autonomous applications driven by events.

References

[1] Armbrust, M., et al.: Above the clouds: a Berkeley view of cloud computing. Berkeley: University of California, EECS Department. Technical Report UCB/EECS-2009-28

[2] Buyya, R., et al.: Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems 25(6)

[3] Carzaniga, A., Rosenblum, D.S., Wolf, A.L.: Design and Evaluation of a Wide-Area Event Notification Service. ACM Transactions on Computer Systems 19(3) (2001)

[4] Covington, R.: Event-Driven Architecture (EDA) and SOA: Complimentary architectures for the enterprise. OMG Special Interest Group: SOA, `http://soa.omg.org/Uploaded%20Docs/SIG%20Meetings/Dec.%2005/05-12-04.pdf` (accessed 12.12.2010)

[5] Cugola, G., Di Nitto, E., Fuggetta, A.: The JEDI Event-Based Infrastructure and Its Application to the Development of the OPSS WFMS. IEEE Transactions on Software Engineering 27(9), 827–850 (2001)
[6] Dittrich, K.R., Gatziu, S., Geppert, A.: The Active Database Management System Manifesto: A Rulebase of ADBMS Features. In: Sellis, T.K. (ed.) RIDS 1995. LNCS, vol. 985, Springer, Heidelberg (1995)
[7] Dunkel, J., Eberhart, A., Fischer, S., Kleiner, C., Koschel, A.: Systemarchitekturen für Verteilte Anwendungen. Carl Hanser Press, Munich
[8] Gatziu Grivas, S., Schaaf, M., Kaschesky, M., Bouchard, G.: Cloud-based Event-processing Architecture for Opinion Mining. In: IEEE International Workshop on Management in Cloud Computing (MCC 2011), Washington (July 2011)
[9] Goyal, P., Mikkilineni, R.: Policy-based Event-driven Services-oriented Architecture for Cloud Services Operation & Management. In: CLOUD-II IEEE 2009 International Conference on Cloud Computing, Bangalore, India (September 21-25, 2009)
[10] Laliwala, Z., Chaudhary, S.: Event-driven service-oriented architecture. In: 5th International Conference on Service Systems and Service Management (ICSSSM 2008), Melbourne, Australia (June 30 - July 02, 2008)
[11] Luckham, D.: The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems. Pearson Education, London
[12] Luckham, D.: SOA, EDA, BPM and CEP are all complementary. Part I, http://complexevents.com/wp-content/uploads/2007/05/SOA_EDA_Part_1.pdf (accessed 30.10.2010)
[13] Michelson, B.M.: Event-driven architecture overview: Event-driven SOA is just a part of the EDA story. Patricia Seybold Group, http://dx.doi.org/10.1571/bda2-2-06cc (accessed 26.11.2010)
[14] Ogle, D., et al.: Canonical Situation Data Format: The Common Base Event V 1.0.1. IBM, New York, http://www.eclipse.org/tptp/platform/docments/resources/cbe101spec/CommonBaseEvent_SituationData_V1.0.1.pdf (accessed 05.02.2011)
[15] Raines, G.: Cloud computing and SOA. MITRE, Bedford, http://www.mitre.org/work/tech_papers/tech_papers_09/09_0743/09_0743.pdf (accessed 28.05.2010)
[16] Rosenblum, D.S., Wolf, A.L.: A Design Framework for Internet-Scale Event Observation and Notification. In: Proc. of the 6th European Conference Held Jointly with the 5th ACM SIGSOFT International Symposium on Foundations of Software Engineering, Zurich, Switzerland, pp. 344–360 (September 1997)
[17] Schaaf, M., Koschel, A., Gratziu Grivas, S., Astrova, I.: An Active DBMS Style Activity Service for Cloud Environments. In: Proc. of the International Conference on Cloud Computing, GRIDs, and Virtualization (IARIA 2010), Lisbon, Portugal (November 21-26, 2010)
[18] Schaaf, M., Koschel, A., Gatziu Grivas, S.: Event Processing in the Cloud Environment with well-defined Semantics. In: Proc. of the first International Conference on Cloud Computing and Services Science, CLOSER 2011, Noordwijkerhout, Netherlands (May 2011)
[19] Zhang, L.J., Zhou, Q.: CCOA: Cloud computing open architecture. In: IEEE 7th International Conference on Web Services (ICWS 2009), Los Angeles (July 6-10, 2009)

Effective Content-Based Music Retrieval with Pattern-Based Relevance Feedback

Ja-Hwung Su[1], Tzu-Shiang Hung[1], Chun-Jen Lee[2], Chung-Li Lu[2], Wei-Lun Chang[2], and Vincent S. Tseng[1,*]

[1] Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan, R.O.C.
[2] Telecommunication Laboratories, Chunghwa Telecom, Taoyuan, Taiwan
tsengsm@mail.ncku.edu.tw, cjlee@cht.com.tw

Abstract. To retrieve the preferred music piece from a music database, content-based music retrieval has been studied for several years. However, it is not easy to retrieve the desired music pieces within only one query process. It motivates us to propose a novel query refinement technique called *PBRF* (*Pattern-based Relevance Feedback*) to predict the user's preference on music via a series of feedbacks, which combines three kinds of query refinement strategies, namely QPM (Query Point Movement), QR (Query Reweighting) and QEX (Query Expansion). In this work, each music piece is transformed into a pattern string, and the related discriminability and representability of each pattern can be calculated then. According to the information of discriminability and representability calculated, the user's preference on music can be retrieved by matching patterns of music pieces in the music database with those of a query music piece. In addition, with considering the local-optimal problem, extensive and intensive search methods based on user's feedbacks are proposed to approximate the successful search. Through the integration of QPM, QR, QEX and switch-based search strategies, the user's intention can be captured more effectively. The experimental results reveal that our proposed approach performs better than existing methods in terms of effectiveness.

Keywords: Content-based music retrieval, pattern-based relevance feedback, query point movement, query re-weighting, query expansion.

1 Introduction

With the popularity of Internet, people can easily retrieve music pieces from the Web. People spend much time to listen to music ubiquitously by any mobile devices such as iPod and thereby the need of music acquisition occurs frequently. To handle such need, text-based retrieval and content-based retrieval are two main paradigms in the field of music retrieval. In terms of text-based music retrieval, the primary difficulty comes from that, the gap between music content and query terms is not easy to bridge semantically. For instance, it is difficult to search the related music pieces by using conceptual term "love" because several genres may cover this concept. This difficulty actually limits the related real applications. Accordingly, Content-Based Music

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 285–295, 2011.

Retrieval called CBMR has been widely adopted as a solution to deal with such above problem based on content comparisons. Without any query terms, the user can retrieve the interested music pieces from the music database by submitting a music example. Although some previous solutions have been shown to be promising on CBMR, the user's preference on music is not easy to predict by only one query process. The same problem also exists in the image retrieval field [10].

To aim at this problem, Relevance Feedback also called RF has been studied for many years to connect music content and user interest. Typically, a query session for relevance feedback indicates that, the iterative query process does not stop until the search results can satisfy the user. Namely, a query session contains a series of feedbacks. At each feedback, the user can define preferred and non-preferred music pieces as positive and negative ones, and then re-query the database using the positive and negative information defined. Through a series of interactive queries, the user's interest can be predicted more and more accurately. On the basis of this scenario, we propose a novel Pattern-Based Relevance Feedback for content-based music retrieval called PBRF that integrates Query-Point-Movement (QPM), Query-Reweighting (QR) and Query-Expansion (QEX) ideas to capture the user's preference on music. The empirical evaluations reveal that our proposed approach is more effective than other existing methods in terms of music relevance feedback. The rest of this paper is organized as follows. A review of related work is briefly described in Section 2. In Section 3, we demonstrate our proposed method for pattern-based relevance feedback in great detail. Experimental evaluations of the proposed method are presented in Section 4. Finally, conclusions and future work are stated in Section 5.

2 Previous Work

Over the past few decades, a considerable number of past studies have been made on content-based music retrieval due to the need of music acquisition. Basically, content-based music retrieval systems can be categorized into two types, namely *precise match* and *relevant rank*. Precise match is based on the assumption that, the content of the search result has to be the same as that of the query example, while relevant rank indicates that the search results are similar to the query example on musical content. For precise match, one type is QBH (Query By Humming) which queries the music database by humming the tune of a song [5], and another type is music identification which exploits audio fingerprints to identify the relation between the target and query [3]. For relevant rank [2], the main aspect is to generate a playlist containing a set of relevant music pieces by measuring the content similarity based on the query, e.g., tempo, pitch, rhythmic, timbre, etc.

Although the approaches mentioned above have shown to be effective on content-based music retrieval [10], they still encounter a difficulty that the user's intention is not easy to predict by only one query process. It leads to that, numerous former studies focus their attention on relevance feedback. In the field of relevance feedback, three state-of-the-art techniques, namely Query-Point-Movement (QPM) [7], Query-Reweighting (QR) [8] and Query-Expansion (QEX) [6], were proposed to attack the problems occurring in content-based image retrieval. In terms of QR, the major notion is that the system assigns the higher degree w_i to the i^{th} feature f_i after feeding back

several good examples if f_i is important enough, where w_i is the inversed standard deviation of f_i. In terms of QPM, the major notion is that QPM aggregates the positive examples as a new query point at each feedback. After several forceful changes of locations and contours, query point can be very close to a convex region of user's interest. In terms of QEX, the major notion is that it groups the similar positive examples into several clusters, and selects representative points from these clusters to construct the multipoint query. In spite of the success of visual relevance feedback above, few attempts have been made on content-based music retrieval with relevance feedback up to the present. Hoashi et al. [4] proposed TreeQ to refine a new query by considering positive and negative examples. Shan et al. [9] transformed a music piece into a set of patterns and predicted the user's intention using the association rules discovered from positive and negative patterns. Although the above methods have improved the performance of content-based music retrieval using relevance feedback, the relation of music content and user interest is still not easy to clarify.

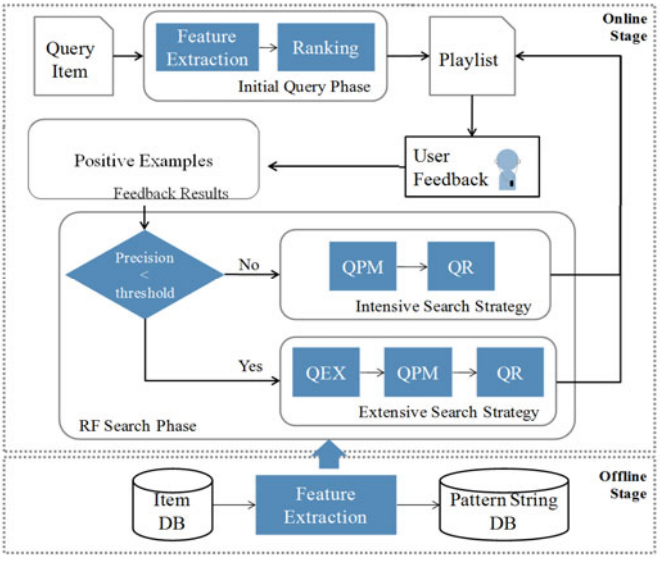


Fig. 1. Framework of the proposed approach

3 Proposed Approach

3.1 Overview of Pattern-Based Relevance Feedback

Basically, the bottleneck of relevance feedback on content-based music retrieval lies in two aspects: 1) the lack of representability and discriminability for musical features, and 2) the problem of local optimal search. From the viewpoint of representability and discriminability, a feature is discriminative if it occurs in few specific music pieces, while it is representative if occurring frequently in a music

piece. From the viewpoint of local optimal search, the search result may converge towards the local optimal area in the search space. As a result, the goal of this paper is two-fold. One is to discover the representative and discriminative musical feature patterns to elicit the user interest hidden in music pieces. The other is to propose a switch-based search strategy that switches between extensive search and intensive search by search results at each feedback, so as to prevent the local optimal problem. To reach such goals, our proposed framework shown in Figure 1 can be divided into two stages, namely offline processing stage and online retrieval stage. In offline processing stage, each music piece in the database is transformed into a set of sequential patterns from low-level musical features to a symbol string. In online retrieval stage, it can be further divided into two parts, namely initial query phase and relevance feedback phase. The main operation in initial query phase includes three steps: extraction of musical features, search of relevant music pieces and generation of a ranking playlist. By the generated playlist, the user can pick up the interested music pieces as positive examples, and then re-submit these positive examples. The RF phase starts with that the positive examples are re-submitted. With the consideration of proposed switch threshold, the system integrates QPM, QR and QEX to search the relevant music pieces using weighted patterns. Overall the RF phase does not stop until the results are satisfactory to the user. The great detail of each phase is described in the succeeding sections.

3.2 Offline Processing Stage

The major operation of this stage is to transform each music piece in the database into a pattern string. First, each music piece is divided into a number of frames and the low-level features of each frame are then extracted, including Modified Discrete Cosine Transform (MDCT), Zero Crossing Rate, Spectral Centroid, Spectral Rolloff, Spectral Flux, Spectral Flatness Measure, Spectral Crest Factor, Linear Prediction Cepstral Coefficients, Chroma and Mel-scale Frequency Cepstral Coefficients (MFCC). Next, for each feature, we calculate four statistical values including maximum, minimum, standard deviation and average of feature. Note that, the statistical values calculated in this step are also classified into two types, which are defined as global attribute and local attribute. A global attribute indicates the statistical value calculated by referring to the whole music piece, while a local attribute refers to a specified subsequence of a music piece. Finally, the statistical values are transformed into symbols according to the equal-sized value range regarding each feature. Let us take a simple example to illustrate this operation. Assume that, a music piece defined as itm_1 is divided into 10 frames where a second is viewed as a local attribute containing 5 frames. Given the low-feature set is {2, 3, 5, 3, 9, 9, 8, 6, 6, 3}. For the global attribute and two local attributes (two seconds), the value sets related to {minimum, maximum, average, standard deviation} are {2, 9, 5.4, 2.5}, {2, 9, 4.4, 1.7} and {3, 9, 6.4, 2.1}, respectively. According to different value ranges of different features, the values are transformed into symbols then. Finally, the pattern string of itm_1 is represented as {A, L, N, R, C, M, P, T, C, M, Q, U}.

3.3 Online Retrieval Stage

This stage primarily concerns with how to predict the user's preference by iterative feedbacks. From the feedbacks, the user's preferable patterns with high representabilities and discriminabilities can be discovered to enhance the quality of music retrieval. On the whole, it can be decomposed into two phases, namely initial query phase and relevance feedback phase.

3.3.1 Initial Query Phase

Once a query is submitted to this system, this phase is triggered. This phase is different from relevance feedback phase since there is no information in addition to the query piece to support the search mechanism in this phase. Assume that there are k music pieces in database D=$\{itm_1, itm_2, \ldots, itm_k\}$. A music piece is also called music item in the followings. For a music item itm_i in D, the degree to query itm_Q can be defined as

$$initial\_degree_{itm_i} = \sum_{ms \in (MS_{itm_Q} \cap MS_{itm_i})} TF_{ms}^{itm_Q} \times TF_{ms}^{itm_i} \times IDF_{ms} \quad (1)$$

where MS_{itm_Q} denotes the set of patterns in itm_Q, MS_{itm_i} denotes the set of patterns in itm_i, $TF_{ms}^{itm_Q}$ denotes the occurrence of pattern ms in itm_Q, $TF_{ms}^{itm_i}$ denotes the occurrence of pattern ms in itm_i, and IDF_{ms} defined as Equation (2) is the number of music items containing pattern ms.

$$IDF_{ms} = \log \frac{|D|}{\sum_{1 \le j \le |D|} df_{ms}^{itm_j}}. \quad (2)$$

where $itm_j \in$ D and

$$df_{ms}^{itm_j} = \begin{cases} 1, if\ ms\ occurs\ in\ itm_j \\ 0, otherwise \end{cases}.$$

Table 1. Example of pattern strings related to music items in database D

	Pattern String					
	Feature 1			Feature 2		
	Global	Local	Local	Global	Local	Local
itm_1	A	C	C	F	H	K
itm_2	B	D	E	F	I	K
itm_3	A	C	E	F	K	I
itm_4	B	D	C	G	I	K
itm_5	A	E	D	F	I	K
itm_6	A	C	D	F	H	I
itm_7	B	D	E	F	K	I
itm_8	A	E	C	F	I	K
itm_9	B	E	E	G	H	K
itm_{10}	A	C	E	F	K	I

For example, assume that there are 10 music items in D and 2 features are extracted from each item. As shown in Table 1, for each feature, every item contains two local attribute patterns and one global attribute pattern. Once a query music itm_Q is submitted to this system, the IDF for each pattern is computed. Then IDF of each

pattern {A, B, C, D, E, F, G, H, I, K} are calculated as {log(10/6), log(10/4), log(10/6), log(10/5), log(10/7), log(10/8), log(10/2), log(10/3), log(10/8), log(10/9)}. By using the calculated TF and IDF, the degrees of the items in D, which represent the similarities between target items and query, can be calculated. Let us take itm_1 as a further example to explain how to calculate the related degree. Suppose the pattern set of itm_1 is {A, C, C, F, H, K} and the related TF set for {A, C, F, H, K} is {1/6, 2/6, 1/6, 1/6, 1/6}={0.167, 0.333, 0.167, 0.167, 0.167}. Also, suppose the pattern set of itm_Q is {A, D, E, G, H, H} and the related TF set for {A, D, E, G, H} is {1/6, 1/6, 1/6, 1/6, 2/6}={0.167, 0. 167, 0. 167, 0. 167, 0.333}. Thereupon the degree of itm_1 to itm_Q is $degree_{itm_1} = TF_{\{A\}}^{itm_Q} \times TF_{\{A\}}^{itm_1} \times IDF_{\{A\}} + TF_{\{H\}}^{itm_Q} \times TF_{\{H\}}^{itm_1} \times IDF_{\{H\}}$ =(0.167*0.167*0.222)+ (0.333*0.167*0.523)=0.035. Afterward degrees of all items in D can be derived. If the number of returned items is set as 5, the final playlist contains itm_9, itm_6, itm_1, itm_4 and itm_5 with the degree set {0.054, 0.044, 0.035, 0.028, 0.017}.

3.3.2 Relevance Feedback Phase

After initial query phase, the user can obtain a playlist. This phase is composed of several feedbacks. A feedback indicates that, the user determines the positive items and then resubmits the positive items to the system. The whole procedure for each feedback starts with positive examples generated at the preceding feedback are received by the system. If the precision of the preceding feedback exceeds the proposed switch threshold, the intensive search is performed. Otherwise, the extensive search is performed. Note that, the precision represents the ratio between the number of positive examples and the number of negative examples. The details of extensive search and intensive search are described as follows.

A. Extensive Search. The major idea behind this search method is to expand the search space while the precision is low. In detail, the search might fall into local space if the precision is low. Basically, this search can be carried out by the following steps.

Step 1. *From the positive examples, find the most different positive example as an expanding query point.* The difference between two positive examples is defined as

$$diff(itm_i, itm_j) = \sqrt{\sum_{ms \in (MS_{itm_i} \cup MS_{itm_j})} (TF_{ms}^{itm_i} - TF_{ms}^{itm_j})^2} \tag{3}$$

where MS_{itm_i} and MS_{itm_j} stand for the pattern sets in itm_i and itm_j, respectively. According to Equation (3), the differences among the possitive items can be derived and the most different possitive item can be determined. Following the above examples, assume that the possitive item set picked by the user at the preceding feedback is {itm_1, itm_6, itm_9} and the precision of the preceding feedback is 3/5 below the assumed switch threshold 0.7. Thus the search strategy is shifted to extensive search. The difference set {*diff*(itm_1, itm_6), *diff*(itm_1, itm_9), *diff*(itm_6, itm_9)} is {$\sqrt{4}$, $\sqrt{12}$, $\sqrt{12}$}. Finally we can thereby obtain the most different item "itm_9".

Step 2. *Generate query points.* In addition to the most different positive item, the other positive items are fused into another query point. In this paper, a query point is composed of several patterns. For a query point generated by the most different item, it can be defined as qp_{new}^{dm}. For a query point fused by other positive items, it can be

defined as qp_{new}^{p}. For example, regarding to Table 1, the extended query point generated by itm_9 is {B, E, G, H, K} and another query point generated by itm_1 and itm_6 is {A, C, D, F, H, K, I}. The TF sets related to qp_{new}^{dm} and qp_{new}^{p} are {1/6, 2/6, 1/6, 1/6, 1/6} and {2/12, 3/12, 1/12, 2/12, 2/12, 1/12, 1/12}, respectively.

Step 3. *Calculate the pattern weights according to the variance of pattern frequency.* It is because the precision in extensive search is low that the search might converge. Our intention in this step is to re-weight the pattern to expand the search space. The basic idea behind this operation is that, the higher the frequency variance of pattern *ms* between two sequential feedbacks, the more important the pattern *ms*. The weight of pattern *ms* is defined as

$$ExWeight(ms)=\sqrt{\frac{\sum_{itm_i \in PA}(TF_{ms}^{itm_i}-TF_{ms}^{qp_{old}})^2}{PN}} \tag{4}$$

where PA is the positive item set at the preceding feedback, *PN* is the cardinality of positive items containing pattern *ms*, and qp_{old} is the query point generated by positive patterns at the preceding feedback. Regarding above examples, let us take pattern C as an example to show the weight of C. The weight of C is $\sqrt{}(((TF_{\{C\}}^{itm_1}-TF_{\{C\}}^{qp_{old}})^2+(TF_{\{C\}}^{itm_6}-TF_{\{C\}}^{qp_{old}})^2)/2)=\sqrt{}(((0.33-0)^2+(0.167-0)^2)/2)=0.263$. Note that, here qp_{old} is itm_Q. Finally, the weights of all patterns {A, C, D, F, H, K, I} are generated and normalized, with respect to {0.016, 0.327, 0.016, 0.016, 0.207, 0.208, 0.208}.

Step 4. *Calculate the degrees of the items in D to* qp_{new}^{p} *and* qp_{new}^{dm}, respectively. In this paper, the degree of item itm_i is defined as

$$degree_{itm_i}=\sum_{ms\in(MS_{itm_i}\cup MS_{qp_{new}})} weight(ms)*TF_{ms}^{qp_{new}}*TF_{ms}^{itm_i}*IDF_{ms}. \tag{5}$$

For example, regarding pattern set {A, C, F, H, K}, the degree of itm_1 to qp_{new}^{p} is ((0.167*0.016*0.167*0.222)+(0.250*0.327*0.333*0.222)+(0.167*0.016*0.167*0.097)+(0.167*0.207*0.167*0.523)+(0.083*0.208*0.167*0.046))=0.0092. Therefore, the degrees of the items in D to qp_{new}^{p} *and* qp_{new}^{dm} are derived by Equation (5), respectively. For example, the degree set of {itm_1, itm_2, itm_3, itm_6, itm_7, itm_8, itm_9, itm_{10}} to {qp_{new}^{p}, qp_{new}^{dm}} is {0.0092, 0.0048, 0.0035, 0.0064, 0.0048, 0.0035, 0.0031, 0.0035} and {0.0026, 0.0035, 0.0016, 0.0024, 0.0035, 0.0016, 0.0106, 0.0016}. Finally, for each item above, the degrees to qp_{new}^{p} *and* qp_{new}^{dm} are aggregated as {0.0118, 0.004, 0.0051, 0.0085, 0.004, 0.0051, 0.0137, 0.0051}. Note that, itm_4 and itm_5 are skipped here because they are identified as negative items at the preceding feedback.

Step 5. *Rank the items by the related degrees and return the ranking list.* For example, the ranking list is {itm_9, itm_1, itm_6, itm_8, itm_{10}} with the degree set {0.0137, 0.0118, 0.0085, 0.0051, 0.0051}.

B. Intensive Search. Mostly, the procedure of intensive search is the same as that of extensive search except step 1 and step 3 above. In terms of step 1, it is because the precision is high enough to make the search move toward the user's interested area in

the search space that the QEX is unnecessary. Therefore, step 1 can be skipped and there is only one query point generated by all positive items in step 2. In terms of step 3, in contrast to the extensive search, the main aspect here is that, the lower the frequency variance of the pattern, the more important the pattern. This aspect can help the search move toward the user's interested area more effectively and efficiently. In intensive search, the weight of pattern *ms* is defined as

$$InWeight(ms)=\sqrt{\frac{PN}{\sum_{itm_i \in PA}(TF_{ms}^{itm_i} - TF_{ms}^{qp_{old}})^2}}. \quad (6)$$

Based on above examples, let us take a series of examples to describe this phase. Assume that, the positive item set at the preceding feedback is $\{itm_1, itm_9, itm_6, itm_8\}$ and the precision is 4/5 exceeding the switch threshold 0.7. As a result, the intensive search is performed. First, generate qp_{new} {A, B, C, D, E, F, G, H, I, K} by the patterns of positive item set $\{itm_1, itm_9, itm_6, itm_8\}$. Second, for patterns {A, B, C, D, E, F, G, H, I, K}, calculate and normalize the related weights as {0.156, 0.078, 0.091, 0.078, 0.053, 0.156, 0.078, 0.053, 0.102, 0.156}. Third, generate the degree set of items $\{itm_1, itm_2, itm_3, itm_6, itm_7, itm_8, itm_9\}$ is {0.0028, 0.0011, 0.002, 0.0024, 0.0011, 0.002, 0.0015}. Note that, in this paper, the accumulated negative items are skipped at each feedback. Hence itm_{10} is further skipped at this feedback. Fourth, the ranking list is $\{itm_1, itm_6, itm_8, itm_3, itm_9\}$. If all items are positive, the relevance feedback phase stops at this feedback. Otherwise, the positive items will be re-submitted to the system.

4 Empirical Evaluation

The experimental data is based on the collection that contains 10 genres which came from Amazon.com, including Blues, Classical, Jazz, Latin, Rock, Metal, Pop, Country, Disco and Hip hop. Each genre contains 150 unique music pieces and the duration of each music piece is 30 sec. For each feedback, the system returns 10 music pieces and the switch threshold is set as 0.5 in the experiments. To make the experimental evaluation complete, the experiments were conducted in two primary parts, namely objective evaluation and subjective user study. For objective evaluation, it is mainly based on the automated comparisons of six kinds of methods, namely PBRF, QPM, QR, RBF [1], TreeQ [4] and Rule-based [9]. In this evaluation, each music piece is adopted as a query and the successful retrieval relies on that, the genre of returned results is the same as that of query. Figure 2 showing the experimental results for objective evaluation delivers some points. First, the precision of our proposed PBRF is the best. Second, after 10 iterations, only our proposed PBRF can exceed precision 90%. Third, QPM is more effective than QR. It delivers an aspect that, although QR is a weight-sensitive method, but the query point does not change, while QPM is able to move toward the user's interest by user's feedbacks. Fourth, traditional QPM and QR using our proposed musical patterns still perform better existing relevance feedback methods [1][4][9] on content-based music retrieval. This is a very important point that reveals our proposed musical patterns are very helpful to music relevance feedback. Fifth, from efficiency point of view, PBRF only needs 4

iterations to reach the specific precision 80%, while QPM and QR need 7 iterations, RBF needs 8 iterations, and TreeQ and Rule need more than 10 iterations. This viewpoint says that, our proposed PBRF is also promising on efficiency of music relevance feedback. As a whole, PBRF is shown to be very effective and efficient on music relevance feedback. Note that, a search process for PBRF can be completed within 1 sec. in the experiments.

In addition to objective evaluation, another issue we are interested in is: does it work well in practice? To address this issue, we implemented a website and invited 17 volunteers to participate in this system evaluation. Each user was requested to test 10 queries for 5 iterations. In this user study, we only compared PBRF and RBF because the experimental results in subjective evaluation have shown that, RBF is better than TreeQ and Rule-based methods. Hence, for each feedback, the system presents 10 results for PBRF and 10 ones for RBF without method information. That is, the participant did not know which result was yielded by PBRF or RBF in this user study. Figure 3 shows that, the precision of PBRF is pretty close to that of RBF at the 1st two iterations, but higher than that of RBF. Moreover, the precision of PBRF can reach precision 80% within 3 feedbacks, but 5 feedbacks for RBF. It tells us the truth that, our proposed PBRF, indeed, works well and outperforms existing methods in real applications.

Let us analyze the results from the statistical viewpoint by Figure 4, which shows the variances of different approaches for objective evaluation. First, the performance of TreeQ is stable because the variance is stable. However, according to Figure 2, its performance cannot reach higher precision even through 10 feedbacks. Second, in contrast to TreeQ, the variances of all the other methods are very close in the first four iterations, but that of PBRF is the lowest after four iterations. Observing Figure 2 together, it is shown that, PBRF is more stable to reach higher precision than other four methods. On average, the set of averaged variances of {RBF, TreeQ, QPM, QR, Rule-based, PBRF} for objective evaluation is {0.06, 0.01, 0.05, 0.05, 0.11, 0.04}. Since the variances are pretty close, the observation that our proposed PBRF outperforms other methods is statistically supported.

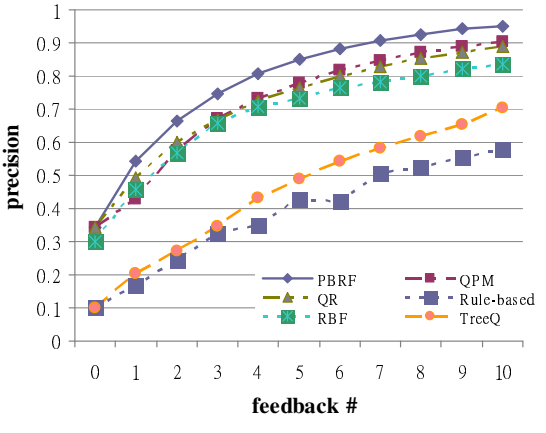


Fig. 2. The precisions of different approaches for objective evaluation

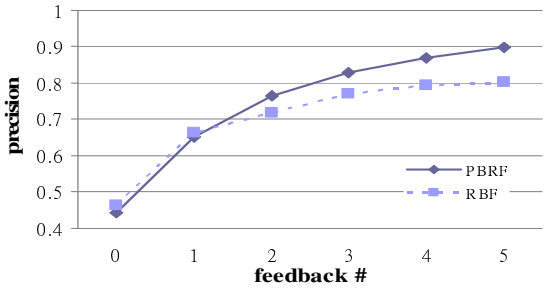


Fig. 3. The precisions of different approaches for subjective user study

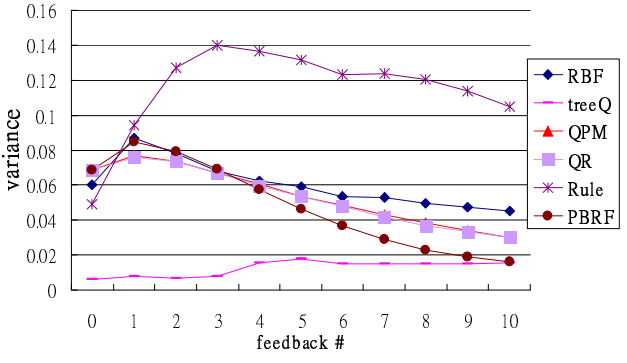


Fig. 4. The variances of different approaches for objective evaluation

5 Conclusions and Future Work

In this paper, we propose a novel pattern-based relevance feedback to enhance content-based music retrieval. The main contribution of the proposed method is that the local optimal problem is prevented by query expansion, query re-weighting and the proposed switch threshold. Aditionally, the user's interest is successfully inferred by iterative feedbacks using our proposed patterns. The experimental results reveal that the proposed method can enhance the retrieval quality more significantly than other non-pattern-based methods such as [1][4][9]. In the future, we will further focus on the following issues: First, the adaptive switch threshold will be investigated in the future. Second, we will design a more efficient data structure to face a large scale dataset. Third, in addition to music, we shall apply the same idea to different multimedia applications such as music recommendation.

Acknowledgement. This research was supported by Telecommunication Laboratories, Chunghwa Telecom, Taiwan.

References

[1] Chen, G., Wang, T.-J., Herrera, P.: A Novel Music Retrieval System with Relevance Feedback. In: Proc. of the 3rd International Conference on Innovative Computing Information and Control (2008)

[2] Fujihara, H., Goto, M.: A music information retrieval system based on singing voice timbre. In: Proc. of International Conference on Music Information Retrieval (2007)
[3] Haitsma, J., Kalker, T.: A Highly Robust Audio Fingerprinting System. In: Proc. of the International Symposium on Music Information Retrieval (2002)
[4] Hoashi, K., Matsumoto, K., Inoue, N.: Personalization of user profiles for content-based music retrieval based on relevance feedback. In: Proc. of the Eleventh ACM International Conference on Multimedia, pp. 110–119 (2003)
[5] Kosugi, N., Nishihara, Y., Sakata, T., Yamamuro, M., Kushima, K.: A practical query-by-humming system for a large music database. In: Proc. of the Eighth ACM International Conference on Multimedia, pp. 333–342 (2000)
[6] Porkaew, K., Chakrabarti, K., Mehrotra, S.: Query refinement for multimedia similarity retrieval in MARS. In: Proc. of ACMMM, pp. 235–238 (1999)
[7] Rocchio, J.J.: Relevance Feedback in Information Retrieval. In: The SMART Retrieval System – Experiments in Automatic Document Processing, pp. 313–323. Prentice Hall, Englewood Cliffs (1971)
[8] Rui, Y., Huang, T., Mehrotra, S.: Content-based image retrieval with relevance feedback in MARS. In: Proc. of the IEEE International Conf. on Image Processing, pp. 815–818 (1997)
[9] Shan, M.-K., Chiang, M.-F., Kuo, F.-F.: Relevance feedback for category search in music retrieval based on semantic concept learning. Multimedia Tools and Applications 39, 243–262 (2008)
[10] Su, J.-H., Huang, W.-J., Yu, P.S., Tseng, V.S.: Efficient Relevance Feedback for Content-based Image Retrieval by Mining User Navigation Patterns. IEEE Transactions on Knowledge and Data Engineering 23(3), 360–372 (2011)

Meta-context: Putting Context-Awareness into Context

Ramón Hervás[1], Jesús Fontecha[1], Vladimir Villarreal[2], and Jose Bravo[1]

[1] Castilla – La Mancha University, Paseo de la Universidad 4,
13071 Ciudad Real, Spain
{Rmon.HLucas,JesusFontecha,Jose Bravo}@uclm.es
[2] Technological University of Panama, Lassonde, David, Chiriquí,
Republic of Panama
vladimir.villarreal@utp.ac.pa

Abstract. Context-awareness advances have evidenced novel challenges related to context management and context-awareness usability. Most of these problems can be supported through general context attributes, i.e. properties that describe issues of the context itself, what we call meta-context. We can identify similar attributes in a variety of proposals to enhance context-aware systems. However, these attributes are usually managed internally. A common and formal model that describes the significant context attributes, their relationships and semantic axioms helps to separate the meta-context from the context model. Moreover, we propose a specific Semantic-Web-based architecture to manage the meta-context at run-time and offers related functionalities to external context-aware applications.

Keywords: Ubiquitous Computing, Context-awareness, Semantic Web, Context Attributes, Metadata.

1 Introduction

Advances in context-awareness have evidenced novel design challenges. Focusing on context management and usability, a variety of new problems have been discovered in context-aware applications. Consequently, new design approaches and functionalities are needed in order to achieve suitable context management autonomy. For this goal, we need mechanisms to provide consistency, transparency and traceability, and also address security and privacy concerns in relation to context information.

There are significant efforts to enhance context-aware systems. In that terms, most of proposals include mechanisms based on the analysis of some context properties, their processing and the launching of pre-configured policies to enhance context functionalities, such as, quality of context, privacy and security, context updating, user control, among others. By analyzing these proposals, we can identify that most of them are based on similar properties of the context. Moreover, these properties not belong to the general domain neither the application domain, the properties are general attributes related to the own context concept, as it is later described, called meta-context. However, we rarely encounter proposals that feature meta-context processing in which general attributes are communicated and fused with other

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 296–305, 2011.

systems to derive common functionalities. In this way, the separation of meta-context attributes handling and the context processing enables to manage the context model independent from the meta-context system.

In this paper, we propose an infrastructure to manage meta-context based on the Semantic Web. Unlike others proposals that identify singular context attributes, we propose an ontological model for conceptualizing attributes, their relationships and their associated axioms.

In general, the adaptation of Semantic Web principles to Ubiquitous Computing offers important benefits [1]. The representation of meta-context, in particular by means of OWL language, can provide a rich and unambiguous definition of relevant concepts. Moreover, regarding to the balance between expressiveness and computability, Semantic Web languages provide mechanisms to achieve this goal and consequently enable a formal knowledge representation to enhance the capabilities of model computational processing, its adaptability, and even promote their massive use. Our proposal enables mechanisms to manage meta-context from diverse and heterogeneous context-aware applications. In this way, general functionalities such as quality, privacy and traceability can be supported across the context information from diverse applications.

This paper is structured as follows. Section 2 is dedicated to analyze related works on meta-context information. Section 3 describes some important challenges on context-awareness regarding to meta-context information. Section 4 introduces the meta-context concept, its scope and the ontological model. Section 5 describes the meta-context infrastructure through Semantic Web principles, and finally, Section 6 concludes this paper.

2 Related Work

In this section we analyze context-aware systems with respect to the kind of upper-abstraction level elements they manage to describe low-abstraction context elements. The meta-context has not been explicitly defined previously, but several authors highlight the need for identify general attributes of the context. The main aim of these attributes is to make high-abstraction level context (HAL context) automatic available to ubiquitous systems, integrating conceptualizations to the context from different sources and enabling mechanisms to enhancing trust on models [2]. We extend this goal by identifying more specific functionalities in next section. Zimmer [3] identifies four main attributes: spatial origin, age, reliability and degree of relationship. Zimmer also defines context attributes as follows: “information on the environment that has been provided from outside the system itself, can not be sensed in any way”. This definition properly highlights the desirable independence between meta-context and the system; however, meta-context should be also independent from the source of acquisition. There are also different proposals for identifying the main context attributes. Wrona and Gomez [4] select the following elements: persistence, which can be static or dynamic; origin of context information, distinguishing between internal and external; and quality of context, compounded by several sub-attributes such as timeliness, frequency and accuracy. Fujinami et al. [5] also identify the quality of context. Moreover, they associate the context attributes with the sensor

specifications, for example, sensitivity, probability of failure and unit. Finally, in Sentient Graffiti project [6] authors manage the following context attributes: identity, creator, precise and proximity location. We present in section 4 an ontological model to define the main meta-context elements formally, and also their properties and their relationships. Moreover, we seek to endow the meta-context management with mechanisms to independent it and the own context model. It is very important to define a set of attributes enough general and independent regarding to different abstraction levels of context models.

3 Context-Awareness Challenges

Context-aware applications have been notorious in the last few years. Many surveys and projects have explored context scope and techniques for acquisition, abstraction, representation and management. However, there are important challenges to be reached. Then, we describe some issues directly related to the meta-context concept:

- Ambiguity and quality treatment: the context information may be inexact, incomplete or uncertain. If context-aware services need to appropriately achieve the user's requirements, it should understand the environment according to trustworthy context information.
- Decoupling from context management and context-aware services: separation between the management of the context and its associated functionalities, and management of the main service engineering process.
- Context information inference: usually, aggregated, derived and inferred information is based on the model language axioms and only uses the domain information through individuals of the model.
- Learning mechanisms: it is important to endow context-aware systems with learning capabilities.
- Information management: a common problem in context-aware systems is the existence of obsolete data. Manual maintenance is unfeasible in complex systems, so autonomous mechanisms must be included.
- Privacy management: privacy guarantee need to be taken into account. Privacy is directly related to concept of scrutability, i.e. mechanisms to let users inspect their related contextual information.
- Semantics extension: in general terms, the main functionality of meta-context is enlarging semantics of context models, including high-level abstraction information which usually is not taken into account or not uniformly managed in context-aware systems.

4 MetaContext: Definition and Scope

Researching on context-awareness modeling investigates the meaning of relevant elements and its representation and management. In [7] the authors distinguish three main abstraction levels: foundational models (e.g. [8,9]), that provides generic and

domain independent specifications, domain models (e.g. [10,4]) to specify the shared conceptualization on a community, and applications models (e.g. [11,12]) to describe specific application issues.

Furthermore, we often forget to describe the context from a meta-data perspective, i.e. the context into its own context. Any information about a contextual concept, and its relationships with the rest of the information, carries certain associated knowledge related to their characteristics into the model, but that is not related to the real world, neither the domain in question and neither the application logic. It is important to emphasize that the meta-context is connected to the model individuals, not to model concepts. We therefore propose this definition:

Definition: Meta-context provides us the information to describe the context of each instance in a context-model.

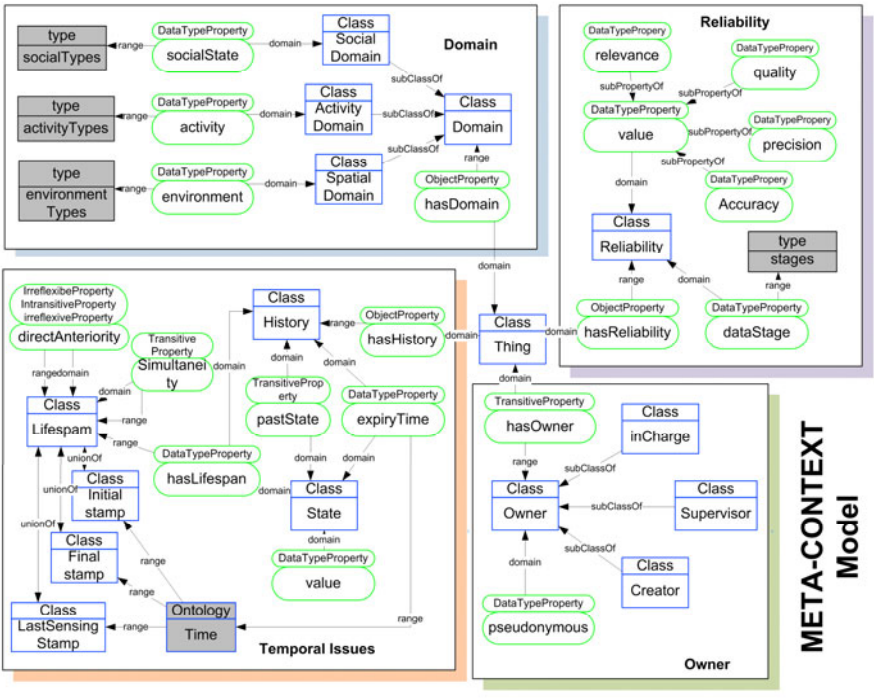


Fig. 1. MetaContext Ontological Model

The survey of previous work has been suitable to identify relevant attributes of the context in Ubiquitous Computing domain and in a user-centered perspective. From these attributes we have analyzed their relationships and we have designed a reference model. The meta-context model is built upon four main elements: context reliability, context owner, temporal scope and user domain-scope (Figure 1).

4.1 Context Reliability

Reliability is the level of trust regarding an instance that can be measured qualitatively or quantitatively. Most context-aware applications assume complete validity of the context information. Despite the advances in sensors and modeling techniques, the context information is not error-free. Ignoring that ambiguity may be a valid approach for simple context-aware applications, but it may be a requirement to consider at design time. Some authors propose mediation [10] as an ambiguity treatment mechanism. It can be a powerful mechanism but requires active interaction with the user. Context models can also include implicit mechanisms to support the ambiguity treatment, for example, delimiting the range of individuals or defining cardinality parameters (e.g., a user can only be in one location at once). The behavior rules can also identify frequent measurement errors and act accordingly. Either way, representation of quality measures related to each individual can enhance the ambiguity treatment. Quality evaluation is a function of data in several phases. Acquisition phase usually involves errors (for example, depending on sensor reliability). Abstraction process may also involve failures because it is based on a combination of data acquired. Context models, as generalization and simplification of the real world, can assume facts that are not always correct, as in the case of behavior rules.

The context reliability sub-model includes the concept of reliability itself. Reliability depends on several kinds of attributes. The sub-model conceptualizes the relevance, i.e. the relation of the context element to the specific requirements in the context-aware system. Relevance may be externally estimated or can depend on meta-context attributes, such as the history properties, the context owners and the user domain. In a similar way, the quality of context property has been defined and again, it can take values thought external mechanisms, based on meta-context elements or approaches that combine both. Finally, we represent two quality measures typically related to context acquired by sensors: precision that indicates the range within which a context value is believed to be true, and accuracy to indicate the error rate.

4.2 Context Owner

Context-aware systems can relate an individual to each other by owner's relationships. Typically, data owner is a user or a device; however, looking for higher generality, the meta-context do not restrict the range of values of owner property, i.e., the owner can be an individual of any valid element in the model. Moreover, the meta-context defines several subclasses in order to increase the information owner semantics, distinguishing between Creator-owner, Supervisor-owner, and inCharge-owner. This meta-context element may enhance the privacy and security mechanisms. Usually, context-aware applications inherit the privacy policies from traditional applications, defining access levels and associating them to user profiles. Context-aware applications need to implement more complex mechanisms for several reasons: (a) the information is distributed and stored by heterogeneous systems and devices, (b) context information is wide and diverse and requires polices based on more complex aspects than simply user profiles, (c) the diversity also supposes a higher granularity of definition, so a policy at the class level is desirable, and (d) it is

difficult to determine the policies at design time due to dynamism and context-evolution of context-aware systems. This meta-context element enables the definition of privacy and security polices based on context-model elements and, consequently, incrementing the granularity and improving the evolution of the policy. Information owners also enhance the maintenance of contextual instances, for example, deleting individuals whenever their owner disappears as valid context element in the model.

4.3 Temporal Scope

Temporal scope is the temporal interval in which a contextual individual holds specific values without significant changes, including past states, the current value or future estimates. It is important to note that irrelevant or insignificant changes (for example in sensor measurements) are not taken into account. Additionally, the temporal scope sub-model conceptualizes the expiration time and the last time that the context element was sensing or generating to its current state. Finally the sub-model includes some attributes to represent the theory of temporal order based on anteriority and simultaneity properties.

Time-scope represents the history of the context that can be a sequence of past contexts that let the actual situation. It can be viewed as a representation of the dynamic character of the context, but usually, since any time representation is granular, context history be represented as a discrete succession of atomic intervals. Sometimes, temporal scope is more important than the particular individual value. For example, a service to detect open doors needs to know the period in which the open-door property has a true value in order to throw an alert if someone forgot to close the door, but not when a user goes in and close the door immediately. On the other hand, time-scope enables the definition of inference rules based on time. Also, It is possible to define expiration properties for model individuals, enhancing the system's autonomy in maintenance. Finally, temporal scope is valuable information for learning engines in context-aware applications, enabling their proactive adaptation based on historical measures at run-time.

4.4 User-Domain Scope

Context model is usually focused on fixed application domains. However, it may be necessary to import sub-models from external domains (for example, by ontology integration). The context-model has to be able to distinguish the user domain or domains where an element is defined and in which of these domains this element is valid. The meta-context definition includes several kinds of user domains: social domains (e.g., workmates, family, friends…) activity domains (work, leisure, etc.) and physical domains (the office, the household, roads, etc.). Domain-issues may be considered at the concept level; however, this can induce ambiguity problems if the same concept is valid in several domains but the individuals are valid in a given domain only. The meta-context engine includes the domain description at the individual level, reducing ambiguity problems.

5 Implementation Based on the Semantic Web

Our proposal is based on the principles of the Semantic Web and provides an infrastructure to allow data to be shared and reused across context models and systems, using formal description of context elements and of their relationships.

5.1 Ontological Model

The meta-context model has been described using OWL. This simplifies the interoperability. The meta-context description is based on a semantic expanse of the owl:Thing superclass. The main meta-context elements have been defined as owl:Thing properties, maintaining the consistency of the full model and enabling future modifications to the meta-context model without affecting the rest of the system. The main elements are reliability, owner, history and user domain, each modeled by a set of OWL classes and properties. The proposed ontological model is shown in Figure 1.

5.2 Architecture

The meta-context is partially dependent on the context model, because its elements take values from valid model instance ranges. In this way we attain interoperability and coherence between meta-context and model classes and individuals. However, the formal definition stand-alone is not enough to ensure the meta-context integration. The meta-context is on a higher level of abstraction, i.e., it describes common properties of the context elements. Consequently, it is necessary to manage the context-model and the meta-context independently. Another difficulty is the range treatment. Some meta-context elements take values from any valid range in the context model, depending on the model class that is being described. For example, the history elements store past values of an individual. The range of the history element is the same as the range of the individual supervised, and it follows that the particular range cannot be known at design time. Meta-context makes a generic definition and a special treatment is required at run time to disambiguate the range.

Taking into account the above issues (that will be discussed ahead), we can enumerate the following responsibilities of the meta-context engine:

- Model adaptation: removing the inconsistencies related to the owl:Thing superclass redefinition. This module supervises the context elements that are subscribed to the meta-context, dynamically.
- Ambiguity resolution: disambiguating the value ranges of the meta-context elements. It is necessary to match the range of meta-context elements with valid elements in specific context model, at runtime.
- Meta-context maintenance: functionalities related to the persistent storage of meta-context individuals in repositories.
- Interface with the functional engines: attending requests and providing meta-context information. The functional engines are abstract modules to perform specific functionalities. A concrete implementation for these modules is presented in [13].

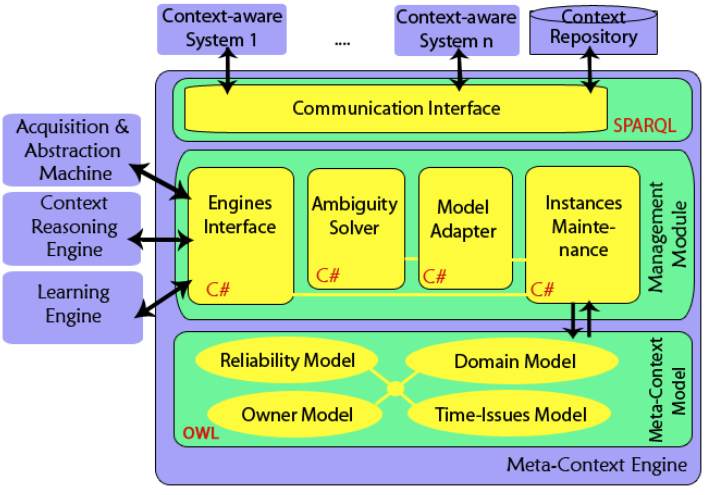


Fig. 2. MetaContext Infraestructure

The infrastructure architecture is shown in Figure 2. The system is built upon .Net Framework. The business logic has been developed in C# language, the persistent storage is managed by using LINQ and the meta-context elements are requested through SPARQL.

The formal definition of meta-context by the OWL language supports its integration with formal ontological models in general and Semantic-Web-based models particularly. However, it requires special treatment because of several reasons. Next, we synthesize the encountered problems, their causes and the proposed solutions:

Thing class omnipresence

- Problem: The meta-context functionalities applied to the whole context model induce consistency problems whenever is applied to elements that have not explicitly related meta-context information.
- Example: The garbage collector may define that a context element asserted as dynamic must to be deleted if its value has not changes in a while. Context elements that not have related meta-context information may be erroneously removed.
- Cause: The superclass Thing in OWL affects to any concept in the model. However, the meta-context should be related to a subset of the model, i.e. only context elements responsive to meta-context should be linked to meta-context information. Consequently, it is necessary to adapt the model to avoid inconsistencies in the redefinition of the super class owl: Thing.
- Solution: The proposed architecture includes a module to maintain which context elements are subscribed to the meta-context, dynamically.

Range ambiguity

- Problem: Some meta-context attributes have been defined ambiguously and, consequently, the consistency of the meta-context model cannot be directly ensured.
- Example: The specific range of elements such as the owner and past value is not defined in the model.
- Cause: The range of some attributes cannot be decided at design-time because it depends of the range of the related context element.
- Solution: The range of those elements has been defined as owl:Thing superclass. This solution does not ensure the consistence for self-serve. For this reason, additionally, a set of behavioral SWRL rules have been include for triggering consistency checking related to context element that need it.

Ontological interoperation

This issue is accord to the declared "interoperability nightmare" [14] regarding to the proliferation of devices that need to be connected in ubiquitous computing environments. The same problem occurs when several context models need to interoperate and share portions of them. One of the most important benefits of Semantic Web is "serendipitous interoperability", the ability of a software system to discover and utilize services it have not seen before, and that were not considered while systems. However, in this direction, there are a variety of challenges to achieve. The proposed architecture no offers direct solutions to this problem far beyond the functionalities that Semantic Web offers. Present and future advances in ontology alignment and interoperation may enhance the proposed infrastructure.

6 Conclusions

High-level abstraction context information is usually forgotten or applied in a general level, as simple context attributes. This paper proposes an OWL meta-context description through the redefinition of the owl:Thing and its associated functionalities, enabling higher-granularity control of the context information. The ontological model includes significant context attributes and also, their relationships and semantic axioms. This model contributes to ubiquitous computing community in order to identify and use context attributes to enhance common requirements such as privacy, quality of context, maintenance and interoperation among others.

Additionally, we describe a preliminary architecture to help identify and understand some problem related to the OWL implementation. This paper describes several solutions to mentioned problems and justifies the use of Web Semantic principles. In fact, we can conclude that the ontological approach based on the Web Semantic offers a variety of benefits in meta-context implementation despite singular identified problems.

Acknowledgments. This work has been financed by the TIN2010-20510-C04-04 projects from the MICINN (Spain).

References

1. Lassila, O., Adler, M.: Semantic Gadgets: Ubiquitous Computing Meets the Semantic Web. In: Spinning the Semantic Web: Bring the World Wide Web to Its Full Potential, pp. 363–376. MIT Press, Cambridge (2003)
2. Wang, Z., Shi, J.: A Model for Urban Distribution System under Disruptions of Vehicle Travel Time Delay. In: Proc. Intelligent Computation Technology, pp. 433–436 (2009)
3. Zimmer, T.: Towards a Better Understanding of Context Attributes. In: Proc. Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, PERCOMW. IEEE Computer Society, Washington, DC (2004)
4. Wrona, K., Gomez, L.: Context-aware security and secure context-awareness in ubiquitous computing environments. In: Proc. Autumn Meeting of Polish Information Processing Society, pp. 255–265 (2005)
5. Fujinami, K., Yamabe, T., Nakajima, T.: Take me with you!: a case study of context-aware application integrating cyber and physical spaces. In: Proc. Symposium on Applied Computing, Nicosia, Cyprus. ACM, New York (2004)
6. Lopez-De-Ipina, D., Vazquez, J.I., Abaitua, J.: A context-aware mobile mash-up platform for ubiquitous web. In: Proc. Intelligent Environments, IE 2007 (2007)
7. Zimmer, T.: Qoc: Quality of context - improving the performance of context-aware applications. In: Advances in Pervasive Computing. Adj. Proc. Pervasive, vol. 207, pp. 209–214 (2006)
8. Schneider, L.: Designing foundational ontologies. In: Song, I.-Y., Liddle, S.W., Ling, T.-W., Scheuermann, P. (eds.) ER 2003. LNCS, vol. 2813, pp. 91–104. Springer, Heidelberg (2003)
9. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening Ontologies with DOLCE. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 166–181. Springer, Heidelberg (2002)
10. Dey, A.K., Mankoff, J.: Designing Mediation for Context-Aware Applications. ACM Transactions on Computer-Human Interaction 12(1), 53–80 (2005)
11. Hervás, R., Bravo, J., Fontecha, J.: A Context Model based on Ontological Languages: a Proposal for Information Visualization. Journal of Universal Computer Science 16(12), 1539–1555 (2010)
12. Lopez-de-Ipina, D., Lo, S.L.: LocALE: a Location-Aware Lifecycle Environment for ubiquitous computing. In: Proc. International Conference on Information Networking (2001)
13. Hervas, R., Bravo, J.: COIVA: Context-aware and Ontology-powered Information Visualization Architecture. Software Pract. Exper. 41(4), 403–426 (2011), doi:10.1002/spe.1011
14. Lassila, O.: Applying Semantic Web in Mobile and Ubiquitous Computing: Will Policy-Awareness Help. In: Proc. Semantic Web and Policy Workshop, 4th International Semantic Web Conference (2005)

Negotiation in Electronic Commerce: A Study in the Latin-American Market

Osvaldo Cairo[1], Juan Gabriel Olarte[2], and Fernando Rivera[3]

[1] Department of Computer Science, ITAM, México DF, México
[2] Department of Computer Science, University College London, United Kingdom
[3] Department of Computer Science, University of Essex, United Kingdom
cairo@itam.mx, gabriel.olarte@cs.ucl.ac.uk, friver@essex.ac.uk

Abstract. A number of research efforts were devoted to deploying agent technology applications in the field of Agent-Mediated Electronic Commerce. On the one hand, there are applications that simplify electronic transactions such as intelligent search engines and browsers, learning agents, recommender agent systems and agents that share knowledge. Thanks to the development and availability of agent software, e-commerce can use more than only telecommunications and online data processing. On the other hand, there are applications that include negotiation as part of their core activities such as the information systems field with negotiation support systems; multi-agent systems field with searching, trading and negotiation agents; and market design field with electronic auctions. Although negotiation is an important business activity, it has not been studied extensively either in traditional business or in e-commerce context. This paper introduces the idea of developing an agent with negotiation capabilities applied to the Latin American market, where both the technological gap and an inappropriate approach to motivate electronic transactions are important factors. We address these issues by presenting a negotiation strategy that allows the interaction between an intelligent agent and consumers with Latin American idiosyncrasy.

Keywords: Intelligent agents, e-commerce, web intelligence, competitive negotiation, virtual negotiation.

1 Introduction

Automated negotiation is an important type of interaction in systems composed of autonomous agents. We define a negotiation in electronic commerce as the process by which two or more parties multilaterally bargain resources for mutual intended gain, using an online platform [1]. Given the agent's ubiquity, such negotiations may exist in many different shapes and forms. We focus on competitive negotiations [2] with a hard exchange strategy rather than on cooperative negotiations or on double dealing strategies.

This paper introduces the description of a virtual intelligent agent [3], capable of negotiating (proposing offers and acting on them) in the business-to-consumer (B2C) and e-commerce transaction models, such as those applications that support

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 306–314, 2011.

commercial transactions among final consumer and virtual enterprises [4, 5]. The agent is also capable of learning user preferences so as to plan future transactions. We use both a formal negotiation protocol that includes a necessary ontology and a defined strategy. We focused our project particularly on the Latin-American market, where both the technological gap and an inappropriate approach to motivate electronic transactions are important factors. We respectfully submit that our new agent should reduce the impact and gap created by commercial technology [6], making it possible for more and more people to get involved in a new and more accessible way of doing electronic commerce.

2 Behavior of the Latin-American Market

The Latin-American market lacks an appropriate tool to perform automated competitive negotiations over the Internet. The way e-commerce is approached is either static (without using its inherent advantages), or dependent upon methods like auctions – mostly English language auctions – which are not well-suited for the particular Latin-American way of thinking [7, 8]. This method has not been successfully exploited due to the following reasons:

- Most people don't know exactly the processes followed in an electronic auction.
- The average Latin-American consumer focuses primarily on the price. Thus, the perception of prices going up instead of going down, due to the extended use of auctions, has a negative impact.
- The fixed closing time in electronic auctions increases the waiting period before an auction winner is declared. The Latin-American consumer does not feel comfortable with waiting to know the outcome of the auction.
- A critical mass of buyers is needed for the auction to work properly; otherwise the reserve price may not be met, leaving the item unsold or underpriced.

These are some of the reasons which suggest that the English language auction model is incompatible with the idiosyncrasies of Latin-American buyers and sellers [9]. This work proposes a negotiation strategy and model based on a "haggling approach" somewhat related to the Dutch auction [10]. We consider that the Latin-American buyer is more used to such a format and that it can provide the following advantages, among others:

- The buyer uses the same strategy as when he or she shops in real life.
- He or she gets the feeling of bargaining for a lower price because the price always goes down.
- The negotiated agreement can be reached within few minutes.
- The need for a critical mass of buyers is eliminated.

3 The Formal Model of the Agent's Domain

The domain is the environment where the interactions between the agent and the users will take place. We use the terminology and notation of set theory to describe all the elements involved and the logic behind their relations [11].

The environment has a set of states $E = \{e_1, e_2, ...\}$ wherein the agent can interact. The agent, on the other hand, has a set of possible actions for each stimulus, defined as $Ac = \{\alpha_1, \alpha_2, ...\}$. Stimulus, in turn, is represented as the collection of information containing the options that the agent perceives from the user within the environment states $St = \{s_1, s_2, \ldots\}$. Based on a set of negotiation rules $R = \{ r_1, r_2, \ldots, r_n\}$ defined in the market environment, and the interactions with the user, the agent chooses an action. This latter comprises the gathering of information that determines which option is most favorable for the user to buy a product.

The visit to the web store (v) can be summarized as a sequence of states and transitions as follows [12]:

$$v : e_{i0} \xrightarrow{\alpha_{i0}} e_{i1} \xrightarrow{\alpha_{i1}} e_{i2} \xrightarrow{\alpha_{i2}} e_{i3} \xrightarrow{\alpha_{i3}} \ldots \xrightarrow{\alpha_{in-1}} e_{in} \quad (1)$$

Where each e_{ij} represents a (possibly different) state and each α_{ij} represents a (possibly different) action.

Formally, we say an environment *Env* is a triplet $Env = \{E, \tau, e_0\}$ where E is a set of environment states, τ is a state transformation function represented concisely, and $e_o \in E$ is the initial state of the visit to the web store. Then the agent is modeled as follows [13]:

$$Ag: R \rightarrow A_c \quad (2)$$

As a result, we can finally define the system as a set that comprises the agent, the environment, and the stimulus:

$$Sys = \{Ag, Env, St\} \quad (3)$$

With (1), (2) and (3), we can design scenarios for the interactions between the agent and the user. We can also determine the protocol of negotiation and the roles for this negotiation. To enhance the user's perception of the interaction, we separated the steps to be developed into two categories. The first one is the design of an embodied virtual agent that interacts in the virtual system. The second is the agent's negotiation ability: sending and receiving proposals, bargaining, and concession-making to the user. These properties intend to adapt the process to the real Latin-American market's behavior.

4 Negotiation Model

Negotiation is a method of dispute resolution. It generally involves a dialogue to promote and motivate an agreement upon mutual courses of action, to bargain for individual or collective advantage, or to reach outcomes that can satisfy various interests. The negotiation can also be observed like a process in which a joint decision is reached by two agents with contradictory demands [14, 15]. An agent can be a person, an organization or an intelligent agent. The participants move towards an

agreement by means of a process of concessions, in search of new alternatives [16]. It is important to observe that our proposed negotiation model was used successfully recently by two important firms in Mexico, Compaq and Elektra.

4.1 Definition of Negotiation Terms

The negotiation mechanism is based on a protocol and a strategy of negotiation. More formally, a negotiation can be represented in the following terms:

a. The initial price P_{ini} is defined as:

$$P_{ini} = f(P_{\min}, RP, COMM, LB) \tag{4}$$

where

- $P_{\min}$ is the minimum price. It is the lowest price at which the store will ever sell the product. It is the clearance price, and is obtained directly from the system database
- RP is the regular price. This is the current market price for the product.
- COMM represents the fees paid by the user of the negotiation system.
- LB represents the buyer's loyalty, a qualification that is granted to the buyer according to his/her consumption record. Its proposed range of value is $0 \leq LB \leq 0.3$.

The value assigned to P_{ini} by the pricing function (4) is not necessarily equal to the market price of the product because it depends on the buyer's loyalty.

b. The reserve price P_{res} is defined as follows:

$$P_{res} = f(P_{\min}, FSD, LB, FS, COMM) \tag{5}$$

where

- FSD represents a supply-demand factor for the product. It is calculated using the number of items already sold, the quantity of remaining items in stock and the number of remaining days in which the items are expected to be sold.
- FS represents the season-of-the-year factor.

The reserve price (5) is the threshold value for a particular negotiation at which the agent may accept a proposal. It is the lowest price at which the product can be sold in such a particular scenario. It is calculated dynamically for each product and for each client and is influenced by supply and demand factors, buyer's loyalty, etc.

c. A negotiation round is a complete negotiation cycle. This includes the agent sending a proposal, the buyer receiving it and then making a counteroffer. It can be defined as:

$$Round_i = (ask_i, bid_i) \quad \text{for} \quad i = 0,...,i_{\max} \tag{6}$$

where

- ask_i represents the agent's proposal at the i-th round.
- bid_i represents the buyer's offer at the i-th round.
- $i_{\max}$ is the maximum number of negotiation rounds.

d. The agent's offer or ask is the proposal that the selling agent makes to the consumer. It is defined as follows:

$$ask_i = \begin{cases} P_{ini} & i = 0 \\ f(bid_{i-1}, ask_{i-1}, HN, P_{res}) & i = 1,...,i_{\max} \end{cases} \quad (7)$$

Where

- bid_{i-1} is the buyer's previous offer.
- ask_{i-1} is the agent's previous proposal.
- HN is the hardness of the negotiation.

e. The consumer offer or bid is the buyer's counteroffer to the agent's offer. The range of possible values is:

$$0 < bid_i \leq ask_i \quad \text{if} \quad i = 0$$

$$bid_{i-1} < bid_i \leq ask_i \quad \text{if } i = 1,2,...,i_{\max}$$

4.2 Loyalty

Loyalty is constituted centrally by perseverance in an association to which a person has become intrinsically committed. We refer to loyalty in electronic commerce as the disposition, perception or feeling by which customers are encouraged to return to an electronic retail system – "the store". We consider it as both a practical disposition and a sentiment [17]. We know that the quality of product or service leads to customer satisfaction, which leads to customer loyalty, which leads to profitability. Potentially, continuous customer return may lead to increased sales. Srinivasan identified eight factors (customization, contact interactivity, care, community, convenience, cultivation, choice, and character) that potentially impact e-loyalty, and he developed scales to measure these factors [18]. Others have emphasized the analysis of the internal database transactions [19]. We analyzed different alternatives, considering the environment and idiosyncrasy of Latin American users. We have decided to bear in mind three additional factors –personalization, presentation and persistence– vis-à-vis the wish to increase the number of clients returning to the store [20]:

1. Personalization. Information is valuable to develop loyalty.
2. Presentation. We focused on the B2C interaction model. Here, loyalty is formed by the human computer interaction and the personal touch.
3. Persistence. The persistence of context is necessary to provide an extension to the personalized information concerning the client's current interests. Recommendations made by a persistent-context agent optimize the knowledge-base based on their preferences, likes and dislikes, thus increasing their loyalty [21].

4.3 Price Determination

The reserve price used by the selling agent can be regarded as a threshold of acceptability of counteroffers [22]. Neither participant (seller or buyer) knows the reserve price. It is also variable because it can change for each negotiation session. For the seller, the reserve price is the lowest price at which they are willing to sell the product, denying any lower offer. As for the buyer, his/her reserve price is the highest price he/she is willing to pay for the product. He/she won't accept any proposal above such price.

It is important to underline that negotiations are based mainly on the reserve price. When it is calculated using the product quantities (sold, in stock), number of days, and the total number of products to be sold, simulations have verified that it is possible to follow the demand of products in a very successful way. Also, it has proven to be one successful method over current systems in order to increase profits [23]. The reserve price can be calculated using the following expression:

$$P_{res} = P_{\min} * (1 + FSD * LB * FS * COMM) \tag{8}$$

And the initial price:

$$P_{ini} = P_{\min} + RP * LB * COMM \tag{9}$$

4.4 Definition of Negotiation Zones

The negotiation zones are bands of prices defined with the objective to locate and to classify the consumer's proposals in order to calculate the agent's answer.

a. Zone A. This area is located between zero and the minimum price scaled down by a constant that depends on the hardness of the negotiation. The relationship of the consumer's bid and this area is defined as:
   If $bid_i \in ZoneA \Rightarrow o < bid_i \leq K \times P_{\min}, K \in (0,1)$

where

- K is defined by the negotiation hardness, which translates the relative importance of the product in the overall negotiation. It is obtained by calculating a normalized score that agent assigns to a value for current product inside the range of acceptable values (the higher the price, the better the agent's utility).

b. A2. This area is between the scaled minimum price and the minimum price itself. The relation of the consumer's bid and the area is:
   If $bid_i \in A_2 \Rightarrow K \times P_{\min} < bid_i \leq P_{\min}, K \in (0,1)$

c. Zone B. This area is bounded by the minimum price and the reserve price.
   If $bid_i \in ZoneB \Rightarrow P_{\min} < bid_i \leq P_{res}$

d. Zone C. This area is located between the reserve price and the initial price.
   If $bid_i \in ZoneC \Rightarrow P_{res} < bid_i \leq P_{ini}$

4.5 Possible Negotiation Actions

The actions taken by the agent at any time are mainly influenced by the buyer's offer and the negotiation zone. They are described in detail below.

1. Haggling options. These options do not appear in all the negotiation rounds. They occur based on the buyer's situation and with certain randomness. The main goals are to render the buyer more interested in the negotiation process, to increase the interaction, and to demonstrate certain possible moves that are common in real life negotiations.
2. Concessions. These are the elements of the negotiation used when the process seems to have reached an impasse after a user's haggling option. By using concessions, the agent accepts the buyer's offer (previously rejected) but can, in turn, receive valuable feedback from the user. The concessions projected are:
    - Answering a poll: The buyer is compromised to answer some quick questions, whose objective is to acquire specific information about him/her (demographic data, personal preferences, etc.). Such information can be used to personalize services, marketing information, ad campaigns, etc.
    - Friend recommendation: in this case, the user has to give out three e-mails of his friends. These will be used for ad campaigns.
    - Buy another product: The acceptance of the offer by the agent is subject to the user buying another product. The agent proposes another closely-related product, complementary of the original one.
3. Offer and counteroffer. The agent gives a price to the buyer, who generally makes a counteroffer. If such is the case then we have two scenarios: if it is considered as interesting, the agent calculates a new asking price; otherwise, the next ask is equal to the prior one.
4. Ending the negotiation. The agent can terminate a negotiation either by accepting the buyer's offer, aborting the negotiation because the buyer's offer is not good enough, or by reaching the maximum number of negotiation rounds without attaining an agreement.

5 Conclusions

This paper presents an agent-based negotiation system for electronic commerce. A competitive negotiation protocol, its strategy, and its formal model are fully described within a web system environment. The system employs a new negotiation strategy based on information stored and retrieved from a knowledge base that covers the negotiation scenarios. We mainly focused on contexts in which the seller agent assists in both the negotiation process and the acquisition of goods and services in a Web fashion.

The system defines in many aspects the electronic marketplace and a number of possible strategies that are employed during negotiation. Here, the core aspect is represented by the agent, which assists the user throughout the whole buying process. The system, in turn, was designed considering two aspects: (1) the idiosyncrasy of

most Latin-American markets and (2) how buyer's behavior differs in nature with respect to other markets around the world.

Our approach proposes the conditions needed to obtain an economically optimal agreement in B2C scenarios, based on a loyalty mechanism (i.e., personalization, presentation of the interactions and persistence of the information regarding the customer's interests and preferences). Furthermore, the method for determining the acceptability in the offer exchange process, in combination with knowledge declarations about the market price and behavior of the products, allows us to obtain an appropriate negotiation protocol that fulfills the system goals. It has been exhaustively evaluated to maximize the benefits of the disputants.

Future work may extend the analysis of the dialogues between agents and users, to determine whether the user is making a good decision or not by appraising the pros and cons of the acquisition of a product. This analysis involves the evaluation of product properties, negotiation parameters and the assessment of the followed strategy to persuade behavior in electronic commerce. Here, argumentation theory may help to improve system efficiency by analyzing and solving conflicts in the evaluation of product properties, or providing a more logical approach in product modeling with the specification and use of ontologies. Either way, we may increase the efficiency of agent-based e-commerce systems. The analysis of efficiency in the process of information exchange would help to establish appropriate negotiation protocols and also would facilitate the decision-making process.

Acknowledgements. This work has been supported by Asociación Mexicana de Cultura A.C.

References

1. Beam, C., Segev, A.: Automated Negotiations: A Survey of the State of the Art. CMIT Working Paper 97-WP-1022 (May 1997)
2. Agrawal, M., Chari, K.: Learning Negotiation Support Systems in Competitive Negotiations: A Study of Negotiation Behaviours and System Impacts. International Journal of Intelligent Information Technologies 5(1), 1–13 (2009)
3. Shen, X., Radakrishnan, T., Georganas, N.: vCOM: Electronic commerce in a collaborative virtual world. Electronic Commerce Research and Applications 1(3), 281–300 (2002)
4. Amazon Web Page, `http://amazon.com`
5. Dell Computers Web Page, `http://dell.com`
6. Rai, V., Kim, D.: Principal–agent problem: a cognitive map approach. Electronic Commerce Research and Applications 1(2), 174–192 (2002)
7. Steenkamp, J., ter Hofstede, F.: International market segmentation: issues and perspectives. International Journal of Research in Marketing 19(3), 185–213 (2002)
8. Kumar, V.: International marketing research. Prentice Hall, Upper Saddle River (2000)
9. Lynch, P., Kent, R., Srinivasan, S.: The global internet shopper: Evidence from shopping tasks in twelve countries. Journal of Advertising Research 41(3), 15–23 (2001)
10. Klemperer, P.: What Really Matters in Auction Design. The Journal of Economic Perspectives 16(1), 169–189 (2002)
11. Papadimitriou, C.: Computational Complexity. Addison-Wesley, Reading (1994)

12. Wooldridge, M., Dunne, P.: Optimistic and Disjunctive Agent Design Problems. In: Castelfranchi, C., Lespérance, Y. (eds.) ATAL 2000. LNCS (LNAI), vol. 1986, pp. 1–14. Springer, Heidelberg (2001)
13. Russell, S., Subramanian, D.: Provably bounded-optimal agents. Journal of Artificial Intelligence Research 2, 575–609 (1995)
14. Narayanan, V., Jennings, N.: An adaptive bilateral negotiation model for e-commerce settings. In: Proceedings of the 7th International Conference on E-Commerce Technology, pp. 34–39 (2005)
15. De Paula, G., Ramos, F., Ramalho, G.: Bilateral negotiation model for agent-mediated electronic commerce. In: Dignum, F.P.M., Cortés, U. (eds.) AMEC 2000. LNCS (LNAI), vol. 2003, pp. 1–14. Springer, Heidelberg (2001)
16. Kowalczyk, R., Bui, V.: On Constraint-Based Reasoning in e-Negotiation Agents. In: Dignum, F.P.M., Cortés, U. (eds.) AMEC 2000. LNCS (LNAI), vol. 2003, pp. 31–36. Springer, Heidelberg (2001)
17. Chiu, C.M., Huang, H.Y., Yen, C.H.: Antecedents of trust in online auctions. Electronic Commerce Research and Applications 9(2), 148–159 (2010)
18. Srinivasan, S., Anderson, R., Ponnavolu, K.: Customer loyalty in e-commerce: an exploration of its antecedents and consequences. Journal of Retailing 78(1), 41–50 (2002)
19. Buckinx, W., Verstraeten, G., Van den Poel, D.: Predicting customer loyalty using the internal transactional database. Expert Systems with Applications 32(1), 125–134 (2007)
20. Arafa, Y., Dionisi, G., Mamdani, A., Pitt, J., Martin, S., Witkowski, M.: Towards Building Loyalty in e-Commerce Applications: Addressing Issues on Personalisation, Persistence & Presentation. In: Proceedings of the Fourth Int. Conference on Autonomous Agents, Agents in Industry Workshop (2000)
21. Schafer, J., Konstan, J., Riedl, J.: Recommender systems in e-commerce. In: Proceedings of the 1st ACM Conference on Electronic Commerce (EC 1999), pp. 158–166 (1999)
22. Faratin, P., Sierra, C., Jennings, N.: Negotiation decision functions for autonomous agents. International Journal of Robotics and Autonomous Systems 24(3–4), 159–182 (1998)
23. Morris, J., Ree, P., Maes, P.: Sardine: dynamic seller strategies in an auction marketplace. In: Proceedings of the 2nd ACM Conference on Electronic Commerce, Minneapolis, Minnesota, United States, pp. 128–134 (2000)

Adjusted Case-Based Software Effort Estimation Using Bees Optimization Algorithm

Mohammad Azzeh

Faculty of Information Technology
Applied Science University
Amman, Jordan P.O. Box 166
m.y.azzeh@asu.edu.jo

Abstract. Case-Based Reasoning (CBR) has achieved a considerable interest from researchers for solving non-trivial or ill-defined problems such as those encountered by project managers including support for software project management in predictions and lesson learned. Software effort estimation is the key factor for successful software project management. In particular, the use of CBR for effort estimation was favored over regression and other machine learning techniques due to its performance in generating reliable estimates. However, this method was subject to variety of design options which therefore has strong impact on the prediction accuracy. Selection of CBR adjustment method and deciding on the number of analogies are such two important decisions for generating accurate and reliable estimates. This paper proposed a new method to adjust the retrieved project efforts and find optimal number of analogies by using Bees optimization algorithm. The Bees algorithm will be used to search for the best number of analogies and features coefficient values that will be used to reduce estimates errors. Results obtained are promising and the proposed method could form a useful extension for Case-based effort prediction model.

Keywords: Case-Based Reasoning, Software Effort Estimation, Bees Algorithm.

1 Introduction

Software effort estimation is very important for successful project bedding and feasibility study and during software development for resource allocation, risk evaluation and progress monitoring [1][2][3][4][5]. So it preserves popularity within research community and became an active research topic in software engineering. Case-Based Reasoning (CBR) is one of efficient methods for software effort estimation because of its outstanding performance and capability of handling noisy datasets [6]. CBR is a knowledge management technology based on premise that history almost repeats itself which leads to problem solving can be based upon retrieval by similarity. CBR has been widely used for many software engineering problems to solve non-trivial or ill-defined problems such as those encountered by project managers including

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 315–324, 2011.

support for software project management in predictions and lesson learned [2][7]. Figure 1 illustrates the process of CBR in software effort estimation. The project in the case base are described by a set of features that may be continues or categorical or both. The solution is always described by the effort needed to accomplish software project in man-month or man-day. However, the performance of CBR is a dataset dependent and has large space of configuration possibilities and design options induced for each individual dataset [8]. So it is not surprise to see contradictory results and different performance figures [9].

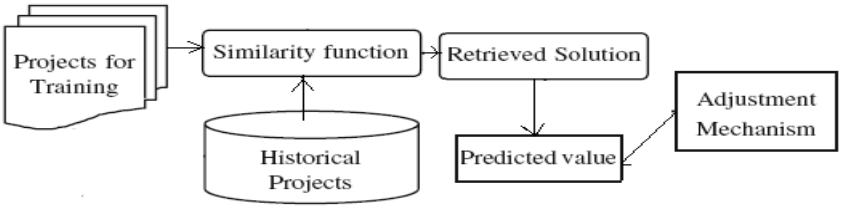


Fig. 1. Process of CBR for software effort estimation

Adjustment method is one of the most influential parameters on CBR as it fits case in hand and minimizes the variation between a new case and retrieved cases. The use of adjustment requires some parameter to be set such as number of analogies (i.e. similar retrieved solutions) and the method to make the adjustment. Our claim is that, we can avoid sticking to a fixed best performing number analogies that change from dataset to dataset. We use an alternative method to calibrate CBR by optimizing the feature similarity coefficients and number of analogies.

This paper employed Bees algorithm (BA) to optimize number of analogies (K) and coefficient values used to adjust feature similarity degrees between new case and other K analogies. The Bees algorithm is a new population-based search algorithm, it was first proposed by [10] in 2005. The algorithm mimics the food foraging behavior of swarms of honey bees. In its basic version, the algorithm performs a kind of neighborhood search combined with random search and can be used for optimization [10]. To the best of our knowledge the BA has not been used in software effort estimation and this paper shows a potential to provide more accurate results.

The rest of this paper is organized as follows: Section 2 gives an overview to Case-based effort estimation and adjustment methods. Section 3 presents the proposed adjustment method. Section 4 presents methodology of this study. Section 5 presents results of empirical validation. Finally, section 6 summarizes our work and outlines the future studies.

2 Background

Case-based effort estimation makes prediction for a new project by retrieving previously completed successful projects that have been encountered and remembered as historical projects [2] [5][7] [8] [11][12]. Although CBR generates successful performance figures in certain datasets, it is still suffered from local tuning problems when

they were to be applied in another setting [13]. Local tuning requires finding appropriate *K* analogies that fits procedure of adjustment and reflects dataset characteristics, where this process is a challenge on its own [8][14]. In literature various methods have been used to determine the best number of analogies such as nearest neighbor algorithms like *k*-NN [15], expert guidance or goal based preference [8]. Idri et al. [16] suggested using all projects that fall within a certain similarity threshold. This approach could ignore some useful cases which might contribute better when similarity between selected and unselected cases is negligible. On the other hand, several researchers suggested using a fixed number of analogies (*K*=1, or 2 or…etc) which is considered somewhat simpler than first one and depends heavily on the estimator intuitions [9][12][15][17]. Another study focusing on analogies selection in context of CBR is conducted by Li et. al. [3]. In their study they perform rigorous trials on actual and artificial datasets and they observe effect of various *K* values.

Concerning the adjustment procedure, there have been various methods developed for effort adjustment such as un-weighted mean [15], weighted mean [12], and median [18] of closest efforts. However, these adjustment methods are global tuning and directly applied to the retrieved efforts without capturing the actual differences between target project and retrieved projects. Li J. [3] proposed another adjustment approach (called AQUA) using similarity degrees of all *K* analogies as weight drivers to indicate the strength of relationship between a similarity metric and effort. Jorgensen et al. [19][20] investigated the use of 'Regression Towards the Mean' (RTM) method based on the adjusted productivity of the new project and productivity of the closest analogies. This method is more suitable when the selected analogues are extreme and availability of coherent groups. They indicated that the adjusted estimation using RTM method follows the same estimation procedure conducted by experts. Chiu & Huang [21] investigated the use of Genetic Algorithms (GA) based project distance to optimize similarity between target project and its closest projects.

However, we believe that reflection on dataset before applying to different algorithms under multiple settings is of more significance. But, this is not enough because the selection of *K* analogies is not only a dataset dependent but also adjustment method dependent. So we can conclude that finding optimal *K* analogies is an optimization problem depending on the choice of adjustment method followed. In this study we propose making use of BA to address this challenge. The illustration of the proposed approach is described in the next section.

3 The Proposed Method (CBR$^+$)

CBR adjustment is a technique used to derive a new solution by minimizing the differences between retrieved analogues and target project [18]. This technique can be represented mathematically as a function (see Eq. 1) that captures differences between problem description of target project and its analogies description in an attempt to generate more reasonable solution.

$$Effort(p_t) = F(p_1, p_2, p_3, \cdots, p_K) \tag{1}$$

Where P_t is the target project and P_1 to P_N is the top K similar projects to the target one. The similarity degree is assessed using Euclidean distance. F is the adjustment function used to capture differences between P_t and all other top similar projects, and then convert these differences into the amount of changes in the effort value. The adjustment function used in this study is illustrated in equations 2 and 3, where w_j is the optimization coefficient. The proposed CBR method (hereafter we will call it **CBR$^+$**) exploits the search capability of the BA to overcome the local tuning problem of effort adjustment. More specifically, the task is to search for appropriate weights (w_j) and K values such that the performance measure is minimized.

$$Effort(p_{ti}) = Effort(p_i) + \sum_{j=1}^{M} w_j \times (f_{tj} - f_{ij}) \tag{2}$$

$$Effort(p_t) = \frac{1}{K} \sum_{i=1}^{K} Effort(p_{ti}) \tag{3}$$

Where M is the number of features, f_{tj} is the j^{th} feature value of the target project. f_{ij} is the j^{th} feature value of the analogy project P_i.

Before starting, the BA parameters must be carefully set [10], these parameters are: problem size (Q), number of scout bees (n), number of sites selected out of n visited sites (s), number of best sites out of s selected sites (e), number of bees recruited for best e sites (nep), number of bees recruited for the other selected sites (nsp), other bees number (osp) and initial size of patches (ngh) which includes site and its neighborhood in addition to Stopping criterion which in our study is to minimize Mean Magnitude Relative Error (MMRE) performance measure. The algorithm starts with an initial population of n scout bees. Each bee represents potential solution as set of K analogy coefficient values. The scout bees are placed randomly in initial search space. The fitness computation process is carried out, using Leave-one out cross validation, for each site visited by a scout bee by calculating MMRE. This step is essential for colony communication which shows the direction in which flower patch will be found, its distance from the hive and its fitness [10]. This information helps the colony to send its bees to flower patches precisely, without using guides or maps. Then, the best sites visited by the highest fittest bees are being selected for neighborhood search. The area of neighborhood search is determined by identifying the radius of search area from best site which is considered the key operation of BA. The algorithm continues searching in the neighborhood of the selected sites, recruiting more bees to search near to the best sites which may have promising solutions. The bees can be chosen directly according to the fitnesses associated with the sites they are visiting. Alternatively, the fitness values are used to determine the probability of the bees being selected. Then the fittest bee from each patch is selected to form the next bee population. Our claim here is to reduce the number of points to be explored. Finally the remaining bees are assigned to search randomly for new potential solutions. These steps are repeated until the criterion of stop is met or the number of iteration has finished. At the end of each iteration, the colony of bees will have two parts to its new population – those that were the fittest representatives from a patch and those that have been sent out randomly. The pseudo code of BA is shown in Figure 2.

```
Input:Q,n,s,e,ngh,nep,nsp
Output:BestBee
Population ← InitializePopulation(n, Q)
While(! StopCondition())
      MMRE=EvaluatePopulation(Population)
      BestBee  ←GetBestSolution(Population)

    NextGeneration ← ø
    ngh ← ( ngh ×PatchDecreasefactor)
    Sitesbest  ←SelectBestSites(Population, s)
    for(Sitei∈ Sitesbest)
            nsp ← ø
            if(i<e)    nsp  ←  nep
            else     nsp ← osp
            Neighborhood← ø
            for( j To  nsp )
              Neighborhood←CreateNeighborhood-
Bee(Sitei,ngh )
            end
            NextGeneration← GetBestSolu-
tion(Neighborhood)
    end for
    RemainingBeesnum ← ( n - s)
    for(j To RemainingBeesnum)
        NextGeneration← CreateRandomBee()
    end for
    Population← NextGeneration()
end while
Return( BestBee )
```

Fig. 2. The BA for CBR adjustment

4 Methodology

The CBR$^+$ has been evaluated using six datasets that exhibit different characteristics. These datasets come from PROMISE online data repository [22]. The descriptive statistics of such datasets are summarized in Table 1. For each dataset we followed the same testing strategy, we used Leave-one out cross validation to identify test and train projects such that, In each run, The prediction accuracy of different techniques is assessed using MMRE, PRED(0.25) performance measure. MMRE computes mean of the absolute percentage of error between actual and predicted project effort values as shown in Eq. 4. PRED(0.25) is used as a complementary criterion to count the percentage of estimates that fall within less than 0.25 (λ) of the actual values (N).

$$MMRE = \frac{1}{N}\sum_{i=1}^{N} \frac{|Actual\ Effort(p_i) - Estimated\ Effort(p_i)|}{Actual\ Effort(p_i)} \tag{4}$$

$$PRED(0.25) = \frac{\lambda}{N} \times 100 \quad (5)$$

We also used Boxplot of absolute residuals and Wilcoxon sum rank test to investigate the statistical significance of all the results. Finally, the obtained results from the proposed approach have benchmarked to other frequently used CBR adjustment techniques such as GA based similarity degree [11], RTM [19], AQUA [3] as well as original CBR [6] method.

Table 1. Descriptive statistics of the datasets

Dataset	Cases	Categorical Features	Numerical Features	Effort mean	Effort Std.
Desharnais	77	1	10	4834	4188
COCOMO	63	2	15	406.4	657
Kemerer	15	2	4	219.25	263
Albrecht	24	1	6	21875	28417
NASA93	18	0	3	49.5	45.7
Telecom	18	0	3	284.3	264.7

5 Results

Before developing CBR$^+$ the parameter values of the BA must be set. Yet, there is no way to identify the best parameters values of BA therefore we did some investigation to select the best values for all datasets. The parameter values used in this study are: (n=30, s=30, e=20, nep=15, nsp=30, osp=20, ngh=15) and stopping criteria is to minimize MMRE). This section presents empirical validation of the CBR$^+$ against original CBR and other adjusted CBR models based on GA, RTM and AQUA, using Leave-one out cross validation. Tables 2 and 3 show the resulting MMRE and PRED (0.25). In five out of six datasets the CBR$^+$ method had the minimum MMRE and larger PRED(0.25) (i.e. the best performance among all investigated methods). However, these findings are indicative of the performance of using BA to optimize and adjust retrieved project efforts which lead to significant improvement on the overall accuracy of Case Based effort prediction. Also from the obtained results we can observe that there is evidence that using adjustment techniques can work better than using null adjustment as in original CBR (i.e. without adjustment).

As to the other adjustment variants, three results deserve some attention. Firstly, it is important to note that both RTM was not suitable for Telecom dataset due to the absence of size feature. The success of RTM rely on the availability of size feature therefore it was impossible to generate productivity feature for RTM. As consequence we did not use both adaptation techniques for Telecom dataset. Also, it has to be considered partitioning the projects into smaller, homogeneous subsets so that the procedure of RTM regresses to a local productivity mean. It was impossible to do so for Albrecht and Telecom datasets as they don't have categorical features therefore the obtained results for these datasets using RTM were not encouraging. These points could be accounted as limitations for this method and may portrait it as less effective method in the case of categorical features unavailability. Secondly, while we were

carrying out the empirical validation for CBR^+, we noticed that the general trend of accuracy improvement when using optimal number of analogies is not clear even though there is a slight improvement for some projects when number of analogies increases and this certainly depends on the choice of adjustment coefficients obtained by BA. Thirdly, There is no consistent results regarding the suitability of CBR^+ for small datasets with categorical features (as in Kemerer dataset) but we can confirm that CBR+ is still comparable to GA in terms of MMRE and PRED(0.25) and have potential to deliver good estimates. In contrast, CBR^+ showed better performance than GA for the other two small datasets (NASA and Telecom) that do not have categorical features.

Table 2. Summary of MMRE performance measure for all models on different datasets

Dataset	CBR+	CBR	AQUA	RTM	GA
Albrecht	**51.68**	71.0	63.5	71.4	55.8
Kemerer	**40.2**	55.9	61.1	81.4	33.7
Desharnais	**42.7**	60.1	60.0	47.2	56.7
COCOMO	**57.0**	157.1	109.7	68.5	76.3
Nasa93	**20.0**	81.2	81.2	23.9	43.8
Telecom	**38.4**	60.0	45.6	N/A	53.1

Table 3. Summary of PRED(0.25)% performance measure for all models on different datasets

Dataset	CBR^+	CBR	AQUA	RTM	GA
Albrecht	**54.2**	29.2	29.2	37.5	45.8
Kemerer	**46.7**	40.0	40.0	40.0	53.33
Desharnais	**44.2**	31.2	31.2	24.7	36.4
COCOMO	**40.6**	12.7	12.7	14.3	34.9
Nasa93	**77.7**	33.3	33.3	66.67	50.0
Telecom	**46.67**	33.33	33.33	N/A	44.4

Figures 3 to 8 show Boxplot of absolute residuals of different models over the employed datasets. These figures show number of interesting findings, sorted based on datasets: (1) Albrecht and Desharnais datasets (Figures 3 and 5): median and box length of absolute residuals of CBR^+ is smaller than others which demonstrate reduced variability of absolute residuals and confirm that CBR^+ is better than other adjustment methods. (2) Kemerer dataset (Figure 4): CBR^+ and GA are at the same level of accuracy because they have relatively similar median of absolute residuals and smaller box length. (3) COCOMO dataset (Figure 6): All methods are relatively at the same level of accuracy, although the use of CBR^+ produced less outliers so the use of CBR^+ is more reliable. (4) NASA93 dataset (Figure 7): median of CBR^+ absolute residuals is slightly smaller than of GA, but the box length of RTM is smaller than that of CBR^+ and CBR^+. Irrespective of that, we can consider that CBR^+, GA are more accurate as they have smaller median of absolute residuals. (5) Telecom dataset (Figure 8): Although the median of absolute residual of CBR^+ and other methods are quite same but the box of absolute residuals is skewed towards minimum value which demonstrates better accuracy.

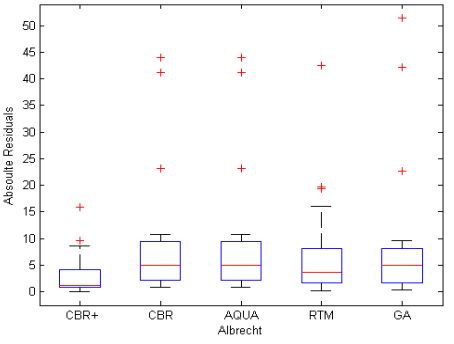


Fig. 3. Boxplot of Albrecht dataset

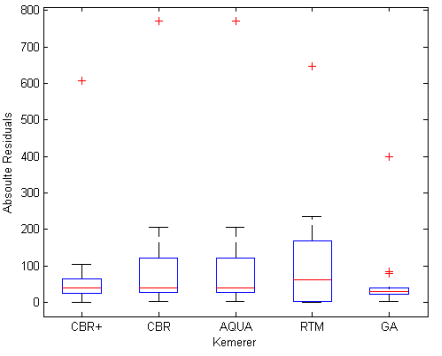


Fig. 4. Boxplot of Kemerer dataset

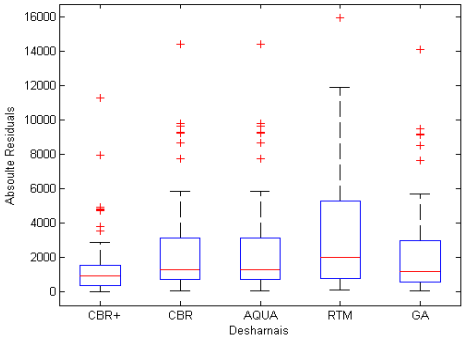


Fig. 5. Boxplot of Desharnais dataset

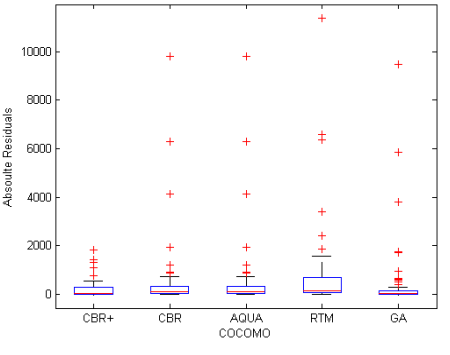


Fig. 6. Boxplot of COCOMO dataset

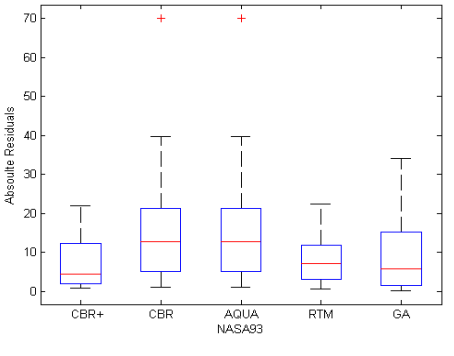


Fig. 7. Boxplot of NASA93 dataset

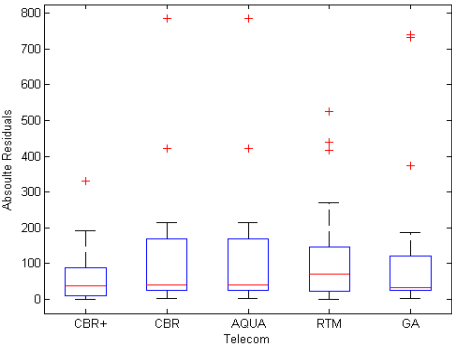


Fig. 8. Boxplot of Telecom dataset

The absolute residuals of all investigated CBR methods are taken and compared using Wilcoxon sum rank test which are presented in Table 4. Predictions based on CBR^{+} model presented statistically significant but necessarily accurate estimations than other techniques over Albrecht and Desharnais. This suggests that there is a significant difference if the prediction generated by CBR^{+} and other methods. For

COCOMO and NASA we can notice that CBR^+ produced significant results than original CBR and AQUA but insignificant than GA. Surprisingly, the statistical test results over Kemerer and Telecom demonstrate that there is no significant difference if the predictions generated by any EBA models. These findings suggest that there is still room for further improvements on CBR^+ to become superior model for the problem of software effort estimation.

Table 4. Wilcoxon sum rank test

Dataset	CBR	AQUA	RTM	GA
Albrecht	**<0.01**	**<0.01**	**0.02**	**0.01**
Kemerer	0.84	0.84	0.64	0.45
Desharnais	**<0.01**	**<0.01**	**<0.01**	**0.01**
COCOMO	**0.03**	**0.04**	**0.026**	0.75
NASA	**0.02**	**0.03**	0.51	0.41
Telecom	0.32	0.31	0.27	0.57

6 Conclusions and Future Works

This paper presents a new method for software effort estimation using combination of CBR and Bees optimization algorithm. We used the BA to automatically specify appropriate set of K analogy coefficient values that are used to adjust retrieved closest analogies. The results obtained showed significant improvements on prediction accuracy for Case-based effort estimation. One of the advantages of the proposed method is that it does not become trapped at locally optimal solutions. This is due to the ability of the BA to perform local and global search simultaneously. While we are not guaranteed that the obtained performance figures are the global optimum, the results we are presenting are the best performance ever obtained when using all features without pruning. Overall, we are encouraged by the results of the present study, which although minor in their own right, they give much more strongly to the value of performing adjustment to Case-based effort estimation when combined with other results. Nevertheless, Publication of raw results is still important so further research is necessary to investigate whether the use feature weights can also help in obtaining accurate estimates.

References

1. Auer, M., Trendowicz, A., Graser, B., Haunschmid, E., Biffl, S.: Optimal project feature weights in analogy based cost estimation: Improvement and limitations. IEEE Trans Software Engineering 32, 83–92 (2006)
2. Azzeh, M., Neagu, D., Cowling, P.: Fuzzy grey relational analysis for software effort estimation. Empirical software Engineering, 60-90 (2010)
3. Li, J.Z., Ruhe, G., Al-Emran, A., Richter, M.: A flexible method for software effort estimation by analogy. Empirical Software Engineering 12(1), 65–106 (2007), doi:10.1007/s10664-006-7552-4
4. Menzies, T., Chen, Z., Hihn, J., Lum, K.: Selecting Best Practices for Effort Estimation. IEEE Transaction on Software Engineering 32, 883–895 (2006)

5. Shepperd, M., Kadoda, G.: Comparing software prediction techniques using simulation. IEEE Trans Software Engineering 27(11), 1014–1022 (2001), doi:10.1109/32.965341.
6. Shepperd, M., Schofield, C.: Estimating software project effort using analogies. IEEE Trans Software Engineering 23, 736–743 (1997), doi:10.1109/32.637387.
7. Azzeh, M., Neagu, D., Cowling, P.: Fuzzy Feature Subset Selection for Software Effort Estimation. In: Proceedings of International Workshop on Software Predictors PROMISE 2008 (part of ICSE 2008), Leipzig, Germany, pp. 71–78 (2008)
8. Kocaguneli, E., Menzies, T., Bener, A., Keung, J.W.: TEAK: Learning Better Case Selection Strategies for Analogy Based Software Cost Estimation. IEEE Transactions on Software Engineering 6(1) (2007)
9. Briand, L.C., El-Emam, K., Surmann, D., Wieczorek, I., Maxwell, K.D.: An assessment and comparison of common cost estimation modeling techniques. In: The 1999 International Conference on Software Engineering, pp. 313–322 (1999)
10. Pham, D.T., Ghanbarzadeh, A., Koç, E., Otri, S., Rahim, S., Zaidi, M.: The Bees Algorithm – A novel tool for complex optimisation problems. In: The 2nd Virtual International Conference on Intelligent Production Machines and Systems (I*PROMS 2006), Cardiff, UK, pp. 454–461 (2006)
11. Chiu, N.H., Huang, S.J.: The adjusted analogy-based software effort estimation based on similarity distances. J. Systems and Software 80, 628–640 (2006)
12. Mendes, E., Watson, I., Triggs, C., Mosley, N., Counsell, S.: A comparative study of cost estimation models for web hypermedia applications. Empir. Softw. Eng. 8, 163–196 (2003)
13. Kirsopp, C., Mendes, E., Premraj, R., Shepperd, M.: An empirical analysis of linear adaptation techniques for case-based prediction. In: Ashley, K.D., Bridge, D.G. (eds.) ICCBR 2003. LNCS, vol. 2689, pp. 231–245. Springer, Heidelberg (2003)
14. Menzies, T., Hihn, J.: Evidence-based cost estimation for better-quality software. IEEE Softw. 23(4), 64–66 (2006)
15. Mendes, E., Mosley, N., Counsell, S.: A replicated assessment of the use of adaptation rules to improve Web cost estimation. In: International Symposium on Empirical Software Engineering, pp. 100-109 (2003)
16. Idri, A., Abran, A., Khoshgoftaar, T.: Fuzzy Analogy: a New Approach for Software Effort Estimation. In:11th International Workshop in Software Measurements, pp. 93-101 (2001)
17. Walkerden, F., Jeffery, D.R.: An empirical study of analogy-based software effort Estimation. Empir. Softw. Eng. 4(2), 135–158 (1999), doi:10.1023/A:1009872202035
18. Li, Y., Xi, M., Goh, T.N.: A study of A study of the non-linear adjustment for analogy based software cost estimation. Empir. Softw. Eng. 14, 603–643 (2009)
19. Jorgensen, M., Indahl, U., Sjoberg, D.: Software effort estimation by analogy and "regression toward the mean". J. Syst. Softw. 68, 253–262 (2003), doi:10.1016/S0164-1212.
20. Shepperd, M., Cartwright, M.: A Replication of the Use of Regression towards the Mean (R2M) as an Adjustment to Effort Estimation Models. In: 11th IEEE International Software Metrics Symposium, METRICS 2005 (2005)
21. Chiu, N.H., Huang, S.J.: The adjusted analogy-based software effort estimation based on similarity distances. J. Syst. Softw. 80, 628–640 (2007)
22. Boetticher, G., Menzies, T., Ostrand, T.: PROMISE Repository of empirical software engineering data repository, West Virginia University, Department of Computer Science, http://promisedata.org/

A New Honeybee Optimization for Constraint Reasoning: Case of Max-CSPs

Ines Methlouthi and Sadok Bouamama

SOIE Laboratory, University of TUNIS, TUNISIA
Methlouthi.ines@yahoo.fr, sadok.bouamama@ensi.rnu.tn

Abstract. In this article, we propose new approaches for maximal constraint satisfaction problems (Max-CSPs), inspired by the marriage process of honeybees. Our approaches consist on honeybees for optimization and constraint reasoning. The first one is centralized and the second one is distributed. Our approaches are enhanced by a new parameter. Experimental comparison between the two approaches and their explanations are provided. Compared to the Dynamic Distributed Double Guided Genetic Algorithm, the Distributed Honeybee Algorithm for Optimization and Constraint Reasoning is better in term of solution quality.

Keywords: Constraint Satisfaction Problem, Max-CSP, Honeybees Optimization, Distributed Approach, Swarm Intelligence, Artificial System.

1 Introduction

Recently, much research has been devoted to developing new metaheuristics to solve combinatorial and numeric optimization problems [14]. These modern metaheuristics can be classified according to the simulated phenomena. This classification includes two important groups. The first one is Evolutionary Algorithms (EA); we can cite the most popular one, Genetic Algorithms (GA) [4], [8]. The second classification is based on swarm intelligence.

Swarm intelligence has attracted many studies to model the behavior of social insects such as ants and bees to use them specifically on the problem solving area where they show impressive collective capabilities to solve those problems [11]. Honeybees are one type of social insect which exhibit many features such as division of labor, communication and cooperative behaviors that distinguish their use as models for intelligent behavior.

This paper is inspired by the marriage behavior of honeybees to solve maximal constraint satisfaction problem. This behavior has already been modeled in [7] but in our approach, we arrange otherwise the tasks of the colony members to produce an optimization search approach. We call the Centralized Honeybee Algorithm for Optimization and Constraint Reasoning C-HoBO (for Centralized Honeybee Optimization), and the Distributed Honeybee Algorithm for Optimization and Constraint Reasoning D-HoBO.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 325–334, 2011.

This paper is organized in the following way: first, we present the max-CSPs formalism. Then we introduce the marriage process in honeybees with some background material. Section 4 presents the centralized honeybee algorithm and its basic concept. Section 5 presents the distributed honeybee algorithm and its agent structure. In section 6 we present experimental setup and the results of comparisons between our two approaches and between D-HoBO and D³G²A (Distributed Dynamic Double Guided Genetic Algorithm) [8]. Finally, conclusions are then drawn.

2 Max-CSPs Formalism

Formally, CSP is a triple (X, D, C):

- X is a finite set of variables $\{x_1, x_2...x_n\}$,
- D is a function which maps each variable in X to its domain of possible values, of any type, and D_{xi} is used to denote the set of objects mapped from xi by D. D can be considered as $D = \{Dx_1, D\ x_2...Dx_n\}$,
- C is a finite, possibly empty, set of constraints on an arbitrary subset of variables in X. These constraints are represented in *Extension* or in *Intention*. [8]

Many CSP extensions have been tackled in the literature, such as Max-CSP.The Max-CSP consists in finding a total assignment satisfying as many constraints as possible. [2]

3 Marriage Process in the Honeybee

Honeybee optimization is based on swarm intelligence and motivated by the intelligent behavior of the honeybee. This optimization has been proposed by the simulation of the model of self-organization of bees [16]. In this model, although each bee performs a single task, it takes the combined efforts of the entire colony to survive and to perform a number of complex tasks such as hive construction, marriage process, harvesting pollen, etc …

In this paper, we are interested in one type of action, which is the marriage process. First we will present the colony structure. Then the marriage process and finally, we will present the artificial analogue model will use in our approaches.

3.1 Colony Structure

Normally a colony has a:

- a single queen: is the only sexually developed female; her primary function is reproduction, and she produces both fertilized and unfertilized eggs,
- fifty to sixty thousand workers at its peak: are the smallest and most numerous bees in the colony; they ensure many functions like foraging for nectar, and feeding the brood (newborn bee),

- several hundred drones (male bees): are the largest bees in the colony; their main function is to fertilize the virgin queen during her mating flight; only a small percentage of drones fulfill this function.

3.2 Marriage Process

The marriage process begins when the queen leaves the hive to mate with drones. She first circles the hive to orient herself to its location, and then leaves the hive.

Drones follow the queen, guided by her chemical odor, and mate with her in the air. A queen mate with seven to twenty drones. In each mating, sperm is accumulated in the sperm theca to form the genetic pool of the colony. Then the queen retrieves at random a mixture of the sperms accumulated in the sperm theca to fertilize the eggs. [10]

3.3 The Artificial Analogue Model

The mating flight can be considered as a set of partial solutions in a search space (the environment) where the queen mates with the drones encountered on that search space. In our approach we suppose that each bee can play the role of queen, drone or worker, and we restrict the functionality of a set of workers to find good drones for the queen.

Our approaches have three main stages:

- the algorithm starts with the dispersion of sets of workers in the search space to find drones using a local search algorithm,
- creation of new broods by crossing the drones' genotypes with the queens,
- replacement of weaker bees by fitter broods.

4 Centralized Honeybee Algorithm for Optimization and Constraint Reasoning for Max-CSPs

In this section, we introduce the basic principles of C-HoBO and its application to the maximal constraint satisfaction problems (Max-CSPs) (e.f Algorithm 1).

C-HoBO begins by randomly generating the initial colony and according to the FV range (Fitness Value: the number of satisfied constraints by a partial solution), it chooses the queen (best FV). Then the reminder of the colony member, one by one, will perform a local search process. This process finds the best neighbor (the best drone) of each member of the colony, using *the method of descent* (which will be later detailed).

Then the queen, who is waiting for the reminder colony members to complete their search process, starts crossing her genotype with the drone's one. To accomplish that, she will use *the haploid crossover*, since all drones are haploid. At the crossover step, we will suppose that the drone lose the half of its genetic information using a genotype marker, [7]. The haploid crossover process will be detailed.

Finally, C-HoBO will choose the new bee for the next generation of the colony based on their FV.

```
Algorithm 1: C-HoBO process
begin
Population  :=GenerateInitialPopulation;
repeat
  GenerationNB ++ ;
  if (compteurLOD == LOD) then
      penalizeSolution (queenGenotype);
      GenerateARandomlySolution;
      CompteurLOD = 0;
  end if
  QueenGenotype:=ChooseQueen;
  foreach bee in Population do
      Drone:=LocalSearchP rocess(BeeGenotype);
      brood :=Crossing(Drone; queenGenotype);
  end foreach
  if (BestSolutionFound == QueenGenotype) then
      compteurLOD ++;
  end if
until GenerationNB is attained;
Announce Best Solution Found;
End
```

To enhance our approach, we add a parameter that we call LOD (for local optima detector) already used in the genetic algorithm [9]. LOD represents the number of generations in which the marriage process does not offer improvement, i.e. if the FV of the queen remains unchanged for a specific number of generations; we can conclude that C-HoBO is trapped in a local optimum. In this situation we will penalize this queen genotype and we replace it by another randomly generated one to try to explore another area of research.

4.1 Local Search Process: The Method of Descent

Several local search algorithms exist in the literature. They explore the space of research point by point, starting from a first solution, and choose with each iteration a solution close to the current solution. Hence, the local search is a heuristic method for solving combinatorial problems. Moreover, local search algorithms are based on their strategy heuristic and meta-heuristic. In this paper we use the method of descent which is a local search method which stops at the first optimum found.

4.2 Haploid Crossover

Eventually, genes exist in pairs, but in our algorithm, we suppose that a haploid drone lose half of its genes because of representation. In the haploid Crossover [6], we randomly generate a genotype marker to specify the marked and unmarked gene from the drone genotype. After the crossover process, the brood genotype inherits the unmarked gene from the drone genotype and the marked gene from the queen genotype.

5 D-HoBO Algorithm for Max-CSP

When we observe a bee colony, we can immediately conclude that each bee has a simple task to do. The combined efforts of the entire colony perform complex work. Let us remember that many works have been done to show the outperforming of distributed [9] and parallel [5] metaheuristics. That is why we propose to distribute C-HoBO using the multi-agents systems. The idea here is to divide C-HoBO process to a sub-process, and each process will be affected to an agent.

5.1 Basics Principles

In our Distributed approach, we integrate the species and ecological niches concepts: the species consists of several organisms having common characteristics whereas the ecological niche represents the task performed by a given species. [8]

Goldberg states that the sexual differentiation based on specialization via both the building of species and the exploitation of ecological niche provide good results [4]. So the idea here is to divide the colony into sub-colonies enthroned by only one queen, using the multi-agents paradigm, and to assign to each sub-colony a bee worker called "Finder". This partition is based on fitness value. Let's explain that the genotype of each bee represents a partial solution for the problem and a sub-colony consists of bees having their fitness value in the same range.

5.2 The Agent Structure

D-HoBO is organized into three types of agents: the queen, the finder and the Interface agent.

Queen Agent
A queen agent has got as acquaintance the interface agent; its static knowledge consists of the max-CSP data (i.e. the variables, their domains of values and the constraints) and the crossover function. Its dynamic knowledge is her genotype which varies from one generation to another, and the drone's genotype.

Finder Agent
The finder represents a local search algorithm. In this paper, he presents the descents methods. Each finder agent receives bees having the same fitness value and he applies the local search algorithm to give the best drones in the neighborhood of those bees. So this agent has as acquaintance the interface agent. Its static knowledge consists of the max-CSP data and the search algorithm. Finally, its dynamic knowledge consists of bees for which he is responsible.

Interface Agent
An interface agent has as acquaintance all the finder agents and the queen. Its static knowledge consists of the max-CSP data and the fitness function range. Its dynamic knowledge is the best partial solution and the member of the colony, which varies from one generation to another.

5.3 D-HoBO Algorithm

The interface agent will randomly generate the initial colony and according to the FV range, it creates a queen (best FV) and partitions the rest of the colony into sub-colonies. Then interface agent asks finders to perform their local search. So each finder will treat each member of this sub-colony separately. It carries out the local search process. This process finds the best neighbor (the best drone) of each member of his sub-colony, using the method of descent, and communicates its result to the interface. The interface saves the genotype of the recruited drone in her memory and communicates them to the queen, who is waiting to start her producing process to give broods (newborn). This process consists of crossing her genotype with the drone's using haploid crossover [7]. Finally, the interface agent will recruit the new bee colony for the next generation based on FV.

6 Experimentation

The purpose of our experimentations is to check the performance of the proposed approaches. The implementation has been done with MadKit [13]. A concurrent object language implemented above the Object Oriented language JAVA. This choice is justified by its convivial aspect and its originality to be based on an organizational model rather than an agent architecture or specific design of interaction.

Experimentations were applied to randomly generated CSPs samples using a random generation model [3]. The samples tests are binary CSPs and they are based on four parameters *(n, d, p, q)* where:

- *n* is the number of the problem variables,
- *d* is the number (or the size) of each domain,
- *p* is the probability between the number of the problem effective constraint to the number of all possible constraints (Density),
- *q* is the probability between the number of forbidden pairs of values to the size of the domain cross product.

In our experimentations, we use as numerical values *n=20, d=20*. Having chosen the following values *0.9, 0.7, 0.5, 0.3, 0.1* for the parameters p and q, we obtain 25 density-tightness combinations. For each combination, we randomly generate 30 examples. Therefore, we have 750 examples. It is significant to note that random costs have been assigned to each constraint. Moreover and considering the random aspect of our approaches, we have performed 10 experimentations per example and taken the average without considering outliers.

Experimental Design
Regarding Honeybee optimization parameters, we use a number of generations equal to 5, the size of the colony equal to 300, and for the LOD value we vary it to *2, 3, 5* and *7*, in order to determine the best value for the remaining experimentations. Regarding D^3G^2A parameters, we use a crossover probability equal to 0.5, a mutation probability equal to 0.2 and a random replacement. It can be noted that the use of these parameters is supported by [12].

The performance is assessed by the two following measures:

- Run time: the CPU time requested for solving a problem instance,
- Satisfaction: the number of satisfied constraints.

The first one refers to the complexity and the second to the quality. In order to have a quick comparison of the relative performance of the two approaches.

In the next section, first, we present experimentations done to determine the best value of Local Optimum Detector. Second, we will check the effects of the distribution by comparing the results given by the centralized and the distributed versions of our approaches. Finally, we will compare our distributed approach and the Dynamic Distributed Double Guided Genetic Algorithm.

6.1 Local Optima Detector

To check the best value of Local Optima Detector, we have assembled the CPU-time averages and fitness values averages for different values of LOD.

Figure 1, shows that the best value of LOD is 2. In fact, the best fitness value is reached for this LOD value. The latter can be explained by the importance of diversification; that is to say, the more we diversify the research the more we improve the solution through the exploration of new areas of research.

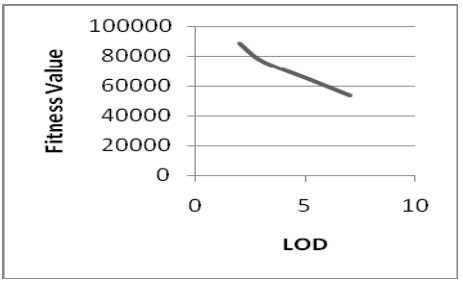


Fig. 1. Fitness value relative to different values of LOD

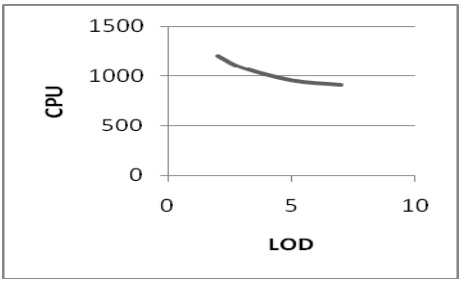


Fig. 2. CPU times relative to different values of LOD

Figure 2 shows the CPU time relative to different values of LOD. We can see that when LOD increases the CPU time decreases due to the support of the LOD.

This can be explained by the extra treatment that would be done when a solution remains unchanged for LOD generations.

In order to choose the best LOD value for a good quality of solution at a reasonable time to our experimentations, we will compute the CPU-Time-fitness compromise. The latter is considered by dividing the average of fitness value by the average of CPU-times. The best value of this compromise is given LOD equal to 2.

6.3 Comparison between D-HoBO's Results and C-HoBO's Results

To compare the performance of our two approaches, we will have recourse to two ratios:

$$\textit{Ratio of satisfaction} = \frac{\textit{Number of satisfied constraints by D-HoBO}}{\textit{Number of satisfied constraints by C-HoBO}}$$

$$\textit{Ratio of run time} = \frac{\textit{CPU time required by C-HoBO}}{\textit{CPU time required by D-HoBO}}$$

Thus, C-HoBO performance is the numerator when measuring the CPU time ratios, and the denominator when measuring fitness value ratio. So, any value greater than 1 of these ratios will indicate a better performance for our D-HoBO.

Figure 3 shows the variations of the satisfaction ratio. We notice that C-HoBO finds a slightly better solution than D-HoBO, especially for hard problems.

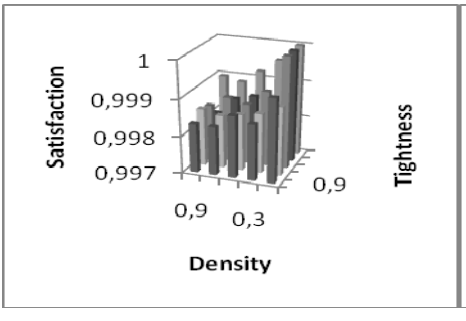


Fig. 3. Variation of satisfaction ratio

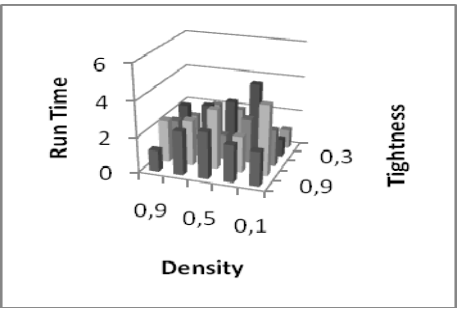


Fig. 4. Variation of run time ratio

From CPU time ratio point of view, Figure 4, D-HoBO requires less times for all types of instance of problems.

So we can conclude that the centralized version finds a slightly better solution than D-HoBO especially for hard problems, but require more time for all type of problems, due to the parallelization of the processes on multiple agents.

6.4 Comparison between D-HoBO's Results and D³G²A's Results

To compare the performance of the two approaches, we will have recourse to two ratios:

$$\textit{Ratio of satisfaction} = \frac{\textit{Number of satisfied constraints by D-HoBO}}{\textit{Number of satisfied constraints by } D^3G^2A}$$

$$\textit{Ratio of run time} = \frac{\textit{CPU time required by } D^3G^2A}{\textit{CPU time required by D-HoBO}}$$

Thus any value greater than 1 of these ratios will indicate a better performance for our D-HoBO.

From CPU time ratio, Figure 5, D³G²A requires less time. We can explain this by the synchronous process of all agents in D-HoBO, which is not the case in D³G²A, because the transition from one generation to another cannot take place unless all agents have accomplished their tasks.

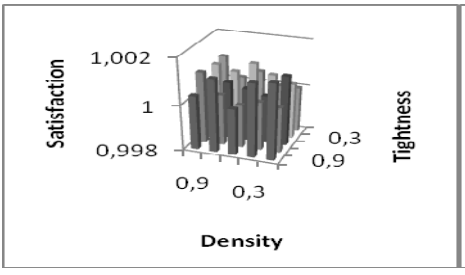


Fig. 5. Variation of run times ratio

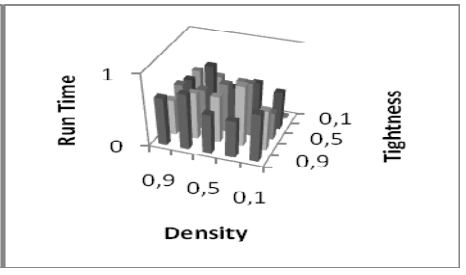


Fig. 6. Variation of satisfaction's ratio

From the solution's quality point of view, Figure 6 shows that D-HoBO finds, for most problems, a better quality of solution than D^3G^2A, especially for the over-constrained and for the most strongly tight set of examples.

So, on one hand, the elapsed time of D-HoBO is due to the synchronous process of all agents. On the other hand, the result quality (satisfaction) given by D-HoBO is explained by the intensification of the optimization process and the diversification thanks to the "Finder" bee.

7 Conclusion and Perspectives

In this paper we propose two versions of a nature inspired metaheuristic, centralized and distributed approaches for solving Max-CSPs. The latter is based on the marriage process in honeybees.

We experimented with this approach for a group of randomly generated CSPs samples. From the analysis of the experimental results, D-HoBO was very successful in terms of execution time compared to its centralized version, due to the distributed process. On the other hand, C-HoBO provides a better solution specifically for hard problems. Compared to D^3G^2A, D-HoBO was very successful, in terms of quality of solutions provided. We can explain this by the intensification sub-process. According to those experiences, we believe that the optimization process which is based on the intelligence of honeybees may be a promising way to give good results for combinatorial problems.

Further work could be focused on implementing these new approaches on real problems known in the literature because the benchmarks used in the experiment are rather academic benchmarks that may be insufficient or do not reflect reality. We can also expand, in future research, the scope of this approach to include other extensions of CSP.

This work is a new starting point for exploring new avenues offered by artificial swarm intelligence.

References

[1] Jonsson, A.P., Klasson, M.: The approximability of three-valued Max-CSP. SIAM Journal on Computing 35 (December 2006)
[2] Bannaceur, A.O.H.: Computing lower bound for Max-CSPs problems. In: IEA/AIE 2003 Proceedings of the 16th International Conference on Developments in Applied Artificial Intelligence (2003)
[3] Smith, B.: Locating the phase transition in binary constraint satisfaction problems. Research report series, Report 94.16 (1994)
[4] Goldberg, D.E.: Genetic algorithms in search optimization, and machine learning. Addison-Wesley, Reading (1989)
[5] Ghazali Talbi, E.: Parallel Combinatorial Optimization. Wiley, Chichester (2006)
[6] Abbass, H.: A single queen single worker honey bees approach to 3-sat. Technical report, Northcott Drive, Canberra ACT, 2600, Australia (2001)
[7] Abbass, H.: An agent based approach to 3-sat using marriage in honey-bees optimization. International Journal of Knowledge-Based Intelligent Engineering Systems (KES) 6(2), 1–8 (2002)
[8] Bouamama, K.S.: A dynamic distributed double guided genetic algorithm for optimization and constraint reasoning. International Journal of Computational Intelligence Research 2, 181–190 (2006)
[9] Bouamama, K.S.: A family of distributed double guided genetic algorithm for Max-CSPs. International Journal of Knowledge-based and Intelligent Engineering Systems 10(5), 363–376 (2006)
[10] Kimsy, L.: Page. Migration and dispersal of spermatozoa in spermathecae of queen honeybees (apis mellifera l.). Experientia 40(2), 182–184 (14 ref.) (1984) ISSN: 0014-4754 Codenexpeam
[11] Bonabeau, M.E., Theraulaz, G.: Swarm intelligence: From natural to artificial systems. Oxford University Press, Oxford (1999)
[12] Bouamama, N.S., Ghedira, K.: Load balancing for the dynamic distributed double guided genetic algorithm for Max-CSPs. International Journal of Artificial Life Research 1, 69–87 (2010)
[13] Gutknecht, O.: Madkit, a generic multi-agent platform. In: Agents 2000 Proceeding of the 4th International Conference on Autonomous Agents, pp. 78–79 (2000)
[14] Karaboga, P.D.: D.T. Intelligent optimization techniques. Springer, London (2000)
[15] Freuder, R.J.W.E.C.: Partial Constraint Satisfaction. Journal Artificial Intelligence. Special volume on constraint-based reasoning 58, 1–3 (1992)
[16] Seeley: The wisdom of the hive: The social physiology of honey bee colonies. Harvard University Press, Massachusetts (1995)

Hybrid Virtual Sensor Based on RBFN or SVR Compared for an Embedded Application

Kuncup Iswandy and Andreas König

Institute of Integrated Sensor Systems, University of Kaiserslautern
Erwin-Schroedinger-str. 12, 67663 Kaiserslautern, Germany
{kuncup,koenig}@eit.uni-kl.de
http://www.eit.uni-kl.de/koenig

Abstract. In numerous practical applications the interesting measurands are not explicitly available by existing sensors or unaffordable due to high cost of explicit sensor principle. Virtual sensors, as one particular means to design intelligent integrated sensory systems (I2S2), offer an solution to this problem, by merging various sources of information to generate the desired measurand for the given environmental stimuli. In this paper, radial-basis-function-networks (RBFN) and support-vector-regression (SVR) are compared for knock-detection in combustion engines with regard to ease of learning, generalization, and resource-efficiency. Additionally, the notion of a hybrid virtual sensor (HVS) is introduced here for invariance and complexity reasons. In our experiments, real-world engine data has been applied for method comparison and recommendations for parameter settings. SVR shows better generalization results than RBFN for the criteria correlation coefficient and absolute mean error are applied. In future work, we will integrate HVS concept in our emerging tool for automated I2S2 design.

Keywords: Support vector machines, regression, virtual sensor, neural networks, combustion engine.

1 Introduction

In contrast to classification tasks, where available data from sensor or other measurement input is mapped or quantized to a limited number of categories or classes, a rich number of applications exist, where a continuous-valued output or figure-of-merit is required. In numerous practical applications the interesting quantities or measurands are not explicitly available by existing sensors or unaffordable due to high cost or realization problems of the explicit sensor principle. Also, feasibility issues, e.g., such as, linearity, accurateness, robustness, stability, etc., could be of importance. Virtual sensors, as one particular means to design intelligent integrated sensory systems (I2S2), offer an solution to this problem, by merging or fusing the input of various sources of information to generate the desired measurand, e.g., product properties or process conditions, for the given environmental stimuli.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 335–344, 2011.

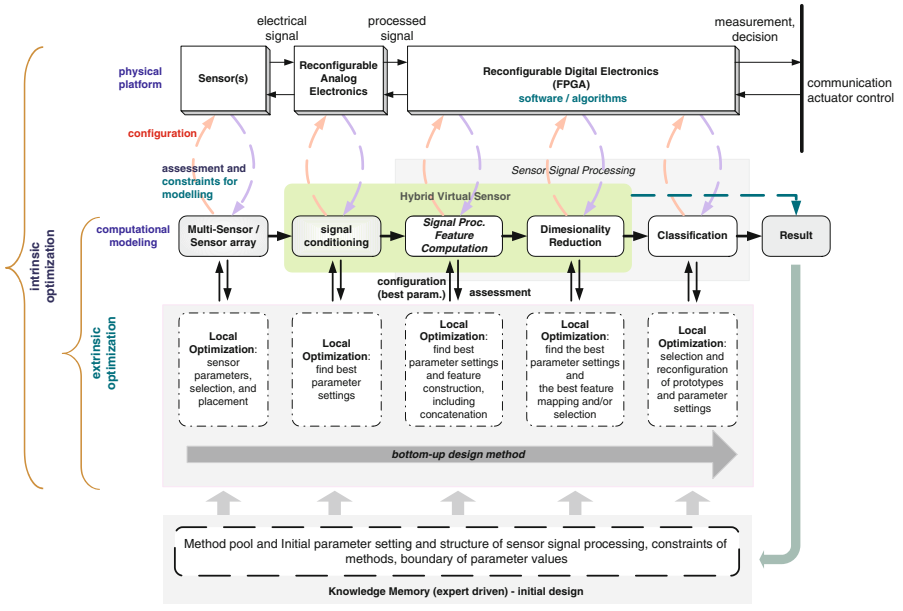


Fig. 1. Virtual sensor in automated design of intelligent sensor systems [3]

Basically, a functional mapping from a multi-dimensional input to one- or multi-dimensional output has to be generated, fusing the information of various sensor sources and extracting the desired relationship. For the realization of a virtual sensor, actually any method for function approximation can be considered, e.g., backpropagation networks [2], cascade-correlation networks, radial-basis-function-networks, or more recently support-vector-regression. The appropriateness of each possible method has to be explored in comparative studies to provide experience and recommendations for new applications. Backpropagation networks and SVR have already found application and comparison in [1]. In this paper, the two kernel methods RBFN and SVR will be compared with regard to ease of learning and parameter settings, generalization capability, and resource-efficiency for a real-world benchmark problem. Knock detection in gasoline combustion engine is employed in this work, which represents an embedded industrial application with resource constraints and high volume. Though it is feasible to achieve the desired virtual sensor by a single adaptive module, robustness, invariance, and complexity advocate the elaboration of application-specific hybrid structures, that incorporate existing knowledge in a mathematical model for data normalization, similar to feature extraction, before passing it to the adaptive module. Thus, the here investigated solution is structured in a mathematical transformation model with heuristic parameter settings for the influential parameters revolution speed and load and an adaptive part based on one of the two regarded kernel approaches. The goal of the work is to compare the methods and give recommendations for appropriate choice and parameter settings in new applications.

The overall architecture of the virtual sensor matches the core blocks of our architecture for the automated design of intelligent integrated sensor systems. This is indicated in Fig. 1. The continuous value obtained from the virtual sensor in some applications can represent the final result, but it can also be the basis of following categorization. In addition to the discussed potential benefits, virtual sensors can also potentially serve for estimation or interpolation of intermediate values on a different time-scale, e.g., in multi-rate signal processing, as well as prediction of future values. The extension of adaptation to deployment and runtime offers the perspective to add so called self-x features, e.g., self-calibration, to the concept.

The remaining paper is organized as follows. In Section 2, the application of knock detection and the general concept of virtual sensors are briefly described. Section 3 discusses the machine learning methods used for conceiving virtual sensors. The comparison in terms of computational complexity and achieved experimental results are given in Section 4. Finally, Section 5 concludes this work and discusses future work for improvement.

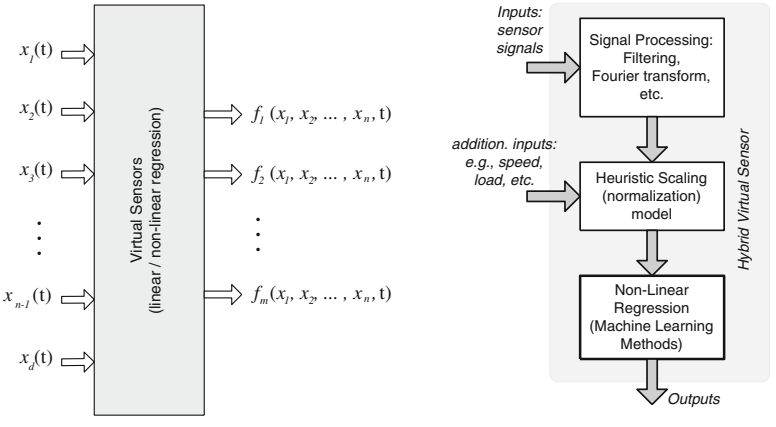
2 Hybrid Virtual Sensor and Its Application in Knock Detection

In general, a virtual sensor is a conceptual device whose either single or many output or inferred variables can be modeled in terms of other parameters relevant to the same process. Figure 2(a) shows this general concept of virtual sensors. In implementation of the virtual sensors, the training set contains elements which consist of paired variables of the independent input variable and the dependent output variable, which needs to be estimated. For example, the independent variable in the functional relation

$$y = f(\mathrm{x}), \tag{1}$$

is x (an input vector) and the dependent variable is y (a scalar). The value of the variable y depends, through the function f, on each of the components of the vector variable $\mathrm{x} = (x_1\ ,\ x_2\ ,\ x_3\ ,\ ...,\ x_d)$.

The current knock sensor technology based on structural-born vibration signals is relative quite cheaper compared to more accurate sensors based on pressure principles, which are used as a reference values. Virtual sensor concept is applied in this application by mapping the output of knock sensor signals to the filtered peak pressure values. The raw sensor signals basically in time domain are preprocessed by signal processing and feature computation methods to obtain a compact and invariant representation. The bandpass filter and Fourier Transformation (FT) are applied to obtain the power of selective frequencies. The speed revolution and load have an influence in the measurement results of combustion engines in which, the number of data sampling and the energy of sensor signals at the same peak pressure values can be different for speed and load variations. Due to these variations at each operating point, the model trained at certain operating point cannot be applied for different operating points. To overcome this



(a) virtual sensors in general (b) hybrid virtual sensor

Fig. 2. General concept of virtual sensors and hybrid virtual sensor in an knock detection application

generalization problem, the power spectrum data has to be scaled up or down by means of a normalization model with heuristic parameter settings based on the speed revolution and load information. After the scaling process, the non-linear regression model is trained by the scaled measurement data. Figure 2(b) shows this processing chain as a concept of hybrid virtual sensor.

3 Standard Machine Learning Techniques

Here two standard machine learning methods used to build virtual sensor, i.e., radial basis function networks and support vector regression are described.

3.1 Radial Basis Function Network

RBF network for regression is used to estimate a function from some examples of input-output pairs without knowledge of the function form. It is known as a supervised learning method. The function is learned from the examples which are given in the training set. The construction of a RBF network contains three layers, i.e., (1) input layer: made up of source nodes that connects the network to its environment; (2) hidden layer: applies a nonlinear transformation, which contains a radial basis function, i.e., Gaussian function; and (3) output layer: creates a linear transformation from hidden unit activations to the target output [8]. The output of hidden layer is the probability of computed distances between the input vector and prototypes selected from the training set. Based on Gaussian function, the output of hidden units is computed as

$$h_i(\mathrm{x}) = \exp(-\frac{\|\mathrm{x} - \mu_i\|^2}{2\sigma_i^2}), \tag{2}$$

where i is the prototype index and μ_i is the vector determining the center of the basis function. The μ_i is trained using the input data. Then, the last layer activation is computed as

$$f(\mathrm{x}) = \sum_{i=1}^{k} w_i h_i(\mathrm{x}) + w_0, \tag{3}$$

where w_0 and w_i are the weight factors, which are optimized in the learning process. The only parameter to be set by user is the spread (σ), which control the sensitivity of the radial basis function.

3.2 Support Vector Regression

In solving the regression problem, SVR solves a convex optimization problem, which has one globally optimal solution. Other non-linear regression methods like MLP often have local optima in their error surface as a function of the parameters of a non-linear model. To solve the overfitting problem, SVR generates a model that approximates all pairs (x_i,y_i) in the training set with ϵ precision in loss function. The error is set to zero if the error is less than some small ϵ, which is usually set to the level of typical noise in the training data. To allow additional error beyond ϵ for each training data, the slack variable ξ is added in the convex optimization problem. The convex optimization problem of support vector regression [4] can be formulated as

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ &\text{subject to} \quad \begin{cases} y_i - \langle \mathrm{w}, \mathrm{x}_i \rangle - b \le \epsilon + \xi_i \\ \langle \mathrm{w}, \mathrm{x}_i \rangle + b - y_i \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \end{aligned} \tag{4}$$

where the constant $C > 0$ is to measure the trade-off between the model complexity or flatness of $f(\mathrm{x})$ and the tolerance of the deviations larger than ϵ. The slack variables ξ and ξ^* represent the upper and lower constraints on the output of the system.

Solving the optimization problem discussed above is by using the Lagrangian function and obtaining the dual formulation of SVR optimization problem, which can be written as follows

$$\begin{aligned} &\max_{\alpha,\alpha^*} \quad L_D = -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\mathrm{x}_i \cdot \mathrm{x}_j) + \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)y_i - \epsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) \\ &\text{subject to} \quad \begin{cases} 0 \le \alpha_i, \alpha_i^* \le C, \text{for all} i \in \{1, 2, ..., l\} \\ \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0, \end{cases} \end{aligned} \tag{5}$$

where α_i and α_i^* are Lagrangian multipliers. The resulting regression estimation is computed as follows

$$f(\mathrm{x}) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)(\mathrm{x} \cdot \mathrm{x}_i) + b, \tag{6}$$

where $\mathrm{x_i}$ are the support vectors.

By using the kernel trick, the support vector algorithm can be applied to solve a non-linear dataset by mapping the given points into a higher dimensional space where it might be approximated linearly. The inputs appear in dot products as shown in the above equation and it is sufficient to be replaced by a kernel $K(\mathrm{x}_i, \mathrm{x}_j) = \Phi(\mathrm{x}_i) \cdot \Phi(\mathrm{x}_j)$. There is no need to explicitly know the transformation function of $\Phi(\cdot)$, if it is known that the kernel function is equivalent to the dot product in other high dimensional space [4]. In this paper, the Gaussian kernel is chosen, since this kernel function works nonlinearly, has only one parameter to be set, i.e., σ, and has fewer numerical difficulties. The Gaussian kernel is computed as

$$K(\mathrm{x}_i, \mathrm{x}_j) = \exp(-\frac{\|\mathrm{x}_i - \mathrm{x}_j\|}{2\sigma^2}). \tag{7}$$

To find the optimum model, three parameters that need to be set are C, ϵ, and σ. This can be done by Grid search method [5] or employing the evolutionary computation [7], e.g., genetic algorithms or swarm optimization.

4 Experimental Results

In our experiment, the benchmark data sets of different steady state operating points based on measurements of three-cylinder-engines were applied [4]. The operating points of engines were from 1400 to 5400 rpm with half or full load. The examination process of virtual sensor methods are shown in Fig. 3. For compact and invariant representation, the raw sensor data was processed by an adequate signal processing and feature computation. The Fourier Transformation (FT) and spectral filtering were sufficiently applied in this case. The spectral coefficients were then used for training and testing the model of virtual sensor compared with the filtered cylinder peak pressure values as the target. The parameter of RBFN was easier to be set than the parameter settings of SVR. Time required to find an optimal model of both methods is more or less the same effort. Here the Matlab Neural Networks toolbox was used and the spread parameter of Gaussian function was set in the range of [0.1, 5]. The implementation of SVR function in Matlab platform was adopted from [9]. The searching γ parameter of Gaussian kernels was set in the range of [0.4, 1]. Two parameters C and ϵ were set to 10^5 and 0.05, respectively.

In the first set of experiments, each of the measurement files were equally divided into two subsets, i.e., 50% for training data and the rest 50% for test data. However, the 10% furthest examples of data sets were only included in the test set to examine the performance of trained models for the examples, which were out of range. Due to large amount of examples per dataset, the experiments for all data sets were done in single run. The experimental results showed that SVMR performed better regression outputs than RBFN with regard to absolute mean error and correlation coefficient, but required more computational effort, e.g., large number of kernels, than RBFN (see Table 1). Figure 4(a) and 4(b) showed the regression results of engine data at operating points of 2600 and 3000 rpm, respectively, with half load, where the model of RBFN seemed to be overfitting the training sets.

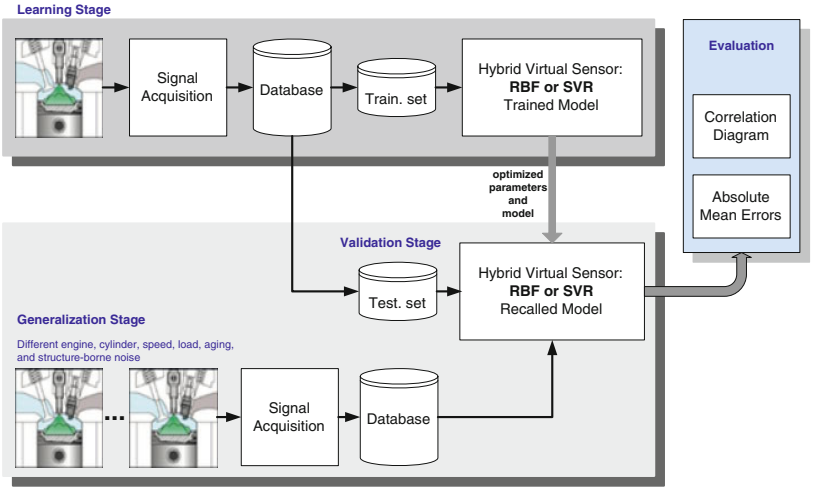
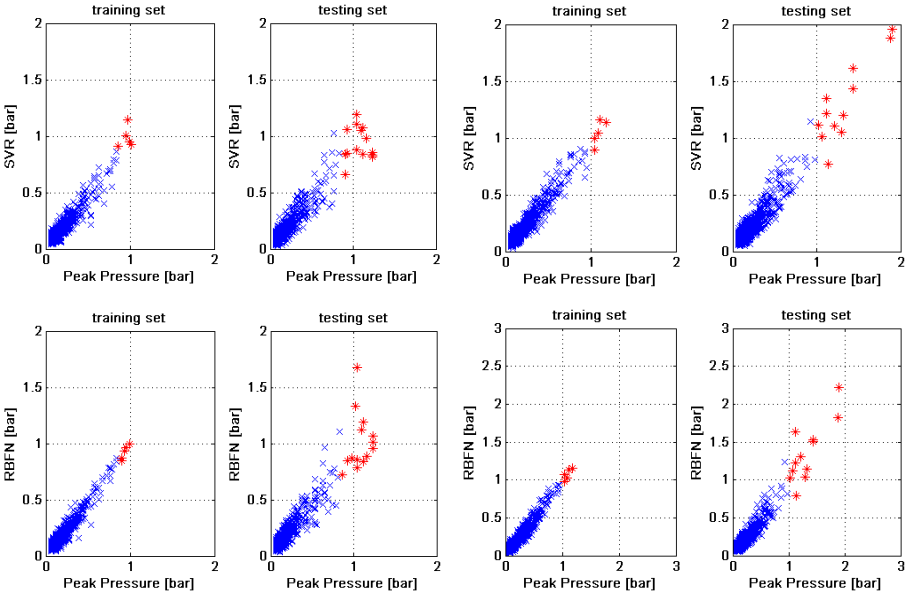


Fig. 3. Examination process of hybrid virtual sensor approach in the application of knock detection

Table 1. SVR vs. RBFN on testing set (50% of complete set) with regard to correlation coefficient, absolute error, and number of kernels

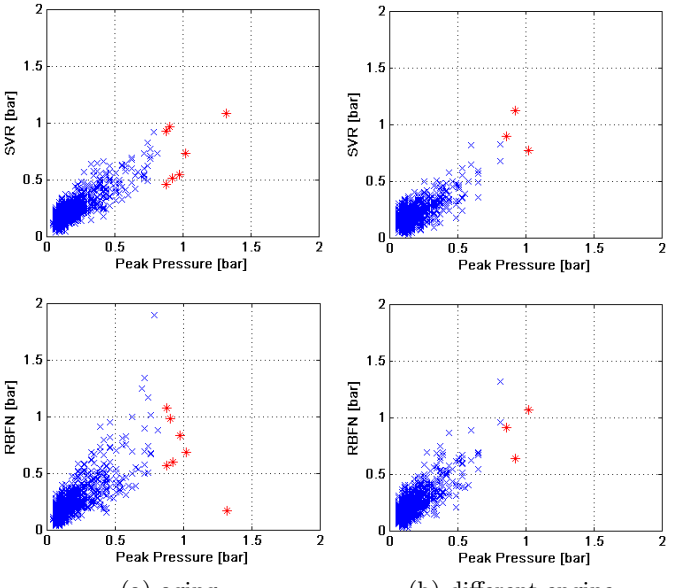
Speed [rpm] Load (H/F)	SVR			RBFN			Total examples
	corr.	$\|error\|$ μ (σ)	kernels	corr.	$\|error\|$ μ (σ)	neurons	
1400(H)	0.8665	0.0289 (0.0276)	101	0.8847	0.0263 (0.0270)	100	1474
2200(H)	0.9387	0.0378 (0.0377)	261	0.9362	0.0374 (0.0393)	79	2378
2600(H)	0.9296	0.0368 (0.0426)	259	0.9294	0.0348 (0.0438)	147	2360
3000(H)	0.9397	0.0366 (0.0358)	416	0.9416	0.0344 (0.0372)	112	3549
3800(H)	0.9457	0.0675 (0.0797)	713	0.9493	0.0685 (0.0712)	55	3022
4600(H)	0.9098	0.0480 (0.0504)	550	0.9158	0.0484 (0.0469)	52	3112
5400(H)	0.8488	0.0701 (0.0784)	717	0.8579	0.0710 (0.0728)	125	3372
1400(F)	0.9468	0.0398 (0.0537)	210	0.9365	0.0400 (0.0607)	109	1607
2200(F)	0.9276	0.0266 (0.0274)	131	0.9319	0.0243 (0.0283)	74	2211
3000(F)	0.9432	0.0399 (0.0460)	350	0.9360	0.0420 (0.0494)	95	2776
3800(F)	0.8878	0.0875 (0.0838)	573	0.8887	0.0891 (0.0806)	29	2877
4600(F)	0.8433	0.0698 (0.0701)	947	0.8429	0.0713 (0.0703)	97	3616
5400(F)	0.8502	0.0922 (0.0870)	772	0.8548	0.0923 (0.0842)	94	3256

In the second set of experiments, the generalization examination was focused on the aging and variation aspects. The measurement files from an aged-counterpart of the three-cylinder-engine and the different engine of the same type were available and included as test sets for generalization examination. For example, at the same operating points (e.g., 2600 rpm), the measurement data of new engine was used as a test set by the trained model from old engine. In



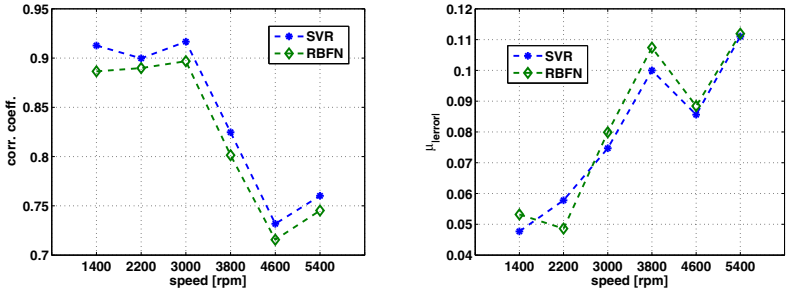
(a) Cylinder 1: 2600rpm (b) Cylinder 1: 3000rpm

Fig. 4. Results for training and test sets: SVR (top) vs. RBFN (bottom)



(a) aging (b) different engine

Fig. 5. Generalization results. SVR: $\mu_{|error|} = 0.048$, $\sigma_{|error|} = 0.044$ (aging) and $\mu_{|error|} = 0.058$, $\sigma_{|error|} = 0.043$ (variation), while RBFN: $\mu_{|error|} = 0.058$ $\sigma_{|error|} = 0.063$ (aging) and $\mu_{|error|} = 0.071$ $\sigma_{|error|} = 0.051$ (different engine).



(a) correlation coefficient (b) absolute mean error

Fig. 6. Assessment criteria for 1400 to 5400 rpm and full load by comparing the SVR and RBFN

these experiments, the RBFN model was failed to perform reasonable outputs for large peak pressure values, while SVR performed well in general for both cases as shown in Fig. 5. The correlation coefficients of SVR were 0.85 for aging and 0.70 for different engine, while the correlation coefficients of RBFN were 0.79 and 0.75 for aging and different engine, respectively.

Moreover, the generalization capability and the robustness of the SVR and RBFN for different engine speed and load ranges were studied and compared. The experiments were done to be able to cover the all ranges with the least possible effort, possibly even with only one trained model of the SVR or RBFN. For example, this was realized by training the SVR and RBFN at one single operating points (3000 rpm, half load) and then applying the model for testing the entire engine operating range. Without a speed-dependent heuristic transformation model, the regression outputs for both SVR and RBFN decreased for the operating points that are not specifically learnt. Therefore, the spectrum signals were normalized by the heuristic engine speed normalization model. These experimental results showed in Fig. 6(a) and 6(b) for correlation coefficient and absolute mean error, respectively, where the SVR performance once again outperformed the RBFN in all operating ranges. However, the RBF was more suitable for embedded application with regard to resource constraints because of a small number of kernels as compared to the SVR.

5 Conclusions

In this paper, virtual sensors based on radial basis function networks and support vector regression have been presented and compared for the application of knock-detection in combustion engines with regard to ease of learning, generalization capability, and resource-efficiency. The concept of a hybrid virtual sensor, as for invariance and complexity reasons, has been structured in a mathematical transformation model with heuristic parameter settings for the influential parameters speed revolution speed and load, and an adaptive part based on one

of the two regarded kernel approaches. The performances of SVR was clearly superior compared to the RBFN, however, the RFBN had less number of kernels or neurons in hidden layer than SVR (Gaussian kernels).

In future work, the new approach of reducing number of support vectors (kernels) will be considered for more resource-efficiency. The hybrid virtual sensor concept will be integrated into our architecture and emerging tool for automated intelligent integrated sensor system design. The heuristic parameters of normalization model along with the adaptive part will be locally optimized.

References

1. Srivastava, A.N., Oza, N.C., Stroeve, J.: Virtual Sensors: Using Data Mining Techniques to Efficiently Estimate Remote Sensing Spectra. IEEE Trans. on Geoscience and Remote Sensing 43(3), 590–600 (2005)
2. Gimmler, H., Gruden, I., Holdgrewe, K., Nester, U., Pischinger, S.: Verfahren zur Bestimmung relevanter Grössen, die den Zylinderdruck in den Zylindern einer Brennkraftmaschine. In: Deutsches Patent- und Markenamt, Offenlegungsschrift DE19741884A1 (1999)
3. Iswandy, K., König, A.: Methodology, Algorithms, and Emerging Tool for Automated Design of Intelligent Integrated Multi-Sensor Systems. Journal Algorithms (Open Access) 2, 1368–1409 (2009), doi:10.3390/a2041368
4. Iswandy, K., Kempf, S., König, A., Sloboda, R.: Robustness Investigation of a SVM-Based Knock Detection Method. Motortechnische Zeitschrift 71(7-8), 20–24 (2010)
5. Bao, Y., Liu, Z.: A Fast Grid Search Method in Support Vector Regression Forecasting Time Series. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 504–511. Springer, Heidelberg (2006)
6. Smola, A.J., Schölkopf, B.: A Tutorial on Support Vector Regression. Statistics and Computing 14, 199–222 (2004)
7. Iswandy, K., König, A.: Fully Evolved Kernel Method Employing SVM Assessment for Feature Computation from Multisensor Signals. International Journal of Computational Intelligence and Applications 8(1), 1–15 (2009)
8. Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice Hall Int., Englewood Cliffs (1999)
9. Gunn, S.: Support Vector Machines for Classification and Regression, http://www.isis.ecs.soton.ac.uk/resources/svminfo/

Skyline Adaptive Fuzzy Query

Wei Yan[1], Cecilia Zanni-Merk[2], and Francois Rousselot[2]

[1]LGECO/INSA Strasbourg, 24 Boulevard de la Victoire,
67084 Strasbourg Cedex, France
[2]LSIIT/FDBT Team (UMR CNRS 7005) - Pôle API BP 10413,
67412 Illkirch Cedex, France
{wei.yan,cecilia.zanni-merk,francois.rousselot}@insa-strasbourg.fr

Abstract. In recent years, Skyline query based on a multi-dimensional space has become a hot topic in the research of database technology according to its potential applications in data mining and visualization of databases. A variety of high-efficient Skyline query approaches is proposed, such as BNL (Blocked Nested Loop), NN (Nearest Neighbour) and BBS (Branch and Bound Skyline). However, these methods always deal with exact values of properties of objects to get the results (the set of points satisfying the user's needs exactly), which can't be carried out with fuzzy information. Also high-performance can't be obtained with the increasing amounts and dimensions of knowledge. In order to solve this problem, this paper proposes the Skyline adaptive fuzzy query method based on the structure of R-trees and the BBS algorithm. It implements a fuzzy inference, and generates rapidly the possibility of getting appropriate results. Finally, in order to improve the accuracy of the reasoning process, Genetic Algorithms are used to study fuzzy rules automatically.

Keywords: Skyline Query, Fuzzy Control, Genetic Algorithms.

1 Introduction

Skyline query, namely Pareto, is to choose a subset from an objective set S in a declared n-dimension space, in which each point is independent of other points in the set; these points are called Skyline Point (SP) [1]. Because of its ability to depict the whole framework of massive data and multi-objective dataset and query the fixed data object, Skyline query is widely used in multi-standard decision support systems, city navigation systems, data mining and user preference queries.

At present, a large number of efficient Skyline query algorithms are proposed. The four major researches on Skyline query are uni-Skyline query method, multi-Skyline query method, Skyline query under different applications and the extension of Skyline query [2]. Especially for uni-Skyline query, there exists a variety of research achievements, in which Branch and Bound Skyline (BBS) is considered as one of the best uni-Skyline query algorithms.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 345–354, 2011.

There exist various Skyline query algorithms, most of them deal with accurate sub-attribute values and calculate the SP set the user prefers. However, during the query process, especially for massive data, it is difficult to obtain exact sub-attribute values, and because of that, the performance becomes much lower. Compared with an exact low-efficiency SP query, users prefer an approximate high-efficiency SP query. For example, Skyline query is usually used to retrieve P2P network neighbors by choosing as input two sub-attributes: the similarity of the message [1] and the hit rate of the node [2] [3]. It is difficult to compute their exact values because they vary with time. However, by considering approximate sub-attributes values, if we can obtain that the similarity of the message and the hit rate are high, then this node can be selected as one of the best objects to be transferred, based on the experience and the historical database. We need to make a decision rapidly to reduce the network delay, so it is more efficient for us to take a node as a candidate to be transferred according to pre-fixed thresholds instead of calculating the exact SP set.

This paper takes into account the mentioned works together, proposes the Skyline adaptive fuzzy query, which operates the Skyline query based on a fuzzy control mechanism. The remainder of the paper is organized as follows. Section 2 gives a brief introduction of the Skyline query and the adaptive fuzzy control. Section 3 describes the Skyline adaptive fuzzy query in detail. Section 4 analyzes the experiment results of our method. In Section 5, we conclude with a summary and outline some directions for future research.

2 The Background Research

2.1 The Skyline Query

Skyline query is a typical multi-objective optimization problem, which results in an algorithm with query complexity $O(nlog_2 n)$ for two- or three-dimensional data and with complexity $O(n(log_2 n)^{d-2})$ for more than three dimensions. It has been proposed by Kung in 1975 [4]. The first proposition to use Skyline with databases appears in [5]. During the last 10 years, research on Skyline query has developed Skyline query on integrated database, such as Branch and Bound Skyline (BBS) [6], Skyline query on distributed databases, designed according to the specific distributed circumstances, and other Skyline queries, for example, Skyline query on data stream.

The main problem of Skyline query is to deal with fuzzy requirements of users. With the high-dimensional and massive data and high performance users need, it is necessary for users to use fuzzy information in Skyline queries. Therefore, in this paper, according to the process of classical adaptive fuzzy control system, we will implement Skyline queries with fuzzy information.

[1] That is, if it is possible for a node to respond to the current retrieval if it has responded to similar retrievals before.

[2] The higher the hit rate of the node in the last period, the more possibly it hits again.

2.2 The Adaptive Fuzzy Control

Zadeh introduced fuzzy sets in 1965[7]. Compared with regular sets, fuzzy sets provide the transition between the complete membership and the complete exclusion, using a membership function ranked in the real interval $[0, 1]$. Any universe element is provided of a degree that represents its membership to the fuzzy set.

With the development of fuzzy sets, the adaptive fuzzy control, namely a self-organizing fuzzy control, is proposed. It can generate and modify fuzzy rules automatically to improve the performance of the system.

The process of the adaptive fuzzy control consists in five parts:

1. the fuzzy interface - a mapping from a fuzzy variable to a single value,
2. the fuzzy reasoning - to do fuzzy reasoning based on the input,
3. the de-fuzzy interface - a mapping from a fuzzy output to a accurate one,
4. the adaptive study of fuzzy rules - to implement the adaptive study process of the rules, and finally,
5. the knowledge base - includes database and rule base.

The architecture of the adaptive fuzzy control system is shown as Fig.1.

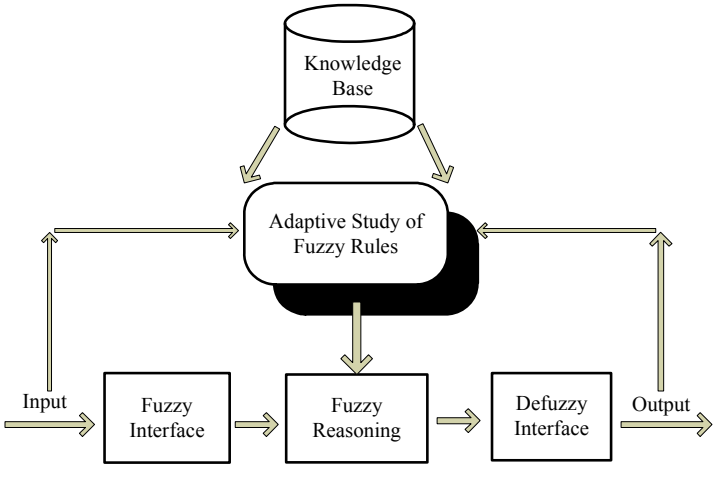


Fig. 1. The architecture of the adaptive fuzzy control system

3 The Skyline Adaptive Fuzzy Query

We will present the Skyline adaptive fuzzy query in the following three parts: Section 3.1 introduces the process of Skyline fuzzy reasoning in detail, and then Section 3.2 describes concretely steps of the adaptive study of fuzzy rules in Skyline query. Finally, Section 3.3 gives the pseudo-code of Skyline adaptive fuzzy query algorithm.

3.1 The Skyline Fuzzy Reasoning

The Skyline query estimates whether a single point is an SP according to the corresponding many-to-one mapping from the input space to the output space. We will use MISO [8] to establish our model. We assume that the input sub-attributes are n-dimensional, that is, $x_1, x_2 \cdots x_n$, and that output y represents the degree of membership in the SP set. Firstly, we implement a fuzzy partition on the input space. We assume that the sub-space which the i^{th} sub-attribute x_i corresponds to is partitioned into $A_{i1}, A_{i2} \cdots A_{iK_i}, i = 1, 2 \cdots n$, and therefore the n-dimensional input space is divided into $K_1 \times K_2 \cdots \times K_n$ fuzzy sub-spaces:

$$(A_{1j_1}, A_{2j_2} \cdots A_{nj_n})$$

$j_1 = 1, 2 \cdots K_1, j_2 = 1, 2 \cdots K_2 \cdots j_n = 1, 2 \cdots K_n$. For example, for a 2-dimensional input space, a fuzzy sub-space $(A_{1j_1}, A_{2j_2})$is shown in Fig.2.

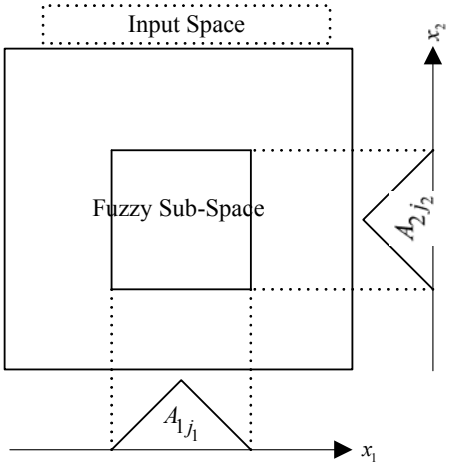


Fig. 2. A fuzzy sub-space of a 2-dimensional input space

In the process of fuzzy reasoning, users provide the degree of membership in the certain interval of each sub-attribute based on their experience, and then obtain the degree of membership of the SP set, according to the reasoning rules of the corresponding sub-space.

Given an input space indexed by R-Trees [9], which we partition to obtain a group of sub-spaces. Only the sub-space, which intersects with the Minimum Bounding Rectangle (MBR) [10] of the dataset, can correspond to a fuzzy rule. We use MBR to represent the dataset because of its simplicity of expression and its ease of use for searching. In contrast, for the sub-space which has no intersection with the MBR, we can denote $y = 0$ because it has no data. Fig.3 shows an example of the intersection in 2-dimensional input space.

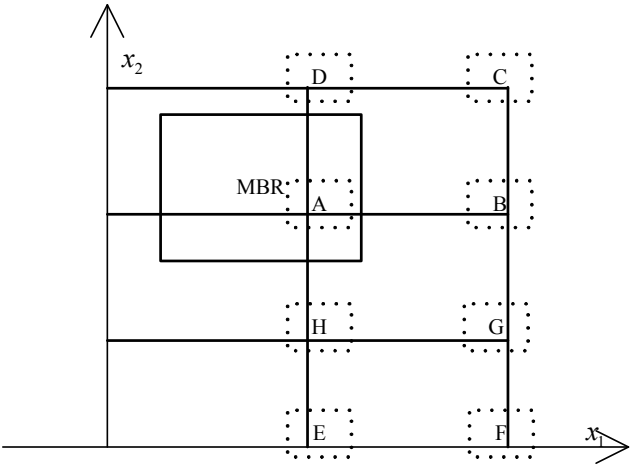


Fig. 3. An example of the intersection in 2-dimensional input space

Fuzzy rules have the following form:

$$R\text{: If } x_1 \text{ is } A_{1j_1} \text{ and } x_2 \text{ is } A_{2j_2} \text{ and } \cdots \text{ and } x_n \text{ is } A_{nj_n}$$
$$\text{Then } y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

Fuzzy rules vary the possibilities that points in different sub-spaces belong to the SP set with different coefficients $\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_n$, which are generated and modified through the adaptive study of fuzzy rules based on Genetic Algorithms as we will see in section 3.2.

3.2 The Adaptive Study of Fuzzy Rules in Skyline Query

As stated in 3.1, the adaptive study of fuzzy rules can be considered as the combinatory optimization problem in the input and output spaces, so we can use Genetic Algorithms, an efficient technology of optimization, to find the permutation and combination to classically adjust the coefficients and get the optimal fitness function value [11]. The detailed operations are as follows:

Step 1. The binary string expression of the output space.
A group of fuzzy rules are recorded as permutation and combination of α_i: $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$, with the fixed degrees of membership in certain interval of each sub-attribute $x_i, i = 1 \cdots n$. We transform $\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_n$ to their binary string expressions,$0 \leq \alpha_i \leq 1$,accurate to the second decimal place, and then multiple by tenfold to be a 7-stage binary string, for example, 0.65 is transformed to be 1000001. We also use a mapping from 127 to 100 to make sure

the corresponding α_i satisfies $0 \leq \alpha_i \leq 1$. The binary strings of $(n+1)\alpha_i$ join with each other to be a 7(n+1)-stage binary string γ_i. The population of γ_i is generated randomly to initialize the algorithm.

Step 2. The functions of Genetic operators.
There are mainly three operators in Genetic Algorithms: selection, crossover and mutation.

1. Selection. A sample can be replicated several times, and the occurrence time $P(\gamma_i)$ of each sample in the next generation is calculated according to its fitness function value:

$$P(\gamma_i) = \frac{f(\gamma_i)}{\sum_{j=1}^{m} f(\gamma_j)} \bullet m \ . \tag{1}$$

where m is the current population size. From this formula we can obtain that the higher the fitness function value of a sample, the more frequently it appears in the next generation.

2. Crossover. The probability of the crossover is 0.5-0.8, and we choose randomly the binary bits of two samples as the objects for this operation.

3. Mutation. The probability of the mutation is 0.1-0.2, and the sample and the binary bits of the mutation are selected randomly. The mutation operation can be simply considered as the inversion of a bit.

Step 3. The estimation of the fuzzy rules.
There are two kinds of points in this paper: sample points and test points. In this section,we use the sample points set to train the fuzzy rule and use the test points set to verify the accuracy of the obtained combinations of coefficients for the fuzzy rule in the experiment. For the generated fuzzy rules, we test and estimate them using their fitness function values. In accordance with the definition of SP and the algorithm BBS [12], we use the proportion of *mindist* of the sub-space in the sum of *mindist* of all the sub-spaces to indicate the possibility of containing an SP. The *mindist* of a data point is the sum of its abscissa and ordinate, while for a data rectangle it equals to the *mindist* of its left bottom data point.

Given a sub-space, we define the fitness function as follows:

$$f(\gamma_i) = 1 - |y(x_1, x_2, \cdots x_n) - \frac{mindist}{\sum_{j=1}^{c} mindist_j}| \ . \tag{2}$$

where $(x_1, x_2 \cdots x_n)$is a random data point in the sub-space, and it satisfies $y(x_1, x_2 \cdots x_n) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$, γ_i is the corresponding binary string of $\alpha_0^i, \alpha_1^i, \alpha_2^i, \cdots \alpha_n^i$, *mindist* corresponds to the current sub-space, and c is the number of the sub-spaces intersecting with the MBR.

From the fitness function, we can obtain that the closer the $y(x_1, x_2 \cdots x_n)$ to the possibility of the sub-space containing SP, the higher fitness function value $f(\gamma_i)$ and through this process, the coefficients $\alpha_0, \alpha_1, \alpha_2, \cdots \alpha_n$ are modified to better reflect the circumstance for special sub-space of containing SPs.

3.3 The Skyline Adaptive Fuzzy Query Algorithm

As stated above, according to the fuzzy reasoning, we estimate rapidly whether a fixed data point is an SP, and so we can go through all the data points of the MBR to find SPs with high performance. For each data point, we firstly fix the sub-space, which contains it, and then provide the degree of membership in the corresponding interval of each sub-attribute. The fuzzy reasoning process will be operated based on these values, and generate the possibility of the point being an SP. Finally, we compare this possibility with the threshold to decide whether the point should be an SP or not.

The description of the Skyline adaptive fuzzy query algorithm is shown in Fig. 4.

Algorithm *Skyline Adaptive Fuzzy Query*
Input: R-tree *t, Data d, MBR of the data set, e*
// t is the R-tree index of the input space, e is the threshold
Output: S // S is the SP set
1. S =∅ ; // Initialize S
2. For each d in MBR
3. Do Depth-First traverse on t, for each leaf-node l
4. if d ⊆ l ⋂ MBR
5. m = The Skyline Fuzzy Reasoning (d) using the fuzzy rules of l;
6. if m >e, insert d into S in descending order of m ;
7. Return S;
8. End

Fig. 4. The Skyline adaptive fuzzy query algorithm

4 Experiment Results

These simulations have been developed in Matlab version 6.5, on a Windows environment, taking an initial population of 5 genes and 2-dimensional sub-attribute space. We assume that the possibility of the crossover and the mutation are 0.6 and 0.2. Fig. 5 and Fig. 6 show the variation of the average and the maximum fitness function with 50 and 80 iterations respectively, in which $avg(f)$ is the average value of the fitness function while $max(f)$ is the maximum value of the fitness function in the current population. According to these figures, we can see that the value of the fitness function ($avg(f)$ and $max(f)$) increases gradually with the increasing number of iterations, that is, with the improvement of the population, the generated fuzzy rules can reflect the special sub-space more objectively.

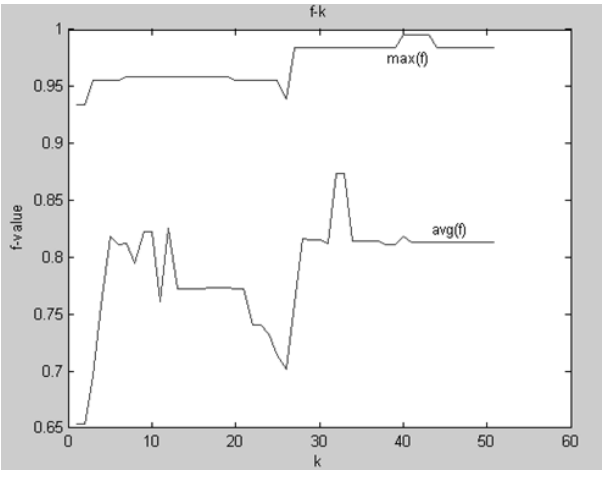


Fig. 5. The value of fitness function with 50 iterations

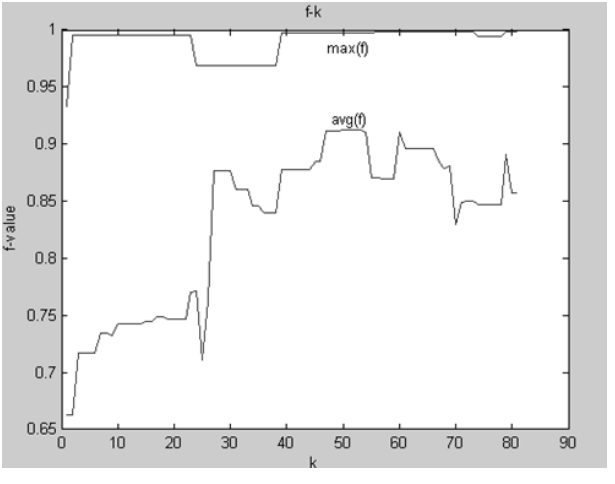


Fig. 6. The value of fitness function with 80 iterations

To compute the query error of the experiment, we use the ratio of the number of points satisfying the user's need to the total number of the obtained points. Fig. 7 and Fig. 8 show the query error with 40 and 80 iterations respectively, from which we can deduce that the query error reduces gradually with the increasing number of iterations and finally becomes stable in a certain period, that is, with the improvement of the population, the result can reflect the requirement of the user more accurately.

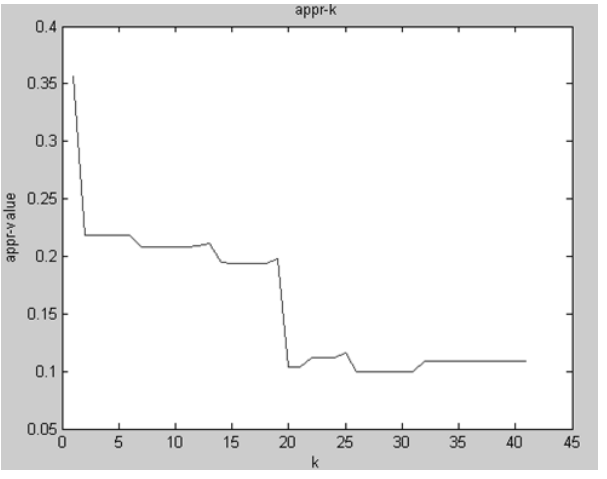


Fig. 7. The query error with 40 iterations

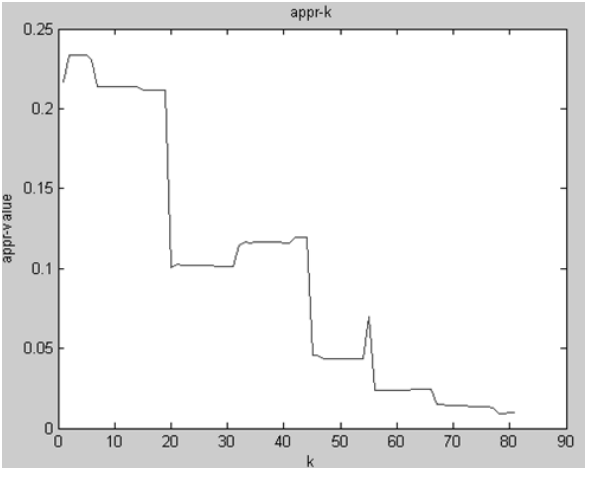


Fig. 8. The query error with 80 iterations

5 Conclusion

Skyline query is widely used in the research of database technology according to its potential applications in data mining and visualization of databases. Many high-efficient Skyline query approaches are proposed, such as BNL (Blocked Nested Loop), NN (Nearest Neighbour) and BBS (Branch and Bound Skyline). However, with the increasing amounts and dimensions of knowledge, it is difficult for Skyline query to obtain a high performance. In order to solve this problem, this paper proposes the Skyline adaptive fuzzy query method based on

the structure of R-trees and the BBS algorithm. It implements fuzzy inference, and generates rapidly the possibility of a point being a SP. Finally, in order to improve the accuracy of the reasoning process, Genetic Algorithms are used to study fuzzy rules automatically.

References

1. Chomicki, J., Godfrey, P., Gryz, J., Liang, D.: Skyline with Presorting: Theory and Optimization. In: Proceedings of the Intelligent Information Systems Conference (IIS): New Trends in Intelligent Information Processing and Web Mining, Advances in Soft Computing, pp. 593–602 (2005)
2. Wei, X.J., Yang, J., Li, C.P., Chen, H.: Skyline Query Processing. Journal of Software 19, 1386–1400 (2008)
3. Wang, X.S., Cui, X.W., Dong, L.G., Li, C.F.: P2P Intelligent Search Algorithm based on Skyline Query Technology. Computer Engineering 7, 122–124 (2009)
4. Kung, H.T., Luccio, F., Preparata, F.P.: On Finding the Maxima of a Set of Vectors. Journal of the ACM 22, 469–476 (1975)
5. Borzsonyi, S., Kossmann, D., Stocker, K.: The Skyline Operator. In: Proceedings of the ICDE, Germany, pp. 421– 430 (2001)
6. Zhu, L., Guan, J.H., Zhou, S.G.: Skyline Computation Survey. Computer Engineering and Applications 44, 160–165 (2008)
7. Zadeh, L.A.: Fuzzy Sets, Fuzzy Logic, Fuzzy Systems. World Scientific Press, Singapore (1996)
8. Ding, Y., Ying, H., Shao, S.: Necessary Conditions on Minimal System Configuration for General MISO Fuzzy Systems as Universal Approximators. In: Proceedings of 1997 IEEE SMC Conference (1997)
9. Guttman, A.: R-Trees: A Dynamic Index Structure for Spatial Searching. In: SIGMOD Conference, pp. 47–57 (1984)
10. Papadias, D., Tao, Y., Fu, G.: An Optimal and Progressive Algorithm for Skyline Queries. In: Proc of the ACM SIGMOD, pp. 467–478 (2003)
11. Zhang, L., Zhang, B.: Research on the Mechanism of Genetic Algorithms. Journal of Software 7, 945–952 (2000)
12. Papadias, D., Tao, Y., Fu, G., Seeger, B.: Progressive Skyline Computation in Database Systems. ACM Trans on Database Systems 30, 41–82 (2005)

Improved Travel Time Prediction Algorithms for Intelligent Transportation Systems

Nihad K. Chowdhury and Carson K.-S. Leung

Department of Computer Science, The University of Manitoba, Canada
kleung@cs.umanitoba.ca

Abstract. Travel time prediction provides commuters with useful information that enables them to decide whether or not to make necessary changes to their routes or departure times. This explains why travel time prediction has become important to intelligent systems, especially intelligent transportation systems (ITS). Over the past few years, several algorithms have been developed to predict travel time, but some of them suffer from a few problems. In this paper, we propose algorithms that solve these problems and improve the performance and/or accuracy of travel time prediction for ITS.

Keywords: Knowledge-based and intelligent information & engineering systems (KES), intelligent system, intelligent transportation system (ITS), knowledge discovery and data mining, travel time prediction.

1 Introduction

As travel time prediction is crucial to numerous intelligent transportation system (ITS) applications such as dynamic route guidance and trip planning, it plays a key role in ITS. With the development of advanced traveller information systems, travel time prediction has become increasingly important as it can assist commuters to better adjust their travel schedules. Moreover, accurate and reliable prediction of travel time on road networks is also vital for many dynamic route guidance systems because such prediction helps commuters to decide whether or not they need to change the starting time of their routes or even cancel their trip. Furthermore, accurate travel time prediction enables the generation of the fastest path (cf. shortest path) between the origin and the destination. As time-varying features of traffic flow (e.g., daily features that distinguish the peak-hour traffic from off-peak hour one) can influence the estimate of accurate travel time, how to provide reliable travel time information has drawn attention to many researchers [4,7,11,12,16].

To predict travel time, various techniques can be applied. These include time series data analysis as well as data mining techniques. Data mining aims to search for implicit, previously unknown, and potentially useful knowledge from data. Common data mining techniques help discover frequently occurring patterns (e.g., popular routes taken by commuters) and detect anomalies (e.g., abnormally busy travel time). Besides finding frequent patterns [9] and detecting

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 355–365, 2011.

anomalies [10], other data mining techniques can be applicable to travel time prediction. For instance, classification techniques [6] can be applied to learn historical data and make reliable prediction for future data. Similarly, clustering techniques [13] can be applied to group similar data into the same class so that prediction can be made based on a class of similar data (than the entire dataset).

Over the past few years, various travel time prediction algorithms have been proposed [3,8,15]. For example, a classification algorithm called NBC [6] was proposed in KES 2008 to predict travel time. Two algorithms called SMA and CA [2] were developed in KES 2009 based on the concept of moving average, and they were shown to give more accurate prediction. In KES 2010, a clustering algorithm called MKC [13] was shown to return more accurate predicted travel times than those predicted by SMA and CA [2] as well as NBC [6]. *Along this direction, can we further improve accuracy of the predicted travel time?* Moreover, a careful analysis on these algorithms reveals that some of them (e.g., SMA and CA [2], MKC [13]) suffer from a few problems. *Can we solve these problems?*

In this paper, our ***key contributions*** are to propose several travel time prediction algorithms that further improve the accuracy and/or performance of the existing algorithms. Specifically, we propose fast and accurate algorithms to improve travel time prediction for ITS.

The remainder of this paper is organized as follows. The next section provides readers with background and related work on ITS. In Section 3, we propose our algorithms that improve travel time prediction for ITS. Experimental results are shown in Section 4. Finally, Section 5 gives conclusions.

2 Background and Related Work

A road network of a town or city usually consists of multiple road segments. A segment or *link* connects two locations in the network. For a commuter to go from a location A to another location Z, his trajectory or *path* consists of one or more road segments. The end location of a segment may become the start location of another segment (e.g., the path from A to Z may consist of two direct segments: one from A to L and another one from L to Z).

Over the past decade, travel time prediction has drawn attention of many researchers. Numerous works have been proposed to accurately predict travel time. For instance, van Lint et al. [17] used a *nonlinear* state-space neural network based approach to predict the travel time. Conversely, some other researchers focused on *linear* travel time predictors. As an example, Kwon et al. [5] applied the linear regression method to estimate travel time on freeways based on the flow and occupancy data (measured by loop detectors) as well as on historical travel time information. Wu et al. [18] proposed a link-based method, which predicts the travel time for a path by the summing the predicted travel times for each link in the path. To predict the travel time for each link, the method was trained on a specific route. It ignores other alternative routes in road networks. Moreover, such a method requires much memory space and runtime. Schmitt and Jula [14] investigated a switching model, which consists of two linear travel

time predictors. Their switching model focuses on a small predefined set of popular routes (instead of all available routes in road networks). Moreover, finding the switching point can be computational intensive.

Lee et al. [6] proposed the Naïve Bayesian Classification (NBC) algorithm, which uses the data mining technique of naïve Bayesian classification to predict travel time for each velocity class. NBC was trained using all routes available in the road network. Their experimental results showed that NBC led to more accurate predicted travel times than those predicted by the linear regression method [5] or the link-based method [18]. As an extension to NBC, a rule-based Bayesian classification (RBC) algorithm [1] was also proposed. Although the decision rules used in RBC represent the knowledge (in the form of rules) in a more intuitive way, RBC was shown to give less accurate prediction than NBC.

2.1 Successive Moving Average (SMA) and Chain Average (CA) Algorithms

In KES 2009, two travel time prediction algorithms—namely, Successive Moving Average (SMA) and Chain Average (CA) [2]—were developed based on the concept of moving average. Given n historical data (i.e., n travel times $t_1, ..., t_n$) for a given road segment S and time period T (e.g., Segment 1 during 8:00–9:00), both SMA and CA use recursive formulas to predict the travel time for S and T. They do so by first constructing a two-dimensional $n \times n$ matrix M and then returning the value of a specific entry in M as the predicted time. Specifically, SMA constructs the matrix M in such a way that each diagonal entry $M_{i,i}$ (for $1 \le i \le n$) represents each of the n historical travel times (i.e., $M_{i,i} = t_i$ where $t_i \in \{t_1, ..., t_n\}$) and each non-diagonal entry $M_{i,j}$ is the average of all entries in i-th row and j-th column:

$$\begin{cases} M_{i,i} = t_i & \text{for } 1 \le i \le n, \text{ and} \\ M_{i,j} = \frac{\sum_{k=i}^{j-1}(M_{i,k}+M_{k+1,j})}{2(j-i)} & \text{for } 1 \le i < j \le n. \end{cases} \quad (1)$$

Once the matrix M is constructed, SMA then returns the value of $M_{1,n}$ as the predicted travel time $\hat{t}$ for the given S and T. See Example 1. Experimental results [2] showed that SMA led to more accurate predicted travel times than those predicted by NBC [6] and the switching model [14].

Example 1. For illustrative purpose, let us consider n=4 historical travel times (say, 8, 12, 12 & 15 mins) collected for a 6 km road segment during 8:00–9:00. SMA constructs the following 4×4 matrix with 10 non-empty entries:

$$M = \begin{pmatrix} t_1 & \frac{M_{1,1}+M_{2,2}}{2} & \frac{M_{1,1}+M_{1,2}+M_{2,3}+M_{3,3}}{4} & \frac{\boldsymbol{M_{1,1}+M_{1,2}+M_{1,3}+M_{2,4}+M_{3,4}+M_{4,4}}}{\mathbf{6}} \\ & t_2 & \frac{M_{2,2}+M_{3,3}}{2} & \frac{M_{2,2}+M_{2,3}+M_{3,4}+M_{4,4}}{4} \\ & & t_3 & \frac{M_{3,3}+M_{4,4}}{2} \\ & & & t_4 \end{pmatrix}$$

and returns the value of $M_{1,n}$=11.6875 mins as the predicted travel time $\hat{t}$. Note that the computation of $M_{1,n}$ directly depends on six entries (i.e., $M_{1,1}, M_{1,2},$

$M_{1,3}, M_{2,4}, M_{3,4}$ & $M_{4,4}$), which in turn depend on the remaining three entries (e.g., $M_{1,3}$ partially depends on $M_{2,3}$ & $M_{3,3}$). In other words, the computation of $M_{1,n}$ directly or indirectly depends on *all* $10 - 1 = 9$ entries. Most of these entries are visited multiple times (e.g., $M_{2,2}$ is used by $M_{1,2}, M_{2,3}$ & $M_{2,4}$). □

Similarly, CA also constructs the matrix M in such a way that each diagonal entry $M_{i,i}$ (for $1 \leq i \leq n$) represents one of the n historical travel times (i.e., $M_{i,i} = t_i$ where $t_i \in \{t_1, ..., t_n\}$), but each non-diagonal entry $M_{i,j}$ is the average of its two neighbouring entries:

$$\begin{cases} M_{i,i} = t_i & \text{for } 1 \leq i \leq n, \\ M_{1,j} = \frac{M_{1,j-1}+M_{2,j}}{2} & \text{for } 2 \leq j \leq n, \text{ and} \\ M_{i,j} = \frac{M_{i-1,j-1}+M_{i+1,j}}{2} & \text{for } 2 \leq i < j \leq n. \end{cases} \quad (2)$$

Again, once the matrix M is constructed, CA returns the value of $M_{1,n}$ as the predicted travel time $\hat{t}$ for the given S and T. See Example 2. Experimental results [2] showed that, between SMA and CA, the latter led to more accurate predicted travel times than the former.

Example 2. Using the same collection of n=4 historical travel times as in Example 1, CA constructs the following 4×4 matrix with 10 non-empty entries:

$$M = \begin{pmatrix} t_1 & \frac{M_{1,1}+M_{2,2}}{2} & \frac{M_{1,2}+M_{2,3}}{2} & \boldsymbol{\frac{M_{1,3}+M_{2,4}}{2}} \\ & t_2 & \frac{M_{1,2}+M_{3,3}}{2} & \frac{M_{1,3}+M_{3,4}}{2} \\ & & t_3 & \frac{M_{2,3}+M_{4,4}}{2} \\ & & & t_4 \end{pmatrix}$$

and returns the value of $M_{1,n}$=11.125 mins as the predicted travel time $\hat{t}$. Note that the computation of $M_{1,n}$ directly depends on two entries (i.e., $M_{1,3}$ & $M_{2,4}$), each of which recursively depends on two other entries (e.g., $M_{1,3}$ depends on $M_{1,2}$ & $M_{2,3}$). Recursively, the computation of $M_{1,n}$ directly or indirectly depends on the values of *all* $10 - 1 = 9$ entries. Among these entries, six of them (in either the diagonal or the last column) are used once and the remaining three are used twice (e.g., $M_{2,3}$ is used by $M_{1,3}$ & $M_{3,4}$) for computing $M_{1,n}$. □

2.2 The Modified k-Means Clustering (MKC) Algorithm

In KES 2010, Nath et al. [13] developed the modified k-means clustering (MKC) algorithm. Given n historical data (i.e., n travel times $t_1, ..., t_n$) for a given road segment S and time period T, MKC first partitions these n data into k=2 clusters using the data mining technique of k-means clustering such that data within each cluster are similar. Dissimilarity between two unique travel times t_i and t_j is measured based on their Manhattan distance in a three-dimensional space consisting of travel times, their frequencies (denoted as f_i & f_j) and velocities:

$$dissimilarity(t_i, t_j) = |t_i - t_j| + |f_i - f_j| + |velocity(t_i) - velocity(t_j)|. \quad (3)$$

MKC then computes the mean of travel times for each cluster. The predicted travel time $\hat{t}$ for the given S and T is the average of the means. See Example 3.

Experimental results [13] showed that MKC led to more accurate predicted travel times than CA [2].

Example 3. Using the same collection of n=4 historical travel times as in Example 1, MKC clusters these three unique travel times (as t_2=t_3=12 mins) using Equation (3). For example, $dissimilarity(t_1, t_2)$ = 4 mins + 1 + 15 km/h = 20 because $velocity(t_1) = \frac{6\,\text{km}}{t_1}$ = 45 km/h and $velocity(t_2) = \frac{6\,\text{km}}{t_2}$ = 30 km/h. At the end, MKC returns the predicted time $\hat{t} = \left(8 + \frac{12+15}{2}\right)/2 = 10.75$ mins. □

3 Our Improved Algorithms for Predicting Travel Time

In this section, we examine the aforementioned SMA, CA and MKC algorithms, and discover their associated problems. Then, we propose several algorithms that solve these problems and further improve accuracy and/or performance of travel time prediction.

3.1 Improved Successive Moving Average (iSMA) and Improved Chain Average (iCA) Algorithms

As illustrated in Examples 1 and 2, both SMA and CA suffer from the following problems.

Problem 1. The amount of space required by SMA or CA is *exponential* to the number of historic travel times. Given n historical travel times, SMA and CA require the construction of an $n \times n$ matrix (or more precisely, an $n \times n$ upper triangular matrix with $\frac{n(n+1)}{2}$ non-empty entries).

Problem 2. SMA cannot compute the predicted time $\hat{t} = M_{1,n}$ directly. Due to the recursive nature of Equation (1), the computation of $M_{1,n}$ directly or indirectly depends on the values of *all* other $\frac{n(n+1)}{2} - 1$ entries in M. This problem applies to CA as well. For Equation (2), the computation of $M_{1,n}$ may appear to (directly) depend on only two entries $M_{1,n-1}$ and $M_{2,n}$, but the computation of each of these two entries depends on another two entries in M. Hence, the computation of $M_{1,n}$ by CA also requires the values of *all* other $\frac{n(n+1)}{2} - 1$ entries.

Solution. To solve the above problems, we first propose the *Improved Chain Average* (*iCA*) algorithm. Observing that the root cause of the problems with CA was the recursive nature of Equation (2), our iCA transforms the recursive equation into a non-recursive one by expanding the terms in the equation. Consequently, given n historical data (i.e., n travel times), the predicted travel time $\hat{t}$ can be expressed in the form of a weighted sum of n travel times:

$$\hat{t} = \sum_{i=1}^{n} w_i \; t_i, \tag{4}$$

where w_i is the weight. By doing so, our iCA returns the same predicted time $\hat{t}$ as CA, but requires less space and time than CA. Note that, due to the simplicity and non-recursive nature of Equation (4), our iCA does not require the

construction of matrix M. Instead, iCA requires only n travel times and their corresponding weights. As such, the amount of space required by iCA is *linear* (than exponential) to n. Moreover, iCA *directly* computes $\hat{t}$ as a sum of n products (of w_i t_i for $1 \leq i \leq n$). As each weight w_i is independent of the value of t_i, the weights can be precomputed once and used multiple times for different collections of n historical data. Furthermore, for each collection of n data, every t_i is used only *once* during the computation of $\hat{t}$. Hence, our iCA greatly reduces the runtime, and it solves Problems 1 & 2. See Example 4.

Example 4. Let us revisit Example 2. Our iCA directly computes the predicted travel time $\hat{t}$ as a weighted sum of n historical travel times: $\hat{t} = \frac{5}{16}t_1 + \frac{5}{16}t_2 + \frac{1}{4}t_3 + \frac{1}{8}t_4 = 11.125$ mins, which is identical to $\hat{t}$ returned by CA. □

When compared with CA, the SMA algorithm suffers the following problem in addition to Problems 1 & 2.

Problem 3. During the computation of $\hat{t}$, most of the entries in M are visited by SMA *multiple* times. An entry $M_{i,j}$ is visited by SMA $(i-1)+(n-j)$ times. (In contrast, CA visits each diagonal entry $M_{i,i}$ & entry in the last column $M_{i,n}$ once, and it visits all other entries in M twice.)

Solution. Following the same logic, we propose the *Improved Successive Moving Average* (*iSMA*) algorithm, which captures Equation (1) but in a simple non-recursive form as in Equation (4). As such, iSMA solves Problems 1–3. It greatly reduces both space and runtime while maintaining the accuracy of $\hat{t}$. See Example 5.

Example 5. Let us revisit Example 1. Our iSMA solves Problems 1–3 as it does not construct M. Instead, to reduce runtime, it *directly* computes $\hat{t}$ as a weighted sum of n historical travel times: $\hat{t} = \frac{5}{16}t_1 + \frac{3}{16}t_2 + \frac{3}{16}t_3 + \frac{5}{16}t_4 = 11.6875$ mins, which is identical to $\hat{t}$ returned by SMA. During the computation, our iSMA uses each t_i only *once* (than multiple times). The required space is *linear* (than exponential) to n. □

3.2 Improved Modified k-Means Clustering (iMKC) Algorithms

As illustrated in Example 3, MKC suffers from the several problems as follows.

Problem 4. The dissimilarity measure was not applied to a normalized three-dimensional space. Instead, it sums the difference between two historical data t_i and t_j in each of the three dimensions. Unfortunately, different units were used for each dimension (e.g., difference in travel times were measured in minutes, difference in velocities were measured in km/h).

Problem 5. Two of the three dimensions—namely, travel time & velocity—used in dissimilarity measure are correlated. For a given road segment (i.e., with a fixed distance), travel time is inversely proportional to its velocity because travel

time × velocity = distance of the road segment. Hence, it becomes redundant to keep both dimensions.

Solution. To solve Problems 4 & 5, we propose our third algorithm. Observing the problems associated with the dissimilarity measure (Equation (3)) used in MKC, our proposed *Improved Modified k-Means Clustering* (*iMKC*) algorithm uses a different but more appropriate dissimilarity measure. Specifically, our iMKC measures the dissimilarity based only on the difference between two unique travel times t_i and t_j:

$$dissimilarity(t_i, t_j) = |t_i - t_j|. \tag{5}$$

By doing so, our iMKC solves Problems 4 & 5 because it avoids measuring two correlated dimensions and it does not sum the differences that are measured in different units. Moreover, it reduces the dimension of clustering space from three to one. See Example 6. As a preview, experimental results in Section 4 show that our iMKC led to more accurate prediction than MKC.

Example 6. Let us revisit Example 3. Our iMKC clusters the three unique travel times using Equation (5), which measures dissimilarity in one-dimensional space. For example, $dissimilarity(t_1, t_2) = 4$ mins. □

Enhancement 1. While our iMKC solves Problems 4 & 5, it can be further enhanced. For instance, observant readers may notice that the dissimilarity measure in Equation (5) only measures the pairwise difference in travel times. However, given a road segment S and time period T, it is not unusual to have duplicate travel times in a collection of n historical travel times. We better capture the frequency of these travel times. Hence, our first enhancement to iMKC is to capture this frequency information. Instead of having an additional dimension and running into potential problem of summing quantities that are measured in different units, we weight the travel times t_i & t_j by their corresponding frequencies f_i & f_j:

$$dissimilarity(t_i, t_j) = |f_i\ t_i - f_j\ t_j|. \tag{6}$$

As a preview, experimental results in Section 4 show that iMKC using this enhancement of *weighted* dissimilarity measure (Equation (6)) led to more accurate prediction than iMKC using the unweighted one in Equation (5).

Enhancement 2. While the first enhancement focuses on the dissimilarity measure, our second enhancement focuses on clustering techniques. As *k-medoid clustering* is often more robust to noise and outliers than k-means clustering, our second enhancement to iMKC is to use k-medoid clustering. The key difference between k-means and k-medoid clustering is that clusters for the former are represented by centroids (i.e., centers of all data within a cluster), whereas clusters for the latter are represented by medoids (i.e., most centrally located datum for each cluster). As iMKC uses iterative refinement for clustering, it requires kn comparisons to assign n travel times into k clusters in each iteration. Although it may appear that n extra comparisons are needed for finding

k medoids, it is quite common that the k medoids remain unchanged while the k centroids were changed from an iteration to another. Thus, as a bonus, the use of k-medoid may reduce the number of iterations. As a preview, experimental results in Section 4 show that iMKC using this enhancement of k-medoid clustering gave more accurate predicted travel times (and more efficiently) than iMKC using k-means clustering.

4 Experimental Evaluation

To evaluate our proposed algorithms, we used a dataset of 2,731,594 travel times produced by a trajectory simulator. These data reflect real traffic patterns of 167,669 trajectories of about 5,000 vehicles collected through their global positioning system (GPS) sensors during weekdays and weekends as well as peak and off-peak hours over a period of one year covering 1,757 unique road segments. The data were divided according to a 12:1 ratio for training and testing purposes. The training data were then partitioned based on their road segments and time periods (e.g., $\langle S$=Segment 1, T=8:00–9:00$\rangle$ partition). For experiments, we arbitrarily picked 8 sets of test data, which represent 8 different time periods in a day. Each test dataset covers a path consisting of about 18 road segments on average. The predicted travel time for a path is the sum of travel time predicted for every road segment in the path.

In the experiments, we compared our five proposed travel time prediction algorithms (i.e., iSMA, iCA, and iMKC & its two enhancements) and three existing algorithms (SMA [2], CA [2] and MKC [13]). All these algorithms were programmed in C++. Experiments were run on a Windows PC with Intel Core i3 processor, 2.26 GHz CPU and 2 GB of RAM.

First, we compared two pairs of algorithms: (i) SMA [2] with our iSMA and (ii) CA [2] with our iCA. Since algorithms within each pair return the same predicted travel time $\hat{t}$, we measured their runtimes (i.e., CPU & I/O times). As shown in Table 1, our proposed iSMA and iCA were much faster than SMA and CA. The reason is that our algorithms directly compute $\hat{t}$ for each road segment

Table 1. Runtimes (in milliseconds)

	SMA [2]	Our iSMA	CA [2]	Our iCA
Path 1	25.34ms	2.29ms	15.97ms	2.28ms
Path 2	46.52ms	1.34ms	5.00ms	1.34ms
Path 3	287.17ms	1.74ms	16.52ms	1.74ms
Path 4	1,637.24ms	3.29ms	29.23ms	3.29ms
Path 5	318.38ms	2.65ms	19.59ms	2.64ms
Path 6	1,133.07ms	2.81ms	41.10ms	2.80ms
Path 7	359.51ms	2.04ms	17.42ms	2.04ms
Path 8	703.30ms	1.96ms	26.93ms	1.96ms

Table 2. Mean absolute relative error (MARE)

	MKC [13]	iMKC	iMKC (w/ weighted time)	iMKC (w/ k-medoid)
Path 1	0.75	0.51	0.51	0.35
Path 2	0.76	0.47	0.38	0.47
Path 3	0.73	0.56	0.50	0.48
Path 4	0.63	0.54	0.42	0.54
Path 5	0.74	0.45	0.39	0.43
Path 6	1.86	1.02	0.94	0.69
Path 7	0.76	0.57	0.46	0.54
Path 8	0.77	0.58	0.37	0.40

in a path as a weighted sum of n travel times (e.g., n=961 for the first segment of Path 1). As iSMA and iCA are only differed by the weights used in the sum, their runtimes did not vary too much. In contrast, SMA and CA construct an $n \times n$ matrix M, compute $O(n^2)$ entries in M, and visit many entries in M multiple times. Between SMA and CA, computation of each $M_{i,j}$ for CA requires lookup of two entries, whereas that for SMA usually requires lookup of more than two entries.

Next, we compared MKC [13] with three variants of our iMKC. To evaluate the accuracy of prediction, we measured the mean absolute relative error (MARE), which sums (over all N segments in a path) the relative difference between the actual travel time t_{S_i} (based on test data) and predicted travel time $\widehat{t}_{S_i}$ (based on historical data) for every road segment S_i in a path:

$$\text{MARE} = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{t_{S_i} - \widehat{t}_{S_i}}{t_{S_i}} \right|. \tag{7}$$

For example, MARE of iMKC on Path 1 consisting of N=17 segments (in which the first segment contains $n = 961$ historical travel times) was 0.51, which was lower than that of MKC (i.e., 0.75). The lower the MARE, the more accurate was the prediction. Results in Table 2 showed that, as our iMKC used an appropriate dissimilarity measure (Equation (5)), it gave more accurate prediction (i.e., lower MARE) than MKC. Regarding SMA and CA [2], they gave higher MAREs than MKC [13] (e.g., MAREs of SMA, CA & MKC for Path 1 were 1.06, 1.04 & 0.63, respectively). Transitively, iMKC led to more accurate predicted travel times (i.e., lower MAREs) than iSMA and iCA (which gave the same MAREs as SMA and CA, respectively). As for the two enhancements to our iMKC (i.e., weighted the travel times by their frequency in the dissimilarity measure, k-medoid clustering), they further improved iMKC as evidenced by the low MAREs in the last two columns of Table 2. For example, for Path 1, iMKC (w/ k-medoid) led to a low MARE of 0.35. In terms of runtimes, it predicted travel times in 51.73ms (cf. 62.86ms by iMKC without enhancement).

5 Conclusions

In this paper, we analyzed three existing algorithms (i.e., CA, SMA, MKC), revealed their problems, and proposed five algorithms—namely, iCA, iSMA, iMKC and its two enhancements—as solutions to solve these problems and to improve travel time prediction for ITS. Our proposed iCA and iSMA algorithms use non-recursive equations to directly compute the predicted travel time. As such, they improve the travel time prediction process by reducing both time and space requirements while maintaining accuracy of the prediction. Our iMKC algorithm lowers the dimension of clustering space and measures dissimilarity in a single dimension of travel times. As such, it improves accuracy and performance of the prediction. The two enhancements further improve iMKC by capturing the frequency of travel times in the dissimilarity measure and/or using k-medoid clustering. Experimental results showed that all our five proposed algorithms improved travel time prediction for ITS.

Acknowledgements. This project is partially supported by NSERC (Canada).

References

1. Chang, J., Chowdhury, N.K., Lee, H.: New travel time prediction algorithms for intelligent transportation systems. J. Intell. Fuzzy Syst. 21(1–2), 5–7 (2010)
2. Chowdhury, N.K., Nath, R.P.D., Lee, H., Chang, J.: Development of an effective travel time prediction method using modified moving average approach. In: Velásquez, J.D., Ríos, S.A., Howlett, R.J., Jain, L.C. (eds.) KES 2009, Part I. LNCS (LNAI), vol. 5711, pp. 130–138. Springer, Heidelberg (2009)
3. Idé, T., Kato, S.: Travel-time prediction using Gaussian process regression: a trajectory-based approach. In: SDM 2009, pp. 1183–1194. SIAM, Philadelphia (2009)
4. Kriegel, H.-P., Renz, M., Schubert, M., Züfle, A.: Statistical density prediction in traffic networks. In: SDM 2008, pp. 692–703. SIAM, Philadelphia (2008)
5. Kwon, J., Coifman, B., Bickel, P.: Day-to-day travel time trends and travel time prediction from loop detector data. TRR Journal 1717, 120–129 (2000)
6. Lee, H., Chowdhury, N.K., Chang, J.: A new travel time prediction method for intelligent transportation system. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part I. LNCS (LNAI), vol. 5177, pp. 473–483. Springer, Heidelberg (2008)
7. Lee, J.-G., Han, J., Whang, K.-Y.: Trajectory clustering: a partition-and-group framework. In: ACM SIGMOD 2007, pp. 593–604. ACM, New York (2007)
8. Lee, W.-H., Tseng, S.-S., Tsai, S.-H.: A knowledge based real-time travel time prediction system for urban network. Expert Systems with Applications 36(3), Part 1, 4239–4247 (2009)
9. Leung, C.K.-S., Brajczuk, D.A.: uCFS$_2$: an enhanced system that mines uncertain data for constrained frequent sets. In: IDEAS 2010, pp. 32–37. ACM, New York (2010)
10. Leung, C.K.-S., Mateo, M.A.F., Nadler, A.J.: CAMEL: an intelligent computational model for agro-meteorological data. In: ICMLC 2007, vol. 4, pp. 1960–1965. IEEE, Piscataway (2007)

11. Liu, W., Wang, Z., Feng, J.: Continuous clustering of moving objects in spatial networks. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 543–550. Springer, Heidelberg (2008)
12. Nakata, T., Takeuchi, J.: Mining traffic data from probe-car system for travel time prediction. In: ACM KDD 2004, pp. 817–822. ACM, New York (2004)
13. Nath, R.P.D., Lee, H.-J., Chowdhury, N.K., Chang, J.-W.: Modified k-means clustering for travel time prediction based on historical traffic data. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part I. LNCS (LNAI), vol. 6276, pp. 511–521. Springer, Heidelberg (2010)
14. Schmitt, E.J., Jula, H.: On the limitations of linear models in predicting travel times. In: IEEE ITSC 2007, pp. 830–835. IEEE, Piscataway (2007)
15. Simroth, A., Zähle, H.: Travel time prediction using floating car data applied to logistics planning. IEEE Trans. Intell. Transp. Syst. 12(1), 243–253 (2011)
16. Takamiya, M., Yamamoto, K., Watanabe, T.: Probabilistic estimation of travel behaviors using zone characteristics. In: Velásquez, J.D., Ríos, S.A., Howlett, R.J., Jain, L.C. (eds.) KES 2009, Part II. LNCS (LNAI), vol. 5712, pp. 615–622. Springer, Heidelberg (2009)
17. van Lint, J.W.C., Hoogendoorn, S.P., van Zuylen, H.J.: Accurate freeway travel time prediction with state-space neural networks under missing data. Transportation Research - Part C: Emerging Technologies 13(5–6), 347–369 (2005)
18. Wu, C.-H., Ho, J.-M., Lee, D.T.: Travel-time prediction with support vector regression. IEEE Trans. Intell. Transp. Syst. 5(4), 276–281 (2004)

An Effective Utilization of Many Neural Networks for Improving the Traditional Technical Analysis in the Stock Market

Norio Baba[1], Kokutan Liu[1], Lee Chen Han[1], Takao Mitsuda[1], Kou Ro[1], and Kou Ninn[2]

[1] Information Science, Osaka Kyoiku University, Asahiga-Oka, 4-698-1, Kashiwara City, Osaka Prefecture, 582-8582, Japan
[2] Toshiba Solution Company, Minatoku, Shibaura 1-1-1, Tokyo, Japan

Abstract. In this paper, we propose a new decision support system for dealing stocks which utilizes information regarding the predictions obtained by NNs concerning the occurrence of the "Golden Cross (GC)" and "Dead Cross (DC)", those (also obtained by NNs) concerning the rate of change of the future stock price several weeks ahead, and that (also obtained by NNs) concerning the relative position of the stock price versus "GC" and "DC". Computer simulation results concerning the dealings of the Nikkei-225 for the last 16 years confirm the effectiveness of our approach.

Keywords: neural networks, traditional technical analysis, prediction of golden cross (GC) and dead cross (DC), Nikkei-225, improved DSS for dealing stocks.

1 Introduction

In recent years, soft computing techniques such as NNs, GAs, and etc. have been studied quite extensively by many researchers and have been applied successfully to various real world problems [1]-[8].

In this paper, we propose a new DSS for dealing stocks which utilizes intelligently predictions (obtained by NNs) concerning the occurrence of Golden Cross (GC) & Dead Cross (DC), those concerning increase (decrease) rate of future stock price several weeks ahead, and that concerning the relative position of the stock price versus crossing point of GC (DC). The outline of this paper is as follows. In the following section, we shall briefly touch upon the traditional technical analysis which predicts future tendency of the stock price by taking the relative relationship between "long term moving average" and "short term moving average" into account. Then, we shall mention that a DSS relying upon only the prediction concerning the occurrence of GC&DC becomes sometimes unreliable. In Section 3, we shall propose a new DSS which utilizes not only the predictions concerning the occurrence of GC & DC, but also predictions concerning the rate of change of the future stock price several weeks ahead and those concerning the relative position of the stock price

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 366–371, 2011.

versus predicted crossing point of GC (DC). In Section 4, computer simulation results concerning the dealings of the Nikkei-225 during the last 16 years which confirm the effectiveness of our approach will be given.

2 Traditional Technical Analysis in the Stock Market

In order to detect the current trend of a stock price, many stock traders have often relied upon the traditional technical analysis which takes the relative relationship between "the Long Term Moving Average (LTMA) " and "the Short Term Moving Average (STMA)" into account. They have believed that the Golden Cross (Dead Cross) which STMA cuts LTMA upwards(downwards) gives a strong sign that suggests the upward (downward) moving of the future stock price.

However, recently, we have noticed that GC & DC are not always reliable in making a forecast of future movement of a stock price and their reliability depends strongly upon the relative changes of the STMA & LTMA near the crosses. Further, we have also noticed that their reliability is also strongly influenced by the relative position of the stock price versus the crossing point. In the following section, we shall propose a new DSS which utilizes predictions (given by NNs) concerning the occurrence of GC (DC), those concerning rate of change of the future stock price several weeks ahead, and that concerning the relative position of the future stock price versus crossing point of the two moving averages.

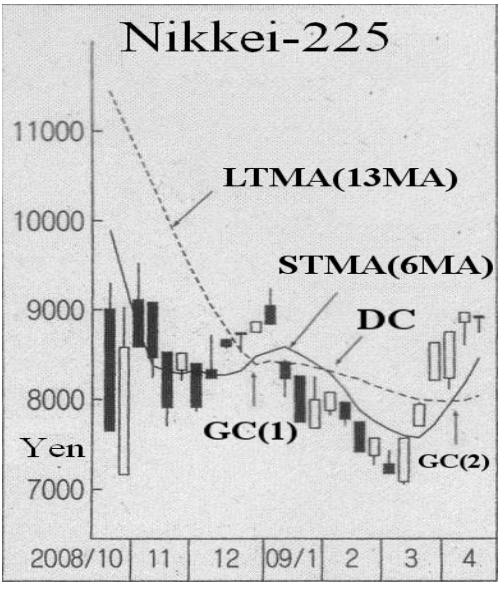


Fig. 1. An Example of the Movements of the LTMA and the STMA Near the Golden Crosses of the Nikkei-225

Remark 2.1: Fig.1 shows the changes of the Nikkei-225 during each week (candle stick), changes of the short term moving average(STMA; solid line; 6 weeks moving average; from October 2008 to April 2009), and changes of the long term moving average (LTMA; dotted line; 13 weeks moving average; from October 2008 to April 2009).

The long term moving average (LTMA) of the Nikkei-225 near the GC(1) moves downwards. On the other hand, the short term moving average (STMA) of the Nikkei-225 near the GC(1) moves upwards. Further, Nikkei-225 at the week when golden cross occurred is not far above the crossing point of GC(1). The upward movement of the Nikkei-225 after GC(1) was not particularly big.

On the other hand, the situation near the GC(2) is quite different from that near GC(1). The LTMA moves upwards. Further, the upward movement of the STMA is remarkable. Also, changes of the Nikkei-225 occurred far above the crossing point of GC(2). We can easily observe that the Nikkei-225 has gone up considerably after the Golden Cross GC(2).

3 A New Decision Support System (NDSS) for Dealing Stocks Which Utilizes Three Kinds of Predictions Obtained by NNs

In the previous section, we mentioned that effectiveness for utilizing predictions of GC & DC in dealing stocks depends strongly upon the directions of STMA & LTMA near the crossing points and the relative position of stock price versus GC & DC.

In the followings, we shall propose a new DSS which utilizes predictions (obtained by NNs) concerning the occurrence of GC & DC, the increase (decrease) rate of the future stock price several weeks ahead, and the relative position of the predicted stock price versus the predicted crossing point of the STMA & LTMA.

NDSS

Carry out dealing "Buy (Sell)" only when the following three conditions A), B) , and C) are satisfied.

Condition A): One of the following conditions (A-1) & (A-2) is satisfied.

(A-1) All of the outputs from n NNs are positive (negative) and the average of their absolute values is above 0.5.
(A-2) All of the following three conditions (2-a), (2-b), and (2-c) are satisfied.
(2-a) The greater part of the outputs from n NNs is positive (negative).
(2-b) The rate of the outputs having values over 0.7 (below - 0.7) exceeds 50 % of the number of the outputs having the same sign.
(2-c) The average of the outputs from n NNs is above 0.5 (below – 0.5)

Condition B): The number of the NNs whose outputs have negative (positive) sign among the m NNs which are prepared for making predictions concerning the rate of change of future stock price several weeks ahead is smaller than (S – 1).

Condition C): The predicted stock price (by NNs) is above (below) the crossing point of the GC(DC).

Remark 3.1. The n NNs in the Condition A) are prepared for making a prediction of GC(DC) several weeks before it occurs. Due to page, we don't go into details. Interested readers are kindly asked to read the paper [7].

Remark 3.2. Signs "positive" and "negative" in Condition A) and B) correspond to buying and selling, respectively.

Remark 3.3. Condition B) is prepared in order to let the proposed system make no response in the case that many NNs do not show any consistant forecast concerning the changes of the future stock price several weeks ahead.

4 Computer Simulations

We have carried out computer simulations concerning dealings of the Nikkei-225 from 1994 to 2008. Table 1 shows the changes of the initial amount of money (10 billion yen) during each year by utilizing the new DSS (NDSS) proposed in the last section, the PDSS [6] proposed several years ago, After Crossing Method (ACM) which carries out dealing based upon the traditional technical analysis (which carries out dealing only after GC (DC) is found), and the Buy-and-Hold method (BHM) [6]. We have also carried out computer simulations concerning dealings of the Nikkei-225 in 2010. Table 2 shows the changes of the initial amount of money (10 billion yen) in 2010 by each method. These computer simulation results confirm the effectiveness of the NDSS.

Remark 4.1: *In our simulations, we have set n = 7 and m = 21. We have used 7 k×k×1 neural network models (where k denotes the number of input variables having been chosen by the sensitivity analysis [8]) for checking whether GC (DC) will occur in several weeks. We have also used 3×7 = 21 neural network models for making predictions concerning the increase (decrease) rate of changes of each individual stock in the Tokyo Stock Market 3 weeks, 4 weeks, and 5 weeks in the future.*

Remark 4.2: *We have carried out neural network training by using the past data for three years. Table 2 shows the learning periods and the prediction periods.*

Remark 4.3: *In the above simulations, we have used the rule which allows "dealing on credit". Due to space, we don't go into details. Interested readers are kindly asked to attend our presentation.*

Remark 4.4: *In the above simulations, we have taken the charge for dealing into account by subtracting (0.001* (total money used for dealing) + 250,000) yen from the total fund for dealing.*

Remark 4.5: *In the above simulations, we have assumed that we could carry out dealing three times as much as the initial cash "10 thousand million yen" each year.*

Remark 4.6: *Due to our careless mistake, we have not been able to prepare the computer simulation results in 2009. We also regret that we are only able to provide partial computer simulation results in 2010 (Sorry! we have only prepared the*

simulation results in the case "s=12".) In the following, we shall show you the simulation results (We shall do our best in order to prepare perfect data in our presentation.):

Buy & Hold: - 523 million yen ; After Crossing: - 465 million yen;
NDSS: 2133 million yen

Table 1. Total Return in Each Year (1994~2008) Which Has Been Obtained by Each Method

(Nikkei-225; Million Yen)

	Buy&Hold	**After Cross**	**PDSS**	**NDSS(s=12)**
A1(1994)	1270	260	67	1,265
A2(1995)	154	42	-545	2,165
A3(1996)	- 630	189	-95	843
A4(1997)	-2,343	-455	7,178	316
A5(1998)	-954	-294	-294	3,054
A6(1999)	3,344	-914	5,052	8,580
A7(2000)	-2,695	2,285	2,684	-623
A8(2001)	-2,379	-437	-1,925	-2,272
A9(2002)	-2,014	-296	6,080	6,497
A10(2003)	2,255	-811	5,341	4,917
A11(2004)	451	-625	-191	0
A12(2005)	3,999	-541	10,505	10,505
A13(2006)	453	-1,063	-1,474	3,856
A14(2007)	-1,036	-337	1,456	-2,198
A15(2008)	-3,924	-604	-514	-2,963
Total Return	**-4,049**	**-3,601**	**33,325**	**33,942**
	NDSS(s=13)	**NDSS(s=14)**	**NDSS(s=15)**	**NDSS(s=16)**
A1(1994)	955	651	468	22
A2(1995)	4,269	4,269	4,269	4,269
A3(1996)	755	755	755	2,516
A4(1997)	2,081	2,081	9,381	9,381
A5(1998)	3,054	3,054	3,054	3,054
A6(1999)	8,580	8,580	8,580	8,580
A7(2000)	-623	-623	-623	-623
A8(2001)	-2,272	-2,272	-2,272	-2,272
A9(2002)	6,210	6,210	6,210	6,210
A10(2003)	4,917	4,917	4,917	4,917
A11(2004)	0	679	1,540	1,540
A12(2005)	10,505	10,505	10,505	10,505
A13(2006)	2,895	2,895	2,895	2,895
A14(2007)	-2,198	-2,198	-2,198	-1,341
A15(2008)	-2,963	-2,963	320	320
Total Return	**36,165**	**36,540**	**47,801**	**49,973**

Table 2. Learning Periods and Prediction Periods

	Learning Period	Prediction Period
A1	January 1991 - December 1993	January 1994 - December 1994
A2	January 1992 - December 1994	January 1995 - December 1995
.	.	.
.	.	.
.	.	.
A15	January 2005 - December 2007	January 2008 - December 2008
A17	January 2007 - December 2009	January 2010 - December 2010

5 Concluding Remarks

A decision support system for dealing stocks which improves the traditional technical analysis by utilizing NNs has been proposed. Computer simulation results having been done for rather long range of years (16 years) suggest the effectiveness of the proposed DSS. However, these simulations have been done only for the Nikkei-225. In order to execute full confirmation concerning the developed NDSS, we need to check whether it can be successfully applied for dealing in the other indexes such as S&P 500, DAX, and etc. For this purpose, we also need to carry out computer simulations concerning various individual stocks. Further, we also need to compare our approach with other approaches such as those utilizing Auto-regressive (AR) linear prediction method, and etc. This is also left for our future study.

References

1. Rumelhart, D.E., et al.: Parallel Distributed Processing. MIT Press, Cambridge (1986)
2. Haykin, S.: Neural Networks. Prentice-Hall, Englewood Cliffs (1998)
3. Baba, N., Kozaki, M.: An intelligent forecasting system of stock price using neural network. In: Proceedings of IJCNN 1992, pp. 371–377 (1992)
4. Refenes, A.-P.N., et al.: Neural Networks in Financial Engineering: A Study in Methodology. IEEE Trans. NNs, 1222-1267 (1997)
5. Chen, S.H., Nin, K. (eds.): Computational Intelligence in Economics and Finance. Springer, Heidelberg (2002)
6. Baba, N., Nomura, T.: An Intelligent Utilization of Neural Networks for Improving the Traditional Technical Analysis in the Stock Markets. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3681, pp. 8–14. Springer, Heidelberg (2005)
7. Baba, N., Nin, K.: Prediction of Golden Cross and Dead Cross by Neural Networks and Its Utilization. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 642–648. Springer, Heidelberg (2007)
8. Zurada, J.M., et al.: Sensitivity analysis for minimization of joint data dimension for feed forward neural network. In: Proceedings of the IEEE International Symposium on Circuits and Systems, pp. 447–450 (1994)

A New Framework for Assets Selection Based on Dimensions Reduction Techniques

Marina Resta

DIEM, sezione di Matematica Finanziaria,
Via Vivaldi 2, 16126, Genova, Italy
{mresta}@unige.it
www.diem.unige.it/Marina_RESTA.html

Abstract. We introduce a model called Asset Drivers Framework (ADF), which combines Dimensions Reduction Techniques (DRT) with a ranking procedure to find out assets to be inserted into a financial portfolio. The basic idea is that market securities can be described by a wider number of determinants, but only a few number of them can effectively characterize the assets to form well–balanced portfolios. The ADF manages this as a dimensions reduction problem, and extrapolates for each asset a reduced number of determinants as natural drivers of theirs. The procedure ends by assigning a score to the assets projected in such dimensionally reduced space, with a method of punishment/reward of the way the securities cluster into it. The beauty of the ADF scheme relies on a number of points: (*i*) it provides a platform to test various dimensions reduction techniques; (*ii*) looking at the performance, ADF makes possible to build portfolios whose returns are aligned to those of the traditional approach, but with lower variance, and hence lower risk.

Keywords: Dimensions Reduction Techniques, Assets selection, Assets Drivers, Clustering.

1 Introduction

The milestone on which Modern Portfolio Theory is built upon is the work by Markowitz [10], introducing the so called Mean–Variance (MV) approach. From the formal viewpoint, if $A = \{S_1, S_2, \ldots, S_n\}$ is the set of all possible assets in the market, each asset S_i $(i = 1, \ldots, n)$, may be associated to a history, i.e. a timeseries of v observable price levels: $\{l_i(t)\}_{t=1,\ldots,v}$, and then to the corresponding series of returns $r_i(t) = [l_i(t) - l_i(t-1)]/l_i(t-1)$, $(t = 2, \ldots, v)$. Investors will choose their portfolio composition by solving the quadratic optimization problem:

$$\min \sum_{i=1}^{n} \sum_{i=j}^{n} \sigma_{ij} x_i x_j \tag{1}$$

$$s.t.$$

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 372–381, 2011.

$$\sum_{i=1}^{n} \mu_i x_i \geq \mu_e \tag{2}$$

$$\sum_{i=1}^{n} x_i = 1 \tag{3}$$

$$x_i \geq 0,\ i = 1, \ldots, n \tag{4}$$

where n is the number of assets, μ_i is the expected return of asset S_i, σ_{ij} is the covariance of expected returns between assets S_i and S_j, and x_i $(i = 1, \ldots, n)$ is the (unknown) quantity that the trader can invest into each stock. The portfolio return is a random variable whose expected return is given by $\sum_{i=1}^{n} \mu_i x_i$, whereas μ_e is the minimum portfolio return the investors are bargaining for. Constraint (3) ensures that asset weights sum up to one, as they are considered as fractions of the whole amount of money to be invested. Finally, constraint (4) prevents investors from short–selling.

Indeed, that original model has been variously modified, introducing upper and lower bounds to the capital to be invested [9], or forcing the number of the portfolio assets to stay below a proper threshold [12]. More recently [8] discussed an approach based on Genetic Algorithms, while the memetic approach was adopted in [1]. Moreover, [4] introduced a definition of risk in the fuzzy environment, thus developing a new framework for the portfolio selection problem. The fuzzy approach was also recently adopted in [2], to develope a mean–variance–skewness model for portfolio optimization. However, our proposal differs from the aforementioned ones. We start by observing that the MV procedure implies that investors are (quite uncounsciously) performing some clustering task whose discriminants are the mean and the variance of the assets. Unfortunately, such clustering task is too much often affected by the curse of dimensionality: the issue is that market assets can be characterized by a wider set of features, including (but not limited to) mean and variance. In our opinion, then, the primary problem is not how to select assets, but rather how to extract the natural drivers of each asset: in practice, the portfolio problem is nothing but a matter of dimensionality reduction. With this in mind, we are going to discuss a scheme we have called the Asset Drivers Framework (ADF), which combines Dimensions Reduction Techniques (DRT) together with a ranking procedure to find out assets to be inserted into a financial portfolio. The aim of the procedure is to examine the features that can characterize the market assets, discarding those which are not really significant. The core of our method relies in the choice of a proper dimensions reduction technique [7] to reach this goal. What remains of the work is organized as follows: in Section 2 we will describe the now introduced ADF. In Section 3 we will run simulations on real data (namely we will create portfolios with the assets composing the Standard & Poor's SP–100 index), and we will discuss our results. Section 4 will conclude.

Input: Set $F_i = \{F_h^{(i)}\} \rightarrow S_i$,$AD = [\,]$,$W = 0$,$\pi = [\,]$
Output: Mapping to Portfolio π a set of assets
foreach *Asset* S_i **do**
  Project F_i with a dimension reduction technique;
  Evaluate the contribution of $F_h^{(i)}$;
  if $F_h^{(i)} \geq \theta_h$ **then**
    $AD = \{AD, F_h^{(i)}\}$;
  else
    Maintain AD unchanged
foreach *Element in AD* **do**
  Rank S_i;
  Assing sc_i
foreach S_i *in* A **do**
  $sc_i \rightarrow w_i$;
  $rL = \{rL, w_i\}$
sort rL;
Rank A according to rL;
set v=length(rL)
foreach $k = 1, \ldots, v$ **do**
  $W = W + rL(k)$;
  while $W < 1$ **do**
    $\pi = \{\pi, A(k)\}$;

Algorithm 1. The ADF algorithm

2 The Assets Driving Framework

The way ADF works is summarised making use of some pseudocode (see Algorithm 1): aiming to help the reader to understand it, we briefly introduce some notational conventions. In order to denote the set of assets in the market, we use the same notation we employed in the previous section: $A = \{S_i\}, (i = 1, \ldots, n < \infty)$ is the set of assets in the market. We also define by: $F_i = \{F_h^{(i)}\}$, $(h = 1, \ldots, z < \infty)$ the set of features that can perspectively drive the asset $S_i \in A$, while AD is the set of effective driving factors. This is initially an empty set, and it can reach a maximal length equal to z (i.e. equal to the number of available features), once the procedure has come to end. Moreover, we denote by θ_h the threshold values used to discriminate features, by $sc_i(j)$ the score of asset S_i according to the driving feature F_j, and by $w_i \in [0, 1]$ the average score of asset S_i over all its driving features. Finally, we set $rL = \{w_i\}$, $(i = 1, \ldots, n)$ as the list of weights associated to each asset, and we assume $W = \sum_i w_i \leq 1$ to be the control variable letting us to add assets to the portfolio π until the budget constraint is reached.

Clearly the aim of the procedure is to get a set AD whose length is sensitively lower than z. The core of the method relies in the choice of a proper technique to reach this goal. However, the main advantage is that the ADF is completely flexible, and a great variety of techniques can be suitable to be inserted into

it. To illustrate the effectiveness of the ADF, we actually focused on three well known methods: (*i*) Diffusion Maps; (*ii*) Elastic Maps, and (*iii*) Self Organizing Maps.

2.1 Diffusion Maps

Diffusion Maps [6] (DM) embed data in a lower–dimensional space, such that the Euclidean distance between points approximates the diffusion distance in the original feature space. DM, in fact, are based on defining a Markov random walk on the graph of the data: by performing the random walk for a number of timesteps, a measure for the proximity of the datapoints is obtained. Using this measure (the so-called diffusion distance) it is possible to build a transition matrix from two points $\mathbf{x_i}$ and $\mathbf{x_j}$ in the input space with entries:

$$p_{ij} = \frac{w_{ij}}{\sum_k w_{ik}} \tag{5}$$

where: $w_{ij} = e^{-\frac{||x_i - x_j||_E^2}{2\alpha^2}}$, and $||\cdot||_E$ is the standard Euclidean norm. Here α is the kernel scale and, by tweaking it we can choose the size of the neighbourhood: for nonlinear lowerdimensional structures, a small neighbourhood will be preferable, while for sparse data, a larger neighbouhood will be more appropriate. Since diffusion maps originate from dynamical systems theory, the resulting matrix $P^{(1)}$ is a Markov matrix that defines the forward transition probability matrix of a dynamical process. Hence, $P^{(1)}$ represents the probability of a transition from one datapoint to another in a single timestep, while for t timesteps we will have: $P^{(t)} = \left(P^{(1)}\right)^t$ and the diffusion distance will be given by:

$$D^{(t)}(x_i, x_j) = \sqrt{\sum_k \frac{(p_{ik}^{(t)} - p_{jk}^{(t)})^2}{\psi(x_k)^{(0)}}} \tag{6}$$

where $\psi(x_k)^{(0)} = \frac{m_i}{\sum_j m_j}$, and $m_i = \sum_j p_{ij}$. The idea is that the transition probabilities that are defined by P reflect the local geometry of the data. The higher t, the further a probability weight can diffuse to other points further away. In such a framework, a cluster is a region in which the probability of escaping this region is low. Moreover the computational solution can be easily found since it has been shown that the low-dimensional representation Y retaining diffusion distances $D^{(t)}(\mathbf{x_i}, \mathbf{x_j})$ as good as possible (under a squared error criterion) is formed by the d nontrivial (i.e. not equal to 1) principal eigenvectors of the eigenproblem:

$$P^{(t)}\mathbf{v} = \lambda\mathbf{v} \tag{7}$$

2.2 Elastic Maps

Formally, an Elastic Map [3] is an undirected graph $G(Y, E)$, where $Y = \{Y^{(j)}, j = 1, \ldots, N\}$ denotes the collection of graph nodes, and $E = \{E^{(j)}, j = 1, \ldots, s\}$ is the set of graph edges, each of them with a beginning node $E^{(j)}(0)$ and final node $E^{(j)}(1)$. Additionally, each pair of adjacent edges $R^{(j)} = \{E^{(j)}, E^{(k)}\}$ is an elementary rib, with a beginning node $R^{(j)}(1)$, end node $R^{(j)}(2)$ and the central node $R^{(j)}(0)$. The algorithm merges a finite input set $\Xi = \{\xi_r\}_{r<\infty}$ into the neural space through an iterative procedure whose main steps are here recalled.

Step 1. Divide input samples into subcollections satisfying the following criterion with respect to map nodes:

$$K_j = \{\xi^{(j)} : \min_Y ||\xi^{(j)} - Y^{(j)}||\} \tag{8}$$

Step 2. Define the graph energy function U, summarizing energies of every node, edge and rib: $U = U^{(Y)} + U^{(E)} + U^{(R)}$, being:

$$\begin{aligned} U^{(Y)} &= \tfrac{1}{N}\sum_{j=1}^{N} \sum_{\xi^{(r)} \in K^j} ||\xi^{(r)} - Y^{(j)}||^2 \\ U^{(E)} &= \sum_{j=1}^{s} \lambda_j ||E^{(j)}(1) - E^{(j)}(0)||^2 \\ U^{(R)} &= \sum_{j=1}^{s} \nu_j ||R^{(j)}(1) + R^{(j)}(2) - 2R^{(j)}(0)||^2 \end{aligned} \tag{9}$$

where values λ_j and ν_j are, respectively, coefficients of stretching elasticity of every edge $E^{(j)}$, and coefficients of bending elasticity of every rib $R^{(j)}$. We assume: $\lambda_1 = \lambda_2 = \ldots = \lambda_s = \lambda(s)$, and $\nu_1 = \nu_2 = \ldots = \nu_r = \nu(r)$. While $U^{(Y)}$ is somewhat self−explaining, $U^{(E)}$ may be correctly thought as the analogue of summary energy of elastic stretching, and $U^{(R)}$ as the summary energy of elastic deformation of the net. This makes every node connected by elastic bonds to the closest input data points, and simultaneously to the adjacent nodes.

Step 3. Minimize the graph energy U: new nodes positions are calculated accordingly. Note that:

$$\begin{aligned} \tfrac{1}{2}\tfrac{\partial U^{(Y)}}{\partial Y^{(j)}} &= N_j Y^{(j)} - \sum_{\xi^{(r)} \in K^{(j)}} \xi^{(r)} \\ \tfrac{1}{2}\tfrac{\partial U^{(E)}}{\partial Y^{(j)}} &= \sum_{k=1}^{N} Y^{(k)} \sum_{i=1}^{s} \lambda_i \Delta E^{ij} \Delta E^{ik} \\ \tfrac{1}{2}\tfrac{\partial U^{(R)}}{\partial Y^{(j)}} &= \sum_{k=1}^{N} Y^{(k)} \sum_{i=1}^{r} \nu_i \Delta R^{ij} \Delta R^{ik} \end{aligned} \tag{10}$$

where N_j is the number of points in $K^{(j)}$.

Step 4. If changes in U values become less than a small (arbitrary) ε, the procedure stops, otherwise steps (1)−(3) are repeated until such threshold value is reached.

2.3 Self Organizing Maps

In the Self Organizing Maps (SOM) [5] a finite set of input patterns is represented by means of a smaller number of nodes (neurons), sharing with inputs the same format, and arranged into a mono or bi–dimensional grid. When an arbitrary input is presented to a SOM, a competitive procedure starts, during which a winner or leader neuron is chosen in the map, as the best matching node, according to a metric previously fixed. In a generic step of the procedure[1] if $\mathbf{X}(t) = \{X_j(t)\}_{j=1,...,n} \in \mathbb{R}^n$ is the input item presented to a map M with q nodes with weights $\mathbf{m}_i(t) = \{m_{i,j}(t)\}_{j=1,...,n} \in \mathbb{R}^n$, $(i = 1, ..., q)$, i_t^* will be claimed the winner neuron at step t iff:

$$i_t^* = argmin_{i \in M} \left(\sum_{i \in M} \sum_{j=1}^{n} |X_j(t) - m_{i,j}(t)|^p \right)^{1/p}, \; p \in \mathbb{N} \tag{11}$$

Once the leader has been identified according to Eq.(11), the correction of nodes in the map takes place; if $N_{i^*}(t)$ is the set of neurons in the map belonging to the *neighbourhood* of i^* (in a topological sense), then:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{i*,i}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)] \tag{12}$$

Here $h_{i^*,i}(\cdot)$ is an interaction function, governing the way the nodes adjust respect to the winning neuron on the grid. After iterating such procedure over a number of epochs, the map should tend to a steady organized state, and neighbouring neurons should represent similar inputs.

3 Experiments and Results Discussion

Our data sample consists of the assets included into the SP–100 index, in the period from April 4th 2007 to October 10th 2010, for an overall number of 901 daily observations. Earlier 700 data have been used as input set, while the remaining data have been used as test set. For each asset we have analyzed 20 different factors that belong to various indicators families: we considered both statistical indicators like mean, variance, and skewness, observed at different time frames (going backward from 50 to 200 days), and financial indicators like the assets Beta: $\sigma_{ij}(k)/\sigma_M^2(k)$, and the Risk to Return Ratio: $\sigma_i/\mu_i(k)$, with $k = 50, 100, 150, 200$. We then moved to perform the steps characterizing the ADF. In the first stage, we examined the aforementioned 20 variables for each asset, discriminating them through a dimensions reduction technique. In that case, a not negligible issue concerned the choice of model parameters requested by the dimensions reduction technique in use. Clearly, we have to select each of these parameters wisely. However, note that working with financial data requests to combine good theoretical knowledge together with some practical expertise,

[1] We will refer to the case of a mono–dimensional SOM, but the layout presented can be easily generalized to higher dimensional grids.

so that the choice of parameters is often guided by common sense. In addition, since each market is a *uniquum* the choice of parameters is quite subjective. With this in mind, in the diffusion framework we were asked to determine the time scale α. In our numerical experiments we selected α to be the mean of the Euclidean distances from each point to its k–nearest neighbour, where k equals to 0.5 percent of the total number of points in the training set. In the case of Elastic Maps, on the other hand, the data were trained in two epochs, the first one with $\lambda_0 = 0.1$, and $\nu_0 = 200$; the second with $\lambda_0 = 0.01$, and $\nu_0 = 25$, without any optimization of the final value of ν_0. Various resolution grids were examined: we run experiments with classical rectangular grids, with dimensions varying from 10×10 to 25×25, as well as with fractal growing structures, and spherical ones. However, the most reliable structures revealed to be a structure based on a 12×12 rectangular grid. For what is concerning the Self Organizing Maps, we trained a bunch of 50 plain SOMs with rectangular grid topology, and dimensions varying from 5×5 to 21×21, for an average number of epochs equal to 10. The SOM with best performances in terms of quantization error (a 21×12 grid) was then isolated and used to our purposes.

All the simulations concerning DM and SOM were run on Matlab. In particular, experiments relying on DM have been done with some slight modifications of the code in the van der Maaten's Toolbox for Dimensionality Reduction [11]; SOM were run with the SOM toolbox [14]. Finally, to implement EM we made use of the ViDaExpert software [16].

Figure 1 shows the projection of assets obtained with the different methods at the end of the first step: the position of each asset is the direct consequence of the mapping with respect to all the drivers of it. We were then able to set apart 4 of the original variables in the case of DM[2], 6 in the case of EM[3], and 12 in the case of SOM[4].

We then performed the final stage of our procedure, and we assigned a score to the assets after isolating their drivers. The main issue is that the scoring procedure must reward both the clusters significance and the diversity of the mapping, i.e. two aspects going towards opposite directions. In fact, the higher the number of assets falling into a cluster, the higher the cluster significance will be. However, this true, the lower will be the diversification allowed among the assets. Having in mind those issues, the score the procedure assigns is formed according the following rule:

$$w_i = w_i^{(init)} + \frac{1}{cr_i - nrc - 1} + \left(1 - \frac{asr_i}{nrn}\right) \cdot resc_i \tag{13}$$

where: nrc is the number of identified clusters, cr_i and asr_i are the cluster and asset i ranking respectively, nrn is the number of nodes within the cluster i belongs to, and, finally, $resc_i$ is the distance between asset i and its corresponding

[2] 50–days variables.

[3] 100–days backward mean, variance, skewness and beta, and 200 days backward mean and variance.

[4] Mean, variance, and beta at all time frames.

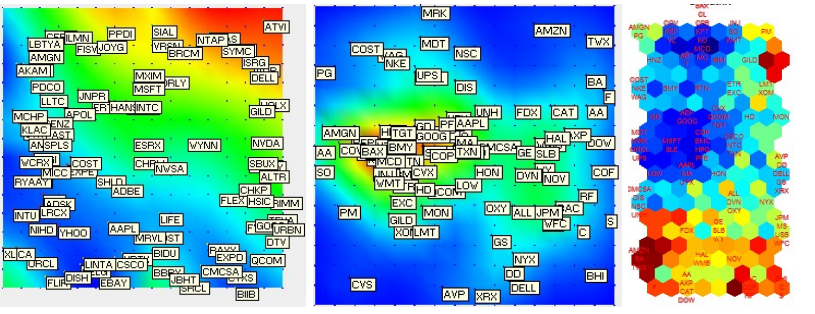


Fig. 1. Overall mapping for the SP–100 components in the ADF with different dimensions reduction techniques. DM is the abbreviation for Density Maps, EM for Elastic Maps, and SOM for Self–Organizing Maps.

node rescaled in the range $[0, 1]$. Here $w_i^{(init)} = 0$, because we are studying a single–period optimization problem. Clearly, in a dynamic framework this variable can be set to assume more meaningfull values. We were then able to assign a weight to each component of the SP–100. The performances obtained by the ADF have been evaluated by analyzing, for each portfolio, the corresponding average return (AR), and two well–known conventional indexes: the Sharpe ratio [13] SR, and the leverage factor L, being:

$$SR = \frac{\mu_\pi - rf}{\sigma_\pi}; \; L = \frac{\sigma_{mkt}}{\sigma_\pi} \tag{14}$$

where μ_π, and σ_π are, respectively, the portfolio mean return and standard deviation, rf is the risk–free rate, and, finally, σ_{mkt} is the reference index (SP–100) standard deviation. We have set $rf = 0.005$, considering it equivalent on bi–monthly basis, to the annual rate $rf_y = 0.03$. The results are given in Table 1, where the scores produced by running classical MV portfolios are compared from those obtained within the ADF.

Table 1. Performances comparison. MV is the abbreviation for Mean–Variance, DM for Density Maps, EM for Elastic Maps, and SOM for Self–Organizing Maps.

	MV	DM	EM	SOM
AR	0.01	0.012	0.02	0.011
SR	0.008	0.021	0.04	0.006
L	0.009	0.016	0.016	0.012

In general, the ADF seems to perform well: of the examined dimensions reduction techniques, only Density Maps provide results sensitively lower than the MV approach. The Sharpe Ratio of the ADF in two cases of three is higher than that of MV. By construction of SR, those results imply that MV is riskier: the higher AR level of MV is counterbalanced by the higher level of portfolio

variance. Moreover, the leverage value suggests the profitability of almost all the examined portfolios.

4 Conclusion

We have presented a scheme called the Assets Drivers Framework (ADF) that allows to select assets according to their natural drivers. The basic idea has been that to combine Dimensions Reduction Techniques (DRT) together with a ranking scheme to find out assets to be inserted into a financial portfolio. We examined the feasibility of the approach with an input set made by the securities composing the SP–100 index, initially described by 20 determinants of various kind. We then isolated the most promising features using a dimension reduction technique. To such purpose, we considered the results for three DRT: Density Maps (DM), Elastic Maps (EM), and Self–Organizing Maps (SOM). The assets now mapped into a reduced dimensional space were then examined again, and to each of them a score was assigned. The results obtained by the aforementioned dimensions reduction techniques into the ADF were there compared to those obtained within the classical Mean–Variance (MV) approach. The results, in our opinion, are interesting from different points of views. Firstly, our scheme is generally more informative than MV, because it allows assets selection through a wider set of significant factors. Not less importantly, the ADF allows the direct comparison of a bunch of data driven models. Looking deepest to the results, we have evidenced that EM and SOM are valid performers, because they allow to insert a limited number of securities in the portfolio (similar to the number of a MV portfolio), maintaining the returns substantially aligned to those of the MV scheme, but with lower levels of variance. As future task we are planning to extend our approach in a dynamic context to study the flexibility of the procedure with respect to sequential portfolio calibrations.

References

1. Aranha, C., Iba, H.: Using Memetic Algorithms to Improve Portfolio Performance in Static and Dynamic Trading Scenarios. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO (2009)
2. Bhattacharyyaa, R., Kara, S., Majumderb, D.D.: Fuzzy mean-variance-skewness portfolio selection models by interval analysis. Computers & Mathematics with Applications 61(1), 126–137 (2011)
3. Gorban, A., Kegl, B., Wunsch, D., Zinovyev, A.: Principal Manifolds for Data Visualisation and Dimension Reduction. Lecture Notes in Computational Science and Engineering, vol. 58. Springer, Heidelberg (2007)
4. Huanga, X.: Risk curve and fuzzy portfolio selection. Computers & Mathematics with Applications 55(6), 1102–1112 (2008)
5. Kohonen, T.: Self-Organizing Maps. Springer, Berlin (2005)
6. Lafon, S., Lee, A.B.: Diffusion maps and coarse–graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(9), 1393–1403 (2006)

7. Lee, J.A., Verleysen, M.: Nonlinear dimensionality reduction. Information Science and Statistics series. Springer, Heidelberg (2007)
8. Lin, C.M., Gen, M.: An effective Decision–Based Genetic Algorithm Approach to Multiobjective Portfolio Optimization Problem. Applied Mathematical Sciences 1(5), 201–210 (2007)
9. Mansini, R., Speranza, M.G.: Heuristic algorithms for the portfolio selection problem with minimum transaction lots. European Journal of Operational Research 114(2), 219–233 (1999)
10. Markowitz, H.: Portfolio selection. Journal of Finance 7(1), 7–91 (1952)
11. Matlab Toolbox for Dimensionality Reduction, http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html
12. Schaerf, A.: Local search techniques for constrained portfolio selection problems. Computational Economics 20(3), 177–190 (2002)
13. Sharpe, F.: Mutual Fund Performance. Journal of Business, 119-138 (January 1966)
14. SOM toolbox, http://www.cis.hut.fi/projects/somtoolbox/download/
15. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
16. ViDaExpert software, http://www.ihes.fr/~zinovyev/vida/vidaexpert.htm
17. Xia, Y., Wang, S., Lai, K.K.: A model for portfolio selection with order of expected returns. Computers & Operations Research 27(5), 409–422 (2000)

Forecasting Stock Price Based on Fuzzy Time-Series with Entropy-Based Discretization Partitioning

Bo-Tsuen Chen[1], Mu-Yen Chen[1], Hsiu-Sen Chiang[1], and Chia-Chen Chen[2]

[1] Department of Information Management, National Taichung Institute of Technology, Taichung, Taiwan (R.O.C)
{s1899b109,mychen,hschiang}@ntit.edu.tw
[2] Department of Information Management, Tunghai University, Taichung, Taiwan (R.O.C)
emily@thu.edu.tw

Abstract. The prediction of stock markets is an important and widely research issue since it could be had significant benefits and impacts. In this paper, we applied entropy-based discretization partitioning to obtain optimized linguistic intervals setting for fuzzy time-series model. In order to evaluate our proposed approach, the dataset collected from Taiwan Stock Exchange (TAIEX). Finally, the experimental results showed that our proposed approach was effective in finding for the better linguistic intervals settings, when the entropy-based discretization partitioning is applied. Furthermore, the performances indicate that the proposed model is superior to the compared models suggested by Chen (1996) and Yu (2005) earlier. It is evident that the entropy partitioning is a good approach to obtain optimized linguistic intervals for fuzzy time-series models.

Keywords: Fuzzy Time-Series, Entropy, Stock Price Forecasting, Linguistic Intervals.

1 Introduction

Mathematical and statistical methods [1, 2] have usually been employed to financial forecasting in early research, such as time series analysis [8] and multiple regression models [3, 4], but traditional time-series requires more historical data along with some assumptions like normality postulates. However, there is considerable evidence that stock market behavior is not fully efficient and is highly nonlinear, so above analysis models cannot meet our research desires. Consequently, using fuzzy time-series to forecast the price activity in stock market has been suited a common method [10, 11, 12].

The fuzzy time series model proposed by Song and Chissom [10], yet different fuzzy time-series models have been proposed to forecast nonlinear data and applied to various applications, such as enrollment, temperature, car road accidents, tourism demand and the stock index, etc. In empirical analysis, various linguistic values are used to evaluate and determine the efficacy one for the proposed model, because the length of linguistic intervals may affect forecasting results from Huarng research [6], and

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 382–391, 2011.

Huarng proposed the distribution-based length method and the average-based length method for setting the length of linguistic intervals.

In this paper, a new fuzzy time-series model applying entropy-based discretization partitioning is proposed to obtain optimized linguistic intervals setting for fuzzy time-series model to improve the forecast accuracy in stock market, and it can adjust the length of each interval in the universe of discourse by entropy-based discretization partitioning each time. Furthermore, the proposed model can get a higher forecasting accuracy rate than the compared models suggested by Chen (1996) and Yu (2005). It is evident that the entropy-based discretization partitioning is a good approach to obtain optimized linguistic intervals for fuzzy time-series models.

The remainder of this paper is organized as follows: Section 2 provides an overview of fuzzy time-series model and entropy-based discretization. Section 3 describes the proposed model. Section 4 presents the experimental results from a simulate dataset. Conclusions are finally drawn in Section 5, along with recommendations for future research.

2 Literature Review

2.1 Fuzzy Time-Series

In this session, we briefly review the concept of fuzzy time series from [10, 11, 12]. Song and Chisoom (1993) first applied the fuzzy theory in time-series and provided the definitions and framework of fuzzy time-series, the main difference of fuzzy time series and traditional time series is that the values of fuzzy time-series are represented by fuzzy sets [14] rather than real values. The definitions and research process are introduced as follows [15]. Let Ube the universe of discourse, where $U = [u_1, u_2, \dots, u_n]$. A fuzzy set defined in the universe of discoursecan be represented as follows:

$$A=f_A(u_1)/u_1+f_A(u_2)/u_2+\ldots+f_A(u_n)/u_n \tag{1}$$

where f_A denotes the membership function of the fuzzy set A, $f_A:U\rightarrow[0, 1]$, and $f_A(u_i)$ denotes the degree of membership of u_i belonging to the fuzzy set A, and $f_A(u_i)\in [0,1]$, and $1 \le i \le n$.

Definition 1. Let $Y(t)(t=\ldots,0,1,2,\ldots)$ be the universe of discourse and be subset of R. Assume $f_i(t)(i= 0,1,2,\ldots)$ are defined on $Y(t)$, and assume that $F(t)$ is a collection of $f_1(t), f_2(t), \dots$, then $F(t)$ is called a fuzzy time-series definition on $Y(t)$.

Definition 2. Assume that $F(t)$ is caused by $F(t\text{-}1)$ only, denoted as $F(t\text{-}1)\rightarrow F(t)$, then this relationship can be expressed as $F(t) = F(t\text{-}1) \circ R(t,t\text{-}1)$, where $F(t) = F(t\text{-}1) \circ R(t,t\text{-}1)$ is called the first-order model of $F(t)$, $R(t,t\text{-}1)$ is the fuzzy relationship between $F(t\text{-}1)$ and $F(t)$, and " $\circ$ " is the Max-min composition operator.

Definition 3. Let $R(t,t\text{-}1)$ be a first-order model of $F(t)$. If for any t, $R(t,t\text{-}1) = R(t\text{-}1,t\text{-}2)$, then $F(t)$ is called a time-invariant fuzzy time-series, otherwise, it is called a time-variant fuzzy time-series.

Definition 4. Let $F(t\text{-}1) = A_i$ and $F(t) = A_j$, it can be denoted by $A_i \rightarrow A_j$, where A_i is called the left-hand side (LHS), A_j is the right-hand side (RHS) and be called the current state of the fuzzy logical relationship (FLR).

2.2 Entropy-Based Discretization

Entropy is one of the most commonly used discretization measures. It was first introduced by Claude Shannon in pioneering work on information theory and the concept of information gain. Entropy-based discretization is a supervised, top-down splitting technique [6]. It explores class distribution information in its calculation and determination of split-points (data values for partitioning an attribute range). To discretize a numerical attribute, A, the method selects the value of that has the minimum entropy as asplit-point, and recursively partitions the resulting intervals to arrive at a hierarchical discretization. Such discretization forms a concept hierarchy for A.

Let D consist of data tuples defined by a set of attributes and a class-label attribute. The class-label attribute provides the class information per tuple. The basic method for entropy-based discretization of an attribute A within the set is as follows:

(1) Each value of A can be considered as a potential interval boundary or split-point to partition the range of A. That is, a split-point for A can partition the tuples in D into two subsets satisfying the conditions $A \leq$ split point and $A >$ split point, respectively, thereby creating a binary discretization.
(2) Entropy-based discretization, as mentioned above, uses information regarding the class label of tuples. To explain the intuition behind entropy-based discretization, we must take a glimpse at classification. Suppose we want to classify the tuples in D by partitioning on attribute A and some split-point. Ideally, we would like this partitioning to result in an exact classification of the tuples. For example, if we hadtwo classes, we would hope that all of the tuples of, say, class C_1 will fall into onepartition, and all of the tuples of class C_2 will fall into the other partition. However, this is unlikely. For example, the first partition may contain many tuples ofC_1, but also some ofC_2. This amount which called the expected information requirement for classifying a tuple in D based on partitioning by A. It is given by:

$$Info_A(D) = \frac{|D_1|}{|D|} Entropy(D_1) + \frac{|D_1|}{|D|} Entropy(D_2) \tag{2}$$

where D_1 and D_2 correspond to the tuples in D satisfying the conditions $A \leq$ split point and $A >$ split point, respectively; $|D|$ is the number of tuples in D. The entropy function for a given set is calculated based on the class distributionof the tuples in the set. For example, given m classes, C_1, C_2, … ,C_m, the entropy of D_1 is

$$Entropy(D_1) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{3}$$

where p_i is the probability of class C_1 inD_1, determined by dividing the number of tuples of class C_i in D_1 by $|D_1|$, the total number of tuples in D_1. Therefore, when selecting a split-point for attribute A, we want to pick the attribute value that gives the

minimum expected information requirement (i.e., min($Info_A(D)$)). This would result in the minimum amount of expected information still required to perfectly classify the tuples after partitioning by $A \leq$ split point and $A >$ split point. This is equivalent to the attribute-value pair with the maximum information gain.

The process of determining a split-point is recursively applied to each partition obtained, until some stopping criterion is met, such as when the minimum information requirement on all candidate split-points is less than a small threshold, ε, or when the number of intervals is greater than a threshold, *max_ interval*.

3 The Research Methodology

As stated in Section 1, there are some drawbacks to the time series models: (1) Proper weight for the fuzzy logic relationship; (2) Reliable length of intervals; (3) Patterns of price charges should be considered. To reconcile these drawbacks, this study considers that the optimization linguistic intervals can significantly affect the forecasting accuracy. Because this model uses the entropy-based discretization partitioning to obtain the reliable length of intervals, and assign a proper weight to individual fuzzy relationships, furthermore, this study considers the patterns of price charges by time-series.

Base on the concept above, this study proposes a new fuzzy time-series model with entropy-based discretization partitioning to forecast the TAIEX stock index. Firstly, this study defines the universe of discourse. Second, optimizes the linguistic intervals with entropy-based discretization partitioning by Eq. (2) and Eq. (3).Third, establish fuzzy logical relationships and group the current states of the data of the fuzzy logical relationships. Finally, apply the Centroid method for defuzzification to get the forecast results, and compare the forecast results with other tradition models. The overall flowchart of the proposed model is shown in Fig.1.

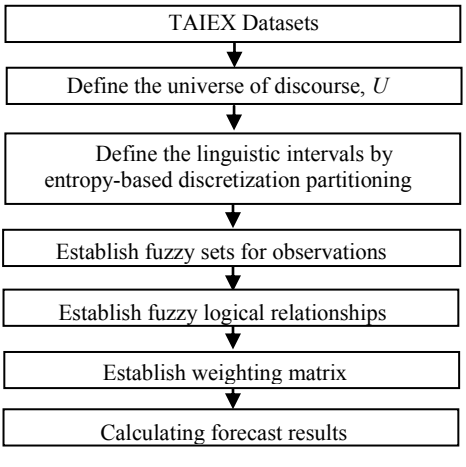


Fig. 1. Research processes of the proposed model

This section uses some numerical data as an example, and the core concept of the proposed algorithm is shown step by step.

Step 1: Define the universe of discourse, *U*

In this step, let D_{min} and D_{max} be the minimum stock price and the maximum stock price of known historical data. Based on D_{min} and D_{max}, define the universe of discourse U as $[D_{min} - D_1, D_{max} + D_2]$, where $D_1$and D_2 are two proper positive numbers. For example, the U of the TAIEX in the training period (1991/1/3-1991/10/30), are 3316 and 6306, respectively.

Step 2: Define the linguistic intervals by entropy-based discretization partitioning

There are two sub-steps implemented in this step: (1) Define the linguistic value for the proposed model and (2) Setting the universe of interval by entropy-based discretization partitioning.

(1) Define the linguistic value for the proposed model

In the first sub-step, the universe of discourse should be partitioned into seven linguistic values, from the research of Miller (1994) [9], he suggested that the appropriate number of category for human shorten memory function is seven, or seven plus or minus two. Hence, we employ seven as the linguistic value for the proposed model.

(2) Setting the universe of interval by entropy-based discretization partitioning

In the initial iteration, the universe of discourse U is divided into seven equal length intervals u_1, u_2, u_3, u_4, u_5, u_6, u_7 by the average-based length method [5],as shown in Table 1, where u_1 = [3316, 3743],u_2 = [3743, 4170], u_3 = [4170, 4597], u_4 = [4597, 5025],u_5 = [5025, 5425],u_6 = [5425, 5879], u_7 = [5879, 6306] in the training period (1991/1/3-1991/10/30) , and the seven linguistic values can be defined as follows: A_1 = (very low stock index), A_2 = (low stock index), A_3 = (little low stock index), A_4 = (normal stock index), A_5 = (little high stock index), A_6 = (high stock index), A_7 = (very high stock index).

After the initial iteration, the universe of intervals should further be setting into optimization definitions by the entropy-based discretization partitioning, because the change of the linguistic interval and the entropy-based discretization method for partitioning the datum more properly, the observations would be fuzzified into the optimization linguistic values more reliably.

Table 1. The seven intervals of TAIEX training dataset (1991)

Intervals	1st iteration	2nd iteration	3rd iteration	4th iteration
u_1	[3316, 3743]	[3316, 3776]	[3316, 3776]	[3316, 3802]
u_2	[3743, 4170]	[3776, 4191]	[3776, 4237]	[3802, 4237]
u_3	[4170, 4597]	[4191, 4622]	[4237, 4622]	[4237, 4639]
u_4	[4597, 5025]	[4622, 5033]	[4622, 5049]	[4639, 5090]
u_5	[5025, 5425]	[5033, 5455]	[5049, 5585]	[5090, 5585]
u_6	[5425, 5879]	[5455, 5900]	[5585, 5900]	[5585, 5925]
u_7	[5879, 6306]	[5900, 6306]	[5900, 6306]	[5925, 6306]

Table 2. The linguistic values of TAIEX training dataset (1991)

Date	Index	1st Iteration	2nd Iteration	3rd Iteration	4th Iteration
1991/2/25	5012.46	A_4	A_4	A_4	A_4
1991/2/28	5033.37	A_5	A_4	A_4	A_4
1991/2/19	5048.48	A_5	A_5	A_4	A_4
1991/7/25	5054.58	A_5	A_5	A_4	A_4
1991/8/03	5060.68	A_5	A_5	A_4	A_4
1991/8/14	5089.95	A_5	A_5	A_5	A_5
1991/8/13	5104.43	A_5	A_5	A_5	A_5

Table 3. The weighting matrix for training dataset (1991)

P(t-1)	P(t)						
	A_1	A_2	A_3	A_4	A_5	A_6	A_7
A_1	0.83	0.17	0	0	0	0	0
A_2	0.04	0.85	0.1	0	0	0	0
A_3	0	0.05	0.74	0.21	0	0	0
A_4	0	0	0.12	0.78	0.1	0	0
A_5	0	0	0	0.19	0.75	0.06	0
A_6	0	0	0	0	0.06	0.74	0.2
A_7	0	0	0	0	0	0.21	0.79

Take the TAIEX training dataset be example as shown in Table 1 and Table 2, after the implements of first interaction, the training dataset all be fuzzified into respective interval values, then we sort the datum from the small to the large, and find out the respective entropy values by partitioning each data. Finally, we can obtain the minimal entropy values for all data; therefore, we can reset the universe of linguistic intervals and the parameter of membership functions by the minimal entropy values, and launch the follow steps of the second iteration.

Step 3: Establish fuzzy sets for observations

The fuzzy sets, A_1, A_2, … , A_k of the universe of discourse is established by Eq. (4). The value of a_{ij} indicates the grade of membership of u_j in fuzzy set A_i, where $a_{ij} \in [0,1]$, $1 \le i \le k$ and $1 \le j \le m$. The degree of each stock index is used to establish its belonging fuzzy set, $A_i (i = 1,\ldots,m)$. If the maximum membership of the stock index is under A_k, then the fuzzified stock index is labeled as A_k. Table 2 shows seven linguistic values A_1-A_7 by using Eq. (4). Usually, we select the triangular-shaped as the A_2, A_3, A_4, A_5, A_6 membership functions, and select the trapezoidal-shaped as the A_1, A_7 membership functions.

$$\begin{aligned} A_1 &= a_{11}/u_1 + a_{12}/u_2 + \ldots + a_{1m}/u_m \\ A_2 &= a_{21}/u_1 + a_{22}/u_2 + \ldots + a_{2m}/u_m \\ &\vdots \\ A_k &= a_{k1}/u_1 + a_{k2}/u_2 + \ldots + a_{km}/u_m \end{aligned} \tag{4}$$

Step 4: Establish fuzzy logical relationships
The FLRs are generated based on the fuzzified observation. One-order FLR is constructed by two consecutive fuzzy sets. Take the second iteration in Table 2 for example, the FLRs are demonstrated as follows: $A_4 \rightarrow A_4$, $A_4 \rightarrow A_5$, $A_5 \rightarrow A_5$, $A_5 \rightarrow A_5$, $A_5 \rightarrow A_5$, $A_5 \rightarrow A_5$.

Step 5: Establish FLR groups
After all the FLRs are generated, it would construct a trend-weighted matrix for all FLRs. Each column in the matrix represents the occurrence of FLRs. It is represented that the more recent ones have higher weights than the older ones. Take the second iteration in Table 2 for example, the FLR groups are demonstrated as $A_4 \rightarrow A_4$, A_5 and $A_5 \rightarrow A_5$.

Step 6: Establish weighting matrix
The weighting matrix are based on above trend-weighted matrix, and transfer these weights into a normalized weight matrix, $w_i(t)$, which is defined in the following equation (Yu, 2005)[4]:

$$w_i(t) = [w_1^{'}, w_2^{'}, ..., w_m^{'}] = [\frac{w_1}{\sum_{j=1}^{m} w_j}, \frac{w_2}{\sum_{j=1}^{m} w_j}, ..., \frac{w_m}{\sum_{j=1}^{m} w_j}] \quad (5)$$

where w_i is the corresponding weight for fuzzy set A_i, $1 \le i \le k$ and $1 \le j \le m$. By Eq. (5), the weight matrix is determined as Table 4.

Step 7: Calculating forecast results
In this step, the forecasts are produced based on the rules and the normalized weights from Step 5 and 6, and the forecasted value can be obtained by matrix multiplication of the defuzzified matrix and weighting matrix as follows:

$$F(t) = M_{df}(t\text{-}1) \circ w_i(t\text{-}1) \quad (6)$$

where $M_{df}(t\text{-}1)$is the deffuzified matrix, which applying the Centroid method for defuzzification, $w_i(t\text{-}1)$ is the weighting matrix. Take the observation on1991/8/13 from Table 2 for example; the forecasted value can be obtained as follows:

$$F(t) = [4828, 5244, 5678] \circ [0.19, 0.75, 0.06] = 5191$$

4 Experimental Results

This section describes the contents of the estimating TAIEX model and compares its performance. To verify the forecasting performance of the proposed model, large amounts of practical stock index datasets are needed. For this reason we use the daily TAIEX closing prices covering the thirteen-year period from 1991 to 2003 as the dataset (including thirteen sub-datasets). Each year of experimental datasets is splits into two subsets, a training dataset and a testing dataset, where the sub-datasets for the previous ten-month of each year from January to October is used for training data,

while those for November and December are selected for forecasting testing data. To inspect forecasting performance for the propose model, the forecast accuracy is compared by the indicator root of mean squared error (RMSE). Hence, the RMSE is employed as evaluation criterion for the forecasting performance of proposed model and comparison models, which is defined as Eq. (7):

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (actual(t) - forecast(t))^2}{n}} \tag{7}$$

where *actual* (*t*) is the actual stock index on time *t*, the *forecast* (*t*) is the forecasting value for actual stock index (*t*), and *n* denotes the count of forecasts in the testing period.

In order to forecast the fuzzy time-series, we need to determine the starting value for the universe of discourse as in Step 1. To this end, we round down the minimal data for each year to the nearest integer and set the value as the starting, and the set the value of maximal data for each year to the nearest integer as the ending.

With seven linguistic values, we generate forecasting values for the TAIEX in thirteen testing periods. And two fuzzy time-series models, Chen's (1996) and Yu's (2005) are employed as comparison models. After forecasting, the RMSE obtained using the conventional and weighted fuzzy time-series models, a performance comparison table is shown as Table 4. In addition, Fig 2 is produced to illustrate the forecasting performances of the different of whether applying entropy-based discretization partitioning.

On comparison of the three proposed models from Table 4, the proposed model bears all the smallest RMSE in thirteen testing period, it is obvious that our model surpass Chen's (1996) and Yu's (2005) in forecasting performance, and demonstrating that the model reduces forecasting errors more effectively when the entropy-based discretization partitioning are applying.

Table 4. Performance comparisons for TAIEX

Models	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	Average
Chen	80	60	110	112	79	54	148	167	149	176	148	101	74	112
Yu	61	67	105	135	70	54	133	151	142	191	167	75	66	109
Propose	42[a]	38[a]	100[a]	66[a]	47[a]	40[a]	116[a]	109[a]	100[a]	117[a]	93[a]	58[a]	43[a]	75[a]

[a]Best performance among three approaches.

In stock markets, investors usually made their decisions based on recent stock prices on the previous couple of days. However, most of the conventional models only extract fuzzy logical relationships from a long period of historical data to generate forecasting rules. Therefore, the comparison models cannot adapt their forecasts to meet recent price fluctuations to reduce the forecasting error. Hence, we can argue that the proposed method has made a great improvement in forecasting the stock market.

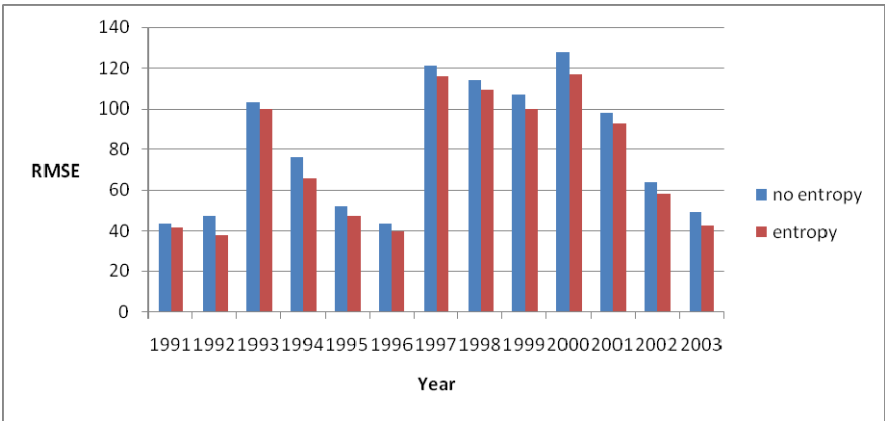


Fig. 2. The performance comparisons for entropy-based discretization partitioning

5 Conclusions and Future Research

In this paper, we have proposed a new model based on fuzzy time-series, this hybrid model combines the trend-weight matrix approach and the entropy-based discretization partitioning to forecast stock markets. Furthermore, the proposed model was compared with two different conventional fuzzy time-series models proposed earlier by Chen's (1996) and Yu's (2005), and the comparison shows that the proposed model surpasses all of thirteen testing dataset in the TAIEX stock markets.

Generally, from the experimental results, three conclusions are given: (1)The weighting matrix based on trend-weighted addresses the issue of lack proper weight for the fuzzy logic relationship; (2) By the entropy-based discretization partitioning, it can refine the reliable universe of intervals objectively, and improve forecasting performance effectively; (3) Patterns of price charges are be considered by fuzzy time-series. Based on the above conclusions, the proposed model which is overcomes the drawbacks mentioned in section 3.

Acknowledgments. The authors thank the support of National Scientific Council (NSC) of the Republic of China (ROC) to this work under Grant No. NSC-99-2410-H-025-011. Moreover, we also thank for STATSOFT STATISTICA 9.0 software to support related experiments.

References

[1] Altman, E.I.: Financial ratios discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance 23, 589–609 (1968)
[2] Beaver, W.H.: Financial ratios as predictors of failure. Journal of Accounting Research 4, 71–111 (1966)
[3] Bollerslev, T.: Generalized autoregressive conditional heteroscedasticity. Journal of Econometrics 31, 307–327 (1986)

[4] Box, G., Jenkins, G.: Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco (1976)
[5] Chen, S.M.: Forecasting enrollments based on fuzzy time series. Fuzzy Sets and Systems 81, 311–319 (1996)
[6] Han, J., Micheline, K.: Data Mining Concepts and Techniques. Morgan Kaufmann, San Francisco (2001)
[7] Huarng, K.H.: Effective lengths of intervals to improve forecasting in fuzzy time series. Fuzzy Sets and Systems 123(3), 387–394 (2001)
[8] Kendall, S.M., Ord, K.: Time series, 3rd edn. Oxford university press, New York (1990)
[9] Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity of processing information. The Psychological Review 101, 343–352 (1994)
[10] Song, Q., Chissom, B.S.: Forecasting enrollments with fuzzy time series–Part 1. Fuzzy Sets and Systems 54, 1–9 (1993a)
[11] Song, Q., Chissom, B.S.: Fuzzy time series and its models. Fuzzy Sets and Systems 54, 269–277 (1993b)
[12] Song, Q., Chissom, B.S.: Forecasting enrollments with fuzzy time series–Part 2. Fuzzy Sets and Systems 62, 1–8 (1994)
[13] Yu, H.K.: Weighted fuzzy time series models for TAIEX forecasting. Physica A, 349609–349624 (2005)
[14] Zadeh, L.A.: The concept of a linguistic variable and its application to approximation reasoning - Part I. Information Sciences 8, 199–249 (1975)

On the Use of Feed-Forward Neural Networks to Discriminate between Models in Financial and Insurance Risk Frameworks

Enrico di Bella

Dipartimento di Economia e Metodi Quantitativi - Sezione di Statistica,
Via Vivaldi, 5, 16100, Genoa, Italy
edibella@economia.unige.it

Abstract. The problem of assessing if a sample is coming from one of two probability distributions is most likely one of the oldest problems in the field of testing statistical hypotheses and a number of papers has been produced over the years without finding a most powerful test for this goal. In financial and insurance risk modeling, this problem is often addressed to identify the best extreme values model in a battery of alternatives or to design the heaviness of the tail of the underlying distribution. Taking advantage of the well known performance in classificatory problems of neural networks, the use of feed-forward neural networks for discrimination between two distributions is herein proposed and the power of a neural goodness-of-fit test is estimated for small, moderate and large sample sizes in a wide range of symmetric and skewed alternatives. The empirical power of the procedure described is compared to the power of eight classic and well known normality tests for a sample to come from a normal distribution against each of twelve close-to normal alternatives. The neural test resulted to be the most powerful in the whole battery and its behavior was consistent with the expected statistical properties.

Keywords: Discrimination between distributions, Extreme values, Feedforward Neural Networks, Goodness-of-fit tests, Tailweight discrimination.

1 Introduction

A relevant problem in statistics is to test whether some given observations follow one of two possible probability distributions. As a matter of fact, in many fields of application requiring frequency modeling, practitioners are faced with the problem of selecting a suitable statistical frequency distribution to fit their data. This problem can be considered a simple hypothesis testing as both the two hypotheses consist of a single distribution. The problem of choosing between frequency distributions, usually faced through the use of goodness-of-fit (gof) statistics or diagnostic plots [11], has been addressed early in the statistical literature. Cox ([7] and [8]) studied the problem of discriminating between the Lognormal and the Exponential distributions deriving an asymptotic probability distribution for the likelihood ratio statistic, thus allowing calculation of the "probability of correct selection" (PCS) given a large sample.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 392–401, 2011.

Similarly, [13], [23] and [24] proposed approaches based on the likelihood function. [43] and [44] provided regularity conditions and proofs for Cox's work. Many other researchers (among the others [6], [20] and [18]) have studied the asymptotic behavior of these tests when the number n of independent and identically distributed observations is sufficiently large. But, generally, two problems arise: very often, in practice, the sample size is not big enough to guarantee the asymptotic approximation to hold and, for many pairs of distributions, PCS approximations cannot be obtained in closed form. These two facts, in conjunction with the increase of computational power of personal computers, led some researchers to use Monte Carlo (MC) simulations to determine the PCS. Through simulation, and based on the likelihood ratio statistic, the case Lognormal vs Weibull (equivalent to Normal and Gumbel) was discussed by [13] and, more recently, by [12], while [3] covered the case Weibull vs Gamma. [23] studied the PCS for the pairs Weibull vs Lognormal, Weibull vs Gamma, and Gamma vs Lognormal and [31] studied the discrimination among generalized exponential, geometric extreme exponential and Weibull distributions.

In the financial and insurance risk design, the heaviness of the tail of the underlying distribution is crucial for the calculations. However, although it seems straightforward theoretically to distinguish between exponential tails and power tails, this requires unexpectedly large samples in practice [19]. One of the issues of risk management is the choice of the distribution of asset returns. Academics and practitioners have assumed for a long time (for more than three decades) that the distribution of asset returns is a Gaussian distribution. Such an assumption has been used in many fields of finance: building optimal portfolio, pricing and hedging derivatives and managing risks. However, real financial data tend to exhibit extreme price changes such as stock market crashes that seem incompatible with the assumption of normality. In the last years, important advances have been made in modeling credit risk at the portfolio level [17]. [5] discuss the optimal portfolio selection in a Value-at-Risk framework under alternative parametric distributions. [28] shows how extreme value theory can be useful to know more precisely the characteristics of the distribution of asset returns and finally help to chose a better model by focusing on the tails of the distribution. Also in these cases, a PCS for the tail model can't be obtained in closed form and therefore a simulation process must be implemented. In this work a computing intensive procedure suggesting the use of artificial neural networks as a instrument to discriminate between two alternative distributions is proposed. Indeed, neural networks have emerged as an important tool for classification [45]. The recent vast research activities in neural classification have established that neural networks are a promising alternative to various conventional classification methods due to various theoretical aspects. Neural networks are data driven self-adaptive methods: they can adjust themselves to data without any explicit specification of the underlying model; they are universal functional approximators, as they can approximate any function with arbitrary accuracy (see [21] and [22]); neural networks are nonlinear models, which makes them flexible in modeling real world complex relationships; neural networks are able to estimate the posterior probabilities, which provides the basis for establishing classification rule and performing statistical analyses [33].

The proposal is to take advantage of the aforesaid classificatory capabilities of feed-forward neural networks to test one sample to come from a normal distribution

against a "close to normal" alternative, comparing the performance of neural networks as gof tests to the results obtained by Taroni [38]. In his work, in fact, Taroni compared the power of eight traditional gof tests ([41], [15], [10], [35], [32],[30], [2] and [42]) used to check the normality of samples of different sizes (n = 20, n = 50 and n = 100) against the twelve "close to normal" alternatives (in the subsequent simply recalled as "alternatives") proposed by [37] and [16] at a significance level $\alpha = 0.1$. The use of batteries of tests to check the power of various methods on a wide range of alternatives is quite common (see, among the others: [34], [39], [25]) but the common conclusion is that a most powerful goodness of fit test does not exist. On the contrary, the results achieved in this work show that the power of a neural networks based gof test is the highest (with one small exception) against all the classic gof tests used by Taroni for all the close to normal alternatives and for any sample size considered.

2 A Neural Network-Based Goodness of Fit Test

Suppose you want to test if a sample $x = (x_1, x_2, \ldots, x_n)$ is coming from a normal distribution or from an alternative distribution chosen in the set of variables used by [37] and [16]. Assumed X to be the set of random variables defined in the unit interval $I = [0,1)$ with a continuous and almost everywhere differentiable Cumulative Density Function G_X, let Φ be the distribution function of the standard normal and U the uniform random variable with support in I. The random variable Y obtained by the transformation (where the symbol $^{\leftarrow}$ denotes the inverse function operator):

$$Y = \Phi^{\leftarrow}(G_X^{\leftarrow}(u)) \tag{1}$$

is normal if and only if $G_X(u) = u$. Alternatives are characterized by the Cumulative Density functions (CDF) $G(x)$ or Probability Density Functions $g(x)$ given in Table 1.

Table 1. Distributions alternative to the normal one

Alternative	CDF $G(x)$ or PDF $g(x)$	Parameters values
1A - 1B	$G(x) = 1 - (1-x)^k \quad 0 \le x \le 1$	1A: k = 1.5 1B: k = 2
2A - 2B	$G(x) = \begin{cases} 2^{k-1}x^k & 0 \le x \le .5 \\ 1 - 2^{k-1}(1-x)^k & .5 \le x \le 1 \end{cases}$	2A: k = 1.5 2B: k = 2
3A - 3B	$G(x) = \begin{cases} .5 - 2^{k-1}(.5-x)^k & 0 \le x \le .5 \\ .5 + 2^{k-1}(.5-x)^k & .5 \le x \le 1 \end{cases}$	3A: k = 1.5 3B: k = 2
4	$g(x) = 1.2 - 0.4x \quad 0 \le x \le 1$	
5A - 5E	$g(x) = \begin{cases} \frac{1}{1+ad} & 0 \le x \le b \\ \frac{1+a}{1+ad} & b < x \le b+d \\ \frac{1}{1+ad} & b+d < x \le 1 \end{cases}$	5A: a = .5, b = 0, d = .5 5B: a = .5, b = .25, d = .5 5C: a = .5, b = 0.1, d = .35 5D: a = –0.3, b = 0.1, d = .35 5E: a = –0.3, b = 0.1, d = .2

All of the alternatives considered, whose skewness and kurtosis indices are shown in Table 2, according to [36] would be classified, except for 3A and 3B, as close to normal.

Consider a feed-forward neural network made of n input neurons, one hidden layer made of h neurons and one single neuron as an output layer. The neural network output S_n is a non linear combination of the hidden neurons outputs that are, themselves, non linear combinations of the ordered statistic $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ input. Let **x** be a random sample of size n coming from a normal distribution or an alternative distribution, both having the same unknown mean and variance. The procedure developed to test if **x** is coming from a normal distribution (H_0) or an alternative one (H_1) is schematized in Table 3 in the form of a pseudocode.

Table 2. Indices of skewness (γ_1) and kurtosis (γ_2) for the alternatives

	1A	**1B**	**2A**	**2B**	**3A**	**3B**	**4**	**5A**	**5B**	**5C**	**5D**	**5E**
γ_1	-0.08	-0.14	0.00	0.00	0.00	0.00	0.05	0.16	0.00	0.22	-0.17	-0.12
γ_2	3.03	3.06	3.24	3.43	2.38	2.07	3.02	3.10	3.51	3.14	2.99	3.09

There are a few key points to be outlined and explained:

1. The procedure herein proposed is not affected by the well known "neural over fitting" as all the datasets are randomly generated for an arbitrary large number of input and output vectors making it virtually impossible for the network to memorize them. Therefore this procedure can be applied also with small samples as the simulation process is done using and arbitrarily large number of new randomly generated samples.
2. The Levenberg-Marquardt (LM) algorithm has been chosen for the network training because of two main reasons: first it gave better results than other traditional the results training algorithms (e.g. Backpropagation or Conjugate Gradients); second because it has interesting theoretical properties which apply fruitfully in this context. Whereas the first reason is simply an opportunity choice based on the results achieved, the second one has a strong mathematical background. This algorithm is a specific method for the minimization of least squares and comes as a combination of Gauss-Newton and the steepest gradient methods proposed by [27] and [29]. In particular it can be shown (among the others [4], page 291) that for a linear network the LM algorithm brings to the global minimum of the mean square error and for nonlinear networks this minimum is reached in the limit of an infinite dataset. Because of the way the datasets are herein generated (with a particular reference to the arbitrary large size of the training set), the use of LM algorithm, along with the mean squared error as a measure of error, gives the chance of training the non linear neural network approaching the mean square error lower limit with an arbitrarily high level of precision, without the overfitting problem to hold.
3. The Glivenko-Cantelli Theorem (e.g. [40]) determines the asymptotic behavior of the empirical distribution function as the number of independent and identically distributed observations grows: the empirical distributions of the test statistic under

the null $(S_n|H_0)$ and the alternative $(S_n|H_1)$ hypotheses as well as their mixture S_n can be estimated with an arbitrarily high level of precision given by the number k of samples generated.

4. Another crucial aspect of the procedure is the use of sorted samples for the input vectors. This choice resulted to be fundamental to achieve the results herein presented. Realistically this action helps the network to reach the best weights structures during the training process, but further studies will be focused on this particular point.

Table 3. Pseudocode to use neural networks as a gof test

STEP	PSEUDOCODE
1	x = original sample set n (sample size) and k (number of samples per distribution) z = rand(n, k) A = sort(norminv(g(z)))
2	N = sort(normrnd(mean(z), var(z)^.5, [n, k]))
3	Input=[N,A] Target=[zeros(1,k),ones(1,k)] Total_set = shuffle[Input; Target]
4	Train_set = select[Total_set, t] Validation_set = select[Total_set, v] Test_set = select[Total_set, ts]
5	net = train.nn(Train_set, Validation_set, h, 'logsig', 'tansig', 'trainlm')
6	Test_set_null = select[Total_set, ts, Target = 0] Test_set_alt = select[Total_set, ts, Target = 1] Output_null = sim(net, Test_set_null) Output_alt = sim(net, Test_set_alt)
7	count = 0 set alpha for i=1:ts if Output_alt(i) <= quantile(Output_null(i)<= alpha) count=count+1 else count = count end; power = 1-count/ts
8	Output = sim(net, x) if Output <= quantile(Output_null(i)<= alpha)accept H0 else accept H1

3 Simulations

The simulation study compared twelve alternatives for three sample sizes: n = 20, 50 and 100. For any of these comparisons five datasets of k = 30,000 samples (t = 15,000, v = 5,000 z = 10,000) have been created according to the rules discussed in paragraph 3 and used to train five networks for an overall number of 180 networks. For each of these networks an estimate of the power of the test has been computed and compared to the highest values of the classic tests for each alternative (MPCT: Most Powerful Classic Test). The results, shown in Table 4, put in evidence that the

proposal is always (only in case it is not but only for a little) the most powerful test in the battery. The architecture of the networks used to obtain the results shown in Table 4 has been chosen after training a large number of networks with different settings. The neural literature (e.g. [4]) states that clear and universal criteria for the choice of neural networks architectures do not exist. In this work the search of the best performing network has been restricted to the class of multi-layer perceptron having two layers of weights with full connectivity between adjacent layers and no direct input-output connections, a space of possible architectures which gave good results in classificatory problems ([45]). The results shown in Table 4 have been achieved for a number of hidden neurons equal to 8 for n = 20, 20 for n = 50 and 40 for n = 100.

Table 4. Estimates of the power of the gof test at the level α = 10% and n = 20, 50 and 100

		1A	1B	2A	2B	3A	3B	4	5A	5B	5C	5D	5E
n = 20	**MPCT**	**12**	**13**	**14**	**16**	**26**	**49**	**13**	**14**	**18**	**19**	**14**	**14**
	Net 1	13,5	15,1	16,0	21,1	49,8	83,2	12,0	21,1	27,8	33,7	26,8	22,7
	Net 2	12,9	14,9	15,8	20,5	49,4	82,9	11,8	20,4	27,0	32,7	26,8	21,1
	Net 3	12,7	14,9	15,6	20,5	49,2	81,4	11,6	20,3	26,6	32,6	26,4	21,0
	Net 4	12,3	14,8	15,5	20,4	48,9	81,2	11,5	19,9	25,9	32,3	26,4	20,9
	Net 5	12,1	14,7	15,5	19,4	48,6	77,5	11,5	19,5	25,3	32,1	26,2	20,6
n=50	**MPCT**	**12**	**13**	**15**	**21**	**45**	**90**	**11**	**17**	**25**	**25**	**19**	**16**
	Net 1	14,8	19,1	20,0	29,2	80,9	99,3	12,8	31,2	43,4	53,6	42,2	33,3
	Net 2	14,1	18,2	19,0	28,4	79,6	99,2	12,8	30,6	42,1	53,4	42,0	33,1
	Net 3	13,9	18,1	19,0	28,1	79,3	99,1	11,8	30,0	41,8	52,1	41,6	32,2
	Net 4	13,6	18,1	18,9	27,5	78,1	98,8	11,1	29,4	41,5	51,7	41,2	32,1
	Net 5	13,5	17,2	18,4	27,1	77,8	98,1	10,9	29,2	40,8	51,1	41,0	31,9
n=100	**MPCT**	**12**	**14**	**18**	**26**	**77**	**100**	**12**	**22**	**31**	**36**	**27**	**19**
	Net 1	16,2	22,0	25,1	39,8	96,8	100,0	15,6	43,4	62,3	73,7	61,6	50,5
	Net 2	15,7	21,7	23,1	39,3	96,4	100,0	14,4	41,2	61,4	72,9	61,1	48,7
	Net 3	13,9	21,0	22,1	37,2	96,4	100,0	13,9	40,0	60,7	72,8	60,1	46,4
	Net 4	13,8	20,1	21,7	35,6	96,4	100,0	13,2	39,8	60,4	72,4	59,6	46,4
	Net 5	13,7	19,7	21,2	34,7	95,3	100,0	12,7	38,8	60,3	71,3	59,5	44,8

As regards the choice of transfer functions, a good performance has been recorded using tan-sigmoid functions for the output neurons of the hidden layer and a log-sigmoid function for the single output neuron The choice of a logistic sigmoid output function has the direct effect of bounding the output into the interval [0, 1]. If the neural network is an unbiased discriminator of the two groups, the distribution of the network output S_n is symmetric around 0.5 and platykurtic whereas the distributions of the two conditional distributions $S_n|H_0$ and $S_n|H_1$ is leptokurtic with the same opposing skewness index (positive the former, negative the ladder). In case of perfect

classification, the distribution of the test statistic should have only two mass points (0 and 1) with equal occurrences. In order to explore these properties, and in particular the neural test consistency [26], consider Figure 1 in which the distributions $S_n|H_0$ and $S_n|H_1$ are shown for a selection alternatives (1A, 3A and 3C) and different values of the sample size. Growing n the two conditional distributions $S_n|H_0$ and $S_n|H_1$ tend to separate one another, in particular in situations of higher test powers.

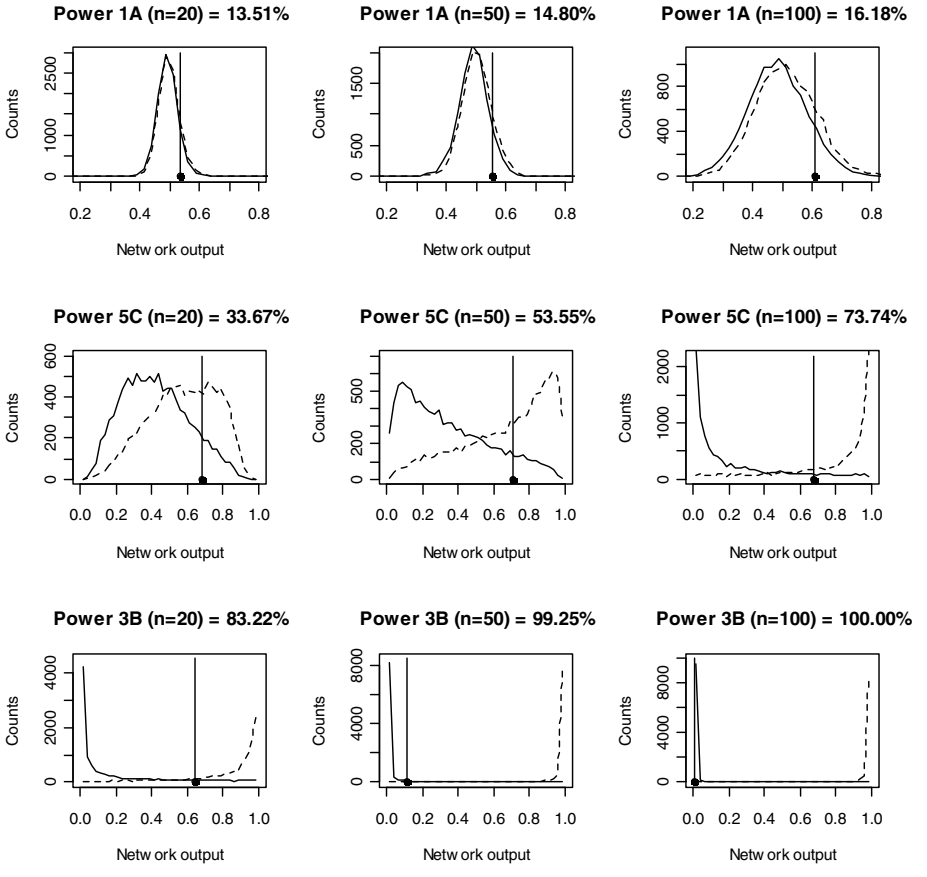


Fig. 1. Distributions of the test statistics under H_0 and H_1 (network output) for a selection of alternatives and n=20, 50 and 100

4 Conclusions

As already stated in the introduction, the problem of choosing between a strict number of models is very common in many different fields and in particular in financial and insurance risk analysis and in lifetime models. The returns on most financial assets exhibit kurtosis and many also have probability distributions that

possess skewness as well (see, among the others, [1]). The methodology herein presented allows to check various alternative distributions and to build the empirical distribution of the goodness of fit statistic.

Although neural networks are well known classificatory tools, their use as goodness of fit test has not been proposed yet in literature. The most relevant result achieved in this work is that neural networks were the uniformly most powerful tests in the battery analyzed and the power levels reached are much higher than the highest powers of the classic tests. This kind of the result has never been achieved using classic goodness-of-fit test as they are generally the most powerful on a specific class of alternatives but not on all the alternatives. Future works will explore other alternatives in order to see if the results herein achieved can be generalized to a wider class of alternatives, in particular the ones more relating the tails modeling in risk and insurance analysis.

In this work, in order to have a clean vision of the usability of neural networks as a gof test, a very basic implementation of feed-forward neural networks has been done in order to compare the results herein presented with results coming from strongly consolidated statistical techniques. But a number of developments, variants and technical studies can be considered for future works. First of all this work can be easily generalized to more than two alternative distributions. Using a network output with two neurons, it would possible to have one null distribution and three alternatives, for example setting (0, 0) as a target value for the null distribution and (1,0), (0,1) and (1,1) for the alternatives. Secondly this technique can be applied to the more specific discriminations, such as the ones recalled in the introduction and in particular relating the Lognormal distribution vs Gamma, Exponential, Weibull and Normal distributions. There are also some technical aspects which may be studied more in depth such as the exact role of the sample ordering in the input vector on the network performance, the use of different feed-forward classes of networks and the use of self-organizing maps (Kohonen maps) to this specific problem, in particular in the multi-alternative framework. Finally, as we herein stated the goodness of fit test as a binary classification problem, it is possible to compare the results given in Table 4 with the ones obtained through the application of support vector machines (e.g. [9]).

References

1. Adcock, C.J.: Asset pricing and portfolio selection based on the multivariate extended skew-Student-t distribution. Annals of Operations Research 176(1), 221–234 (2010)
2. Anderson, T.W., Darling, D.A.: A test of goodness of fit –. Journal of the American Statistical Association 49, 765–769 (1954)
3. Bain, L.J., Engelhardt, M.: Probability of correct selection of Weibull versus gamma based on likelihood ratio. Communication in Statistics: Theory and Methods 9, 375–381 (1980)
4. Bishop, C.M.: Neural Networks for pattern recognition. Oxford University Press, Oxford (1985)
5. Campbell, R., Huisman, R., Koedijk, K.: Optimal portfolio selection in a Value-at-Risk framework. Journal of Banking & Finance 25(9), 1789–1804 (2001)
6. Chernoff, H.: A Measure of Asymptotic Efficiency for Tests of a Hypothesis based on the Sum of Observations. Annals of Mathematical Statistics 23, 493–507 (1952)

7. Cox, D.R.: Tests of separate families of hypotheses. In: Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability, pp. 105–123. University of California Press, Berkeley (1961)
8. Cox, D.R.: Further results on tests of separate families of hypotheses. Journal of Royal Statistical Society 24, 406–424 (1962)
9. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)
10. D'Agostino, R.B.: An omnibus test for normality for moderate and large size samples. Biometrika 58(2), 341–348 (1971)
11. D'Agostino, R.B., Stephens, M.A.: Goodness-of-fit techniques. Marcel Dekker, New York (1986)
12. Dey, A.K., Kundu, D.: Discriminating Among the Log-Normal, Weibull, and Generalized Exponential Distributions. IEEE Transactions on reliability 58(3), 416–424 (2009)
13. Dumonceaux, R., Antle, C.E.: Discrimination between the lognormal and the Weibull distributions. Technometrics 15(4), 923–926 (1973)
14. Dumonceaux, R., Antle, C.E., Haas, G.: Likelihood ratio test for discrimination between two models with unknown location and scale parameters. Technometrics 15(1), 19–27 (1973)
15. Filliben, J.R.: The probability plot correlation coefficient test for normality. Technometrics 17, 111–117 (1975)
16. Frosini, B.: Distribution and power of a new goodness of fit statistic. Statistica 3, 389–413 (1983)
17. Gordy, M.B.: A comparative anatomy of credit risk models. Journal of Banking & Finance 24(1-2), 119–149 (2000)
18. Han, T.S., Kobayashi, K.: The strong converse theorem for hypothesis testing. IEEE Transactions on Information Theory 35, 178–180 (1989)
19. Heyde, C., Khanhav, A.: On the Problem of Discriminating between the Tails of Distributions. In: Contributions to Probability and Statistics: Applications and Challenges, Proceedings of the International Statistics Workshop, University of Canberra, April 4 - 5, pp. 246–258 (2005)
20. Hoeffding, W.: Asymptotically optimal test for multinomial distributions. Annals of Mathematical Statistics 36, 369–401 (1965)
21. Hornik, K.: Approximation capabilities of multilayer feedforward networks. Neural Networks 4, 251–257 (1991)
22. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366 (1989)
23. Kappenman, R.F.: On a method for selecting a distributional model. Communication in Statistics: Theory and Methods 11, 663–672 (1982)
24. Kappenman, R.F.: A simple method for choosing between the lognormal and Weibull models. Statistics & Probability Letters 7, 123–126 (1989)
25. Landry, L., Leparge, Y.: Empirical behavior of some tests for normality. Communications in Statistics: Simulation and Computation 21, 971–999 (1992)
26. Lehman, E.L., Romano, J.P.: Testing Statistical Hypotheses. Springer, Heidelberg (2005)
27. Levenberg, K.: A Method for the Solution of Certain Non-Linear Problems in Least Squares. The Quarterly of Applied Mathematics 2, 164–168 (1944)
28. Longin, F.: The choice of the distribution of asset returns: How extreme value theory can help? Journal of Banking & Finance 29(4), 1017–1035 (2005)
29. Marquardt, D.: An Algorithm for Least-Squares Estimation of Nonlinear Parameters. SIAM Journal on Applied Mathematics 11(2), 431–441 (1963)

30. Pace, L.: Un test per la verifica dell'ipotesi funzionale complessa con particolare riferimento al controllo della normalità. Rivista di Statistica Applicata 4, 299–308 (1983)
31. Pakyari, R.: Discriminating between generalized exponential, geometric extreme exponential and Weibull distributions. Journal of statistical computation and simulation 80(12), 1403–1412 (2010)
32. Pesarin, F.: An asymptotically distribution free goodness-of-fit test for statistical distribution depending on two parameters. In: Taillie, C. (ed.) Proc. NATO Adv. Study Inst., Statistical Distributions in Scientific Work, rieste/Italy, vol. 5, pp. 51–56 (1981)
33. Richard, M.D., Lippmann, R.: Neural network classifiers estimate Bayesian a posteriori probabilities. Neural Computing 3, 461–483 (1991)
34. Seier, E.: Comparison of tests for univariate normality. Inter-Stat (London) 1, 1–17 (2002)
35. Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika 52(3-4), 591–611 (1965)
36. Shapiro, S.S., Wilk, M.B., Chen, H.J.: A comparative study of various tests for normality. Journal of the American Statistical Association 63, 1343–1372 (1968)
37. Stephens, M.A.: EDF Statistics for goodness of fit and some comparison. Journal of the American Statistical Association 69, 730–737 (1974)
38. Taroni, G.: Studio comparativo dei test di normalità per alternative prossime alla normale. Statistica applicata 9(2), 231–246 (1997)
39. Thode Jr., H.C., Smith, L.A., Finch, S.J.: Power of tests of normality for detecting scale contaminated normal samples. Communications in Statistics - Simulation and Computation 12, 675–695 (1983)
40. Van der Vaart, A.W.: Asymptotic Statistics. Cambridge Series in Probabilistic Mathematics (1998)
41. Vasicek, O.: A test for normality based on sample entropy. Journal of the Royal Statistical Society B 38, 54–59 (1976)
42. Watson, G.S.: Goodness of fit tests on a circle. Biometrika 48, 57–63 (1961)
43. White, H.: Maximum likelihood estimation of misspecified models. Econometrica 50(1), 1–25 (1982a)
44. White, H.: Regularity conditions for Cox's test of non-nested hypotheses. Journal of Econometrics 19, 301–318 (1982b)
45. Zhang, G.P.: Neural Networks for Classification: A Survey. IEEE Transactions on systems, man and cybernetics – Part C: Applications and reviews 30(4), 451–462 (2000)

Interaction Data Management

Benedikt Schmidt and Eicke Godehardt

SAP Research, Bleistraße 8, 64283 Darmstadt, Germany
{benedikt.schmidt,eicke.godehardt}@sap.com

Abstract. Interaction data externalizes activities between an actor and an object, e.g. between a user and a computer system. The data is an important information source for user support and security applications. Nevertheless, generally little attention is given to interaction data. This paper emphasizes the role of interaction data usage for applications in the domain of human computer interaction and security. A consequent requirement is the clear and concise description of interaction data that enables the comparison of different systems that use interaction data. A scheme for classifying and describing processes related to the handling of interaction data is proposed.

The key contribution of this paper is a classification scheme for interaction data and the description of three core processes required for interaction data management, namely interaction data collection, interaction data processing and interaction data organization. The scheme and the three processes are used to describe a concrete system for interaction data management at the computer desktop.

Keywords: human computer interaction, interaction data, interaction history, context-awareness, user support.

1 Introduction

Interaction data provides information about the interaction process of one or more actors with an object. The data represents externalized interaction activities and is an important provider of context information regarding the actor or the object. Although interaction data is used in many applications there is generally little attention given to the processes related to the generation of used interaction data. Often the use of the data is in the focus of attention, but the structure and the attributes of the interaction data are out of scope [10,4]. The scope of this paper is highlighting the relevance of interaction data and hinting to required information to make interaction data comparable.

An important application area for interaction data is the support of human-computer interaction (HCI) [4,11]. Exemplary usage of interaction data in the HCI domain is a record of user operations that enables undo and redo functionalities of a software. Other application areas are in the domain of security monitoring [1], e.g. fraud or intrusion detection.

In the following, interaction data is considered as a continuous data stream of representations of interaction events. The data stream is generated based on

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 402–409, 2011.

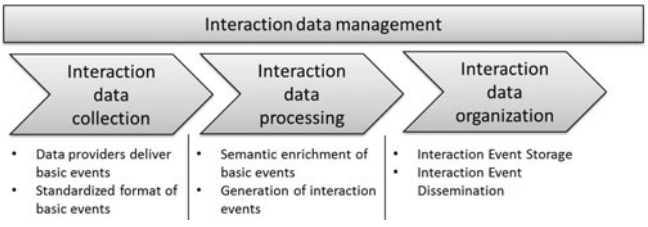


Fig. 1. Interaction data management processes

observations of interactions. An important usage of interaction data is the ex-post analysis. For that reason, interaction data often is referred to as interaction history [4,5]. Here, the term interaction data is used to include an immediate use of that data. Interaction data may be consumed directly or ex-post.

The rest of this paper is organized as follows. First, interaction data management is described by giving a classification scheme and describing three core processes of interaction data management. The processes are interaction data collection, interaction data processing and interaction data organization. Second, the generic discussion of interaction data is accompanied by a real example of interaction data management at the computer workplace. The paper concludes with summary and outlook.

2 Interaction Data Management

Generally, two types of systems that generate interaction data shall be distinguished: systems that generate interaction data and provide it as a service, and systems that generate interaction data for a direct consumption. To our knowledge, the only systems that generate interaction data as service are research prototypes. Examples for these prototypes are the Kimura system [11], the UMEA system [4], the APOSDLE monitor [6], or the monitoring of the UICO system [9]. The generation of interaction data for direct processing is included in many applications, e.g., lists of recently used files or undo and redo operations.

In the following, the aspect of a system that takes care of interaction data is named interaction data management system. For interaction data management, a classification scheme is given and three core processes are described (see figure 1). The processes are: interaction data collection, interaction data processing and interaction data organization.

2.1 Interaction Data Classification

Talking about interaction data implicitly means talking about a set of events following an event representation notation. Generally, an event is a significant change in the state of the universe [2], resulting in an infinite number of possible events based on time as inherent dimension of the universe [3]. Following the

general definition by Chandy, a specification for interaction events is possible. We define interaction events as those state changes that stand for interactions of an actor with a specified object in an environment. The following descriptions hold for the definition:

- Actor: The actor triggers an interaction. An actor can be a system, a natural person or a group of natural persons.
- Object: The object is the thing an actor interacts with. The object can be any physical entity.
- Interaction: An interaction is any kind of activity that is executed by an actor towards an object. The activity not necessarily needs to be reasonable but may be accidental or unplanned.
- Environment: The environment is a spatial or logical limitation for the interactions that are considered as relevant. The environment can be a building or a country, but as well a computer system or even an application.

To classify an interaction data management system it is necessary to describe object, actor, interaction and environment for the interaction events managed by the system. Focusing on the description of interaction events and not on the system purpose follows an important insight: interaction data management systems that are used as service providers have no predefined purpose but are defined by the events they generate.

2.2 Interaction Data Collection

Data providers deliver information related to the favored interaction data. The providers observe the environment interactions take place in and generate basic events as discrete instances of a (possibly continuous) signal [3]. Observation of the environment may be realized based on physical sensors (e.g., temperature), virtual sensors (e.g. software hubs to monitor keyboard interaction) or information streams (e.g., RSS feeds created by a group of actors). All events need to follow a similar standard that should comprise a timestamp and descriptive parameters. General rules that apply are [3]:

- A timestamp may be just one point (point semantics of time) or an interval (interval semantics of time).
- Parameters may be absolute or deltas relative to older reference values.
- Each basic event may contain additional attributes.

The resulting events are input for the interaction data processing.

2.3 Interaction Data Processing

Interaction data processing combines and semantically enriches the basic events delivered by the data providers. In most cases, the data providers will not deliver data that represents the interaction events one is interested in. For example a system shall identify if a user is busy at the workplace, interacting with the office environment. Although the system may use different sensor information related

to the activities of the user (like moving the mouse or sitting on a chair), the required semantics about being busy or not are only implicitly included in observable events. Additionally, the identification of interactions one is interested may only be identifiable based on multiple events (e.g. to state that a user is not busy, information about his work at the computer is not sufficient, additional aspects are required like usage of the telephone and talking in the room). Consequently, a gap between the desired interaction events and the detectable events may occur that requires interaction data processing.

Complex event processing [7] is one suitable method to process the event streams, especially if rules can be generated to build interaction events based on the basic events delivered by the data providers. Another example is classification based on models trained by machine learning. Still, the technique that enriches the events with additional semantics needs to be identified case-by-case.

2.4 Interaction Data Organization

The organization of interaction data as the final step comprises the data storage and data dissemination. The storage of interaction data provides access to interaction histories and enables the ex-post analysis of interaction data. Databases or structured texts like XML are possible storing methods. Dissemination is a service of an interaction data management system. So, dissemination addresses the access to stored interaction data as well as the instant forwarding of interaction data to subscribers. Access rights management is a requirement for the interaction data organization, as interaction data as user information needs to be treated confidential.

3 Interaction Data Management at the Computer Workplace

Based on the scheme and process description given in the upper section, an interaction data management system is presented that identifies interactions at the computer workplace. The presented system is an extension of the APOSDLE monitor application [6]. The system has been implemented as a service that can be subscribed by other applications to use the extracted interaction data.

3.1 System Classification

Purpose of the system is the identification of human computer interactions in terms of the desktop metaphor [8]. That means that all user operations shall be captured on a granularity level of objects (application, file, folder, information object, window) and activities (open, close, rename, delete, cut, paste, print, create, execute, focus) that belong to the desktop metaphor. The actor is the user who interacts with the computer system. The environment is a personal computer with Windows 7 operating system and different standard applications, e.g. the Microsoft Office Suite 2010. Object will be the computer desktop with all running applications and accessible information objects. In the following the above defined three processes implemented by the system are presented.

3.2 Interaction Data Collection

The computer workplace is mediated by the operating system that organizes system input and output. System monitoring benefits from the mediating role of the operating system. Sources for monitoring functionalities may be frameworks to organize data exchange between applications (e.g. the interoperability libraries for windows) and accessibility features (e.g. the UI Automation Framework for windows). For the windows operating system monitors for the following elements can be created by using functionalities of the operating system:

- **Input and Output Devices**
  - Mouse: Mouse movement, mouse wheel operations and mouse clicks is captured by a mouse hub.
  - Keyboard: The Keyboard input stream is captured by a keyboard hub.
  - Printer: Printed documents are captured.
  - Webcam: The webcam image is captured.
  - Sound: The sounds played by the system are captured.
  - Display: The displayed image is captured. The image contains the user interface and can be input, e.g. to OCR to identify the shown text.
- **System Management**
  - Process: Information about running processes can be accessed, partly standing for applications. For each process detailed information can also be accessed, like windows belonging to a process and files that are locked by a process. The most valuable information is the focus window, identifying the process a user is actively working with.
  - Filesystem: The files accessed during the work of a user, including information like creation and modification date can be accessed. Due to many file accesses that are not directly triggered by the user (e.g. scanning of antivirus programs or backup operations), the data can be noisy.
  - Clipboard: The clipboard is used to copy and paste different types of data without application borders. The clipboard data shows which data is transferred by a user between different application contexts.
- **Accessibility Features**
  - Accessibility features provide access to data structures that represent all running instances and potentially visible elements that exist on the operating system. An example is the UI automation framework which provides a tree representation of the user desktop. Each visualized element is part of the tree and may support a set of patterns which can be used to interact with the visualized element (e.g. query displayed text).
- **Special Applications**
  - Application APIs enable the subscription of the events generated by a running instance of a program.

- Application APIs may provide access to the information objects that are displayed by the application. An example for such information objects is content of a textfile displayed by a word processor or the content of website displayed by a web browser.

For interaction data collection a combination of the described sensors has been realized. Process information are used together with mouse and keyboard hubs to trigger data requests. Once a process change, an enter-hit or a mouse click is identified dedicated application monitors are used together with the UI automation framework to extract information about the interaction. The extracted events include information about the interaction type, the process and the user interface element a user interacted with. For many applications the displayed content is extracted, too.

3.3 Interaction Data Processing

The collected events represent the user interactions on a very detailed level. Examples for the captured events are clicks on buttons, selection of text entries in lists and switches between applications. To identify the interaction by means of the desktop metaphor additional processing of the data is required.

Table 1. Desktop operations: possible pairs of operation (OPR) and object (OBJ)

OPR \ OBJ	Application	File	Folder	Information Object	Window
Open	x	x	x		
Close	x	x	x		
Save		x			
Rename		x	x		
Delete		x	x	x	
Cut		x	x	x	
Paste		x	x	x	
Print		x			
Create		x	x	x	
Execute	x				
Focus			x		x

A set of favored events that represent user interaction based on the desktop metaphor has been developed. The events are combinations of operations and objects (see table 1). The challenge of the interaction data processing is the generation of the favored interaction events based on the collected events. The system uses rules to process the collected events to the favored events. Overall, 98 rules were modeled, e.g., 15 to identify close operations. The rules have been implemented for Drools fusion[1], a complex event processing engine that supports temporal reasoning.

[1] http://www.jboss.org/drools/drools-fusion.html

3.4 Interaction Data Organization

The interaction data management system has been designed as a data provider for interaction-sensitive applications and as a data provider for ex-post interaction analysis. Therefore an event broadcasting and a data storage is integrated. The event broadcasting delivers the aggregated interaction events to event subscribers. The data storage is realized by a database or XML files. A scheme for the structure of events has been designed to include name, type and timestamp as required event attributes. The process/window, content and other extracted information are optional event attributes.

3.5 Benefits of the Description

For the interaction data management systems at the computer workplace a classification and a process description has been applied that follows the scheme presented in the previous section. Based on this information, the interaction data management system can now be compared with other existing interaction data management system at the computer workplace, e.g., with the UICO system [9]. Similar data is collected but the processing is different, as UICO focuses on the identification of a predefined set of functionalities applied on resources. This is an important information, once methods that use the interaction data are discussed. A comparison with the interaction data used by [10] or [4] is not possible, as not enough information about the interaction data management of the systems is provided.

4 Conclusion and Outlook

A classification scheme and the basic processes implemented by an interaction data management process have been defined. The elements have been used to describe an interaction data management system for the computer workplace. Thereby two aspects have been in the focus of this paper. On the one hand, stress of the relevance of interaction data. On the other hand, support of the clear and concise description of interaction data. These two aspects are discussed in the following.

Interaction data is an important aspect of context information. It can be used to identify individual attention, make assumptions about upcoming activities or to provide users or objects with a history. Currently, no standard data providers for interaction data exist. In the domain of desktop computing many research prototypes exist that use system libraries and frameworks of the operating system to create system-wide interaction histories. Building such tools is a tedious task that requires high adaptation effort, once new versions of applications or operating systems are available. Operating systems process many interaction related data, therefore the provision of interaction data management as an operating system service is a consequent step. In the domain of ubiquitous computing the selection of data providers for interaction data management is even more

complex, as no common architecture exists to collect interaction data. The perspective provided in this paper may support the creation of reusable patterns of interaction data management for the domain of ubiquitous computing.

The description of systems that use interaction data often lacks detail with respect to the type of interaction data and the applied processes. Nevertheless, a detailed description of interaction data has several advantages, as it supports the comparability of systems. Such comparability is beneficial for the evaluation of systems, as an additional dimension for the evaluation can be realized: the complexity and the amount of interaction data required to realize the system. An example for a comparison is the benchmarking of activity detection systems.

Acknowledgement. Work presented in this paper has been partly funded by the German Federal Ministry of Education and Research, grant no. 01IA08006.

References

1. Bishop, M.: A standard audit trail format. In: National Information Systems Security 1995 Proceedings: Making Security Real, p. 136. DIANE Publishing (1996)
2. Chandy, K.M.: Event-Driven Applications: Costs, Benefits and Design Approaches. Gartner Application Integration and Web Services Summit (2006)
3. Hinze, A., Sachs, K., Buchmann, A.: Event-Based Applications and Enabling Technologies. In: Proceedings of the Third ACM International Conference on Distributed Event-Based Systems (2009)
4. Kaptelinin, V.: UMEA: translating interaction histories into project contexts. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, vol. (5), pp. 353–360. ACM, New York (2003)
5. Kelly, S.U., Davis, J.P.: Desktop History: Time-based Interaction Summaries to Restore Context and Improve Data Access. Human-Computer Interaction (2003)
6. Lokaiczyk, R., Faatz, A., Beckhaus, A., Goertz, M.: Enhancing Just-in-Time E-Learning Through Machine Learning on Desktop Context Sensors. Context, 330–341 (2007)
7. Luckham, D.: The power of events: an introduction to complex event processing in distributed enterprise systems. Addison-Wesley Longman Publishing Co., Inc., Boston (2001)
8. Marcus, A.: Human communications issues in advanced UIs. Communications of the ACM 36(4), 100–109 (1993)
9. Rath, A.S.: UICO: An ontology-based user interaction context model for Automatic Task Detection on the Computer Desktop. In: CIAO 2009: Proceedings of the 1st Workshop on Context, Information and Ontologies, pp. 1–10 (2009)
10. Rattenbury, T., Canny, J.: CAAD: an automatic task support system. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 696–706. ACM, New York (2007)
11. Voida, S., Mynatt, E.D., Macintyre, B.: Supporting Activity in Desktop and Ubiquitous Computing (2007)

SWord: Semantic Annotations Revisited

Oleg Rostanin[1] and Passant al Agroudy[2]

[1] German Research Center for Artificial Intelligence (DFKI GmbH)
Oleg.Rostanin@dfki.de
[2] German University of Cairo
passant.el-agroudy@student.guc.edu.eg

Abstract. Using semantic annotations to support knowledge intensive office work is an important trend in knowledge management since the last decade. Although standards for semantic mark-up such as RDFa were introduced for annotating XML-based resources, a standardized approach to combining unstructured data from office desktop documents with highly-structured semantic annotations still does not exist.

This paper presents an add-in for MS Word allowing to seamlessly integrate semantic categorizations within MS Word documents. The SWord plugin assists user in categorizing Word documents and their parts by assigning them semantic categories from a lightweight knowledge base. Semantic categories are used for convenient navigation in large documents as well as means for better document organization and re-finding when using standard Windows Desktop Search tool.

1 Introduction

The problem of personal information organization was actively discussed in the knowledge management community in recent years. The project EPOS[1] introduced the idea of a personal semantic desktop. The assumption behind the semantic desktop is that knowledge workers would gradually create and maintain their personal information models to categorize information objects on their desktops. Proof-of-concept implementations of semantic desktop showed its feasibility but also discovered some essential deficiencies:

1. Poor integration in a habitual working environment hampers office workers from widely using it.
2. Necessity to create and to maintain the knowledge base (KB) is a heavy barrier to non-IT specialists.
3. Partial annotation of documents with concepts from KB is not possible.
4. Under the OS Windows that keeps leading positions in modern offices, searching for annotated documents by using their semantic categorization is not possible on the OS level and requires special semantic search tools.

In parallel, technologies supporting semantic annotation of document contents were developed. Unlike semantic desktop where documents are regarded as atomic

[1] http://www3.dfki.uni-kl.de/epos/

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 410–419, 2011.

resources that can be referenced by semantic models, semantic annotation allows to mark parts of a document with categories from KB thus enabling finer level information retrieval and more precise navigation in documents. Although standards for semantic mark-up such as RDFa[2] were introduced for describing XML-based resources, a standardized approach to integrating highly-structured semantic annotations into unstructured office desktop documents still does not exist. A large variety of document formats as well as rapid changes in the architecture of standard office software make the task of finding a universal solution extremely difficult (see [10]). After years of research, integrating annotations into office documents remains interesting and challenging task that the knowledge management community has to solve.

In this paper we present an extension for MS Word supporting personalized and collaborative annotations in Word documents. The SWord-plugin is seamlessly integrated into the habitual MS Word environment. It helps office workers to organize their documents by providing means to categorize documents with semantic categories and annotate its passages. SWord uses a lightweight KB build on the top of an intuitive collaborative concept mapping tool LeCoOnt. Documents annotated in SWord can be easily re-found by using standard desktop search tools as the annotations are reflected on the OS-level metadata.

The rest of the paper is build as follows. Section 2 presents related work on semantic document annotations. A short description of a realized prototype SWord is given in section 3. Real-life usage scenario is considered in section 4. Techniques used to realize a semantic document annotation are discussed in section 5. We conclude the paper with discussions of the future work in 6.

2 Related Work

Having documents well organized eases the task of re-finding and re-using them thus saving much time of office workers [4]. With the rush development of semantic web technologies, ontology-based methods for semantic document annotations and categorizations are getting popular [10]. One distinguishes 2 kinds of semantic annotation systems: those that consider documents as atomic resources and those annotating parts of documents and storing the metadata inside a document. Examples of systems belonging to the first category are different implementation of the semantic desktop idea [6],[8].

According to [10], passage-level annotations are more intuitive to knowledge workers who deal a lot with documents. Furthermore, they allow finer document retrieval and reuse. The largest vendors of the office software recognized this fact and made important steps providing developers with mechanisms for integrating semantic annotations into proprietary document formats:

– XMP-metadata introduced into PDF-format from Adobe allows to store semantic annotations as RDF[3]. Eriksson in [3] presents work on integrating RDF-data into PDF documents and referencing it from XMP-metadata.

[2] http://www.w3.org/TR/xhtml-rdfa-primer/, 2011-05-15

[3] http://www.pdfa.org/doku.php?id=artikel:en:pdfa_metadata, 2011-05-15.

- OpenOffice introduced RDF-based metadata and the programming API for manipulating and reading RDF annotations in the document[4].
- MS Office applications were switched from proprietary binary fromat to open XML-based formats allowing storage of structured annotation information as XML in the documents[5]. Fink et al. in [5] describe using XML-parts for storing semantic annotations for medical texts using medical ontologies.

In the current work, we focus on the audience of MS Word users by providing them tools for annotating entire documents as well as their parts using concepts from a personal or corporate KB. Open COM-based architecture of MS Word as well as SmartTag technology inspired many developers to create solutions integrating semantic annotations into Word documents [9],[2]. However, since MS Office 2010, Microsoft reduces support level of the popular SmartTag technology that was widely used to display semantic annotations and to interact with them[6]. Adhere our aim is to support implementing semantic mark-ups in MS Word without using the obsolete SmartTags technology.

3 SWord – Extension for MS Word

The ultimate goal of the SWord add-on is to provide an integrated environment assisting knowledge workers dealing with MS Word documents, like explaining notions, easy navigation in large documents, reusing text passages, diagrams and images. Below we introduce the notion of proactive information delivery that is essential to have in mind before reading about the rest of the approach.

3.1 Proactive Information Delivery

A distinguishing feature of SWord is the proactive information delivery (PID). Based on the current text selection in the document as well on the semantic annotations, SWord generates queries delivering local documents and e-mails relevant to the current selection. Although the PID functionality itself is out of scope for this paper, the work described here creates a solid basis allowing to realize such kind of assistance[7]. Adhere, in the 1st step, SWord has to provide means allowing to easily create and manage semantic document annotations.

3.2 Semantic Annotations

SWord supports user at semi-automatic document annotation with concepts from personal or shared KB (see section 3.3). It supports different granularity levels of annotations listed below.

[4] http://code.google.com/p/bungeni-editor/wiki/RDF_metadata_in_OpenOffice, 2011-05-15.
[5] http://msdn.microsoft.com/en-us/library/bb608618.aspx, 2011-05-15
[6] http://technet.microsoft.com/en-ca/library/cc179199.aspx, 2011-05-15.
[7] For more information on PID see [7].

Passage-level annotations. Passage-level annotation is creating an association between a concept and a certain part of the document. Currently in SWord, such part can be any selected text. Annotating images, shapes and tables is not supported yet. Passage-level annotation is used to easily access document parts related to a certain concept specified by the user. In the future, we are targeting to use passage-level semantic annotations to refine PID (see section 3.1). This approach would allow the reuse of document parts by providing them proactively or upon request during the document creation. In addition, considering spatial relations between the selected text and annotations (e.g. annotations included in the selected text are considered more important than annotations related to the passage including the selected text or document-level annotations) would even more refine PID results and increase their precision.

Document-level annotation. Document-level annotations relate concepts to the entire document rather than to a part of it. Document-level annotation can be used for fast, general classification of documents. On the other hand, it offers a quick, easy solution for users who feel that passage-level annotations are overwhelming and need more effort, time and expertise to perform.

Explicit Concept Highlighting. SWord allows to dynamically highlight exact occurrences of concept labels in the document text by presenting them as embedded hyperlinks. Hyperlinks can be used to see more information about concepts or to open the graphical concept browser LeCoOnt (see section 3.3). The user has a choice to convert dynamically found occurrences into semantically annotated expressions using the mechanism of passage-level annotations.

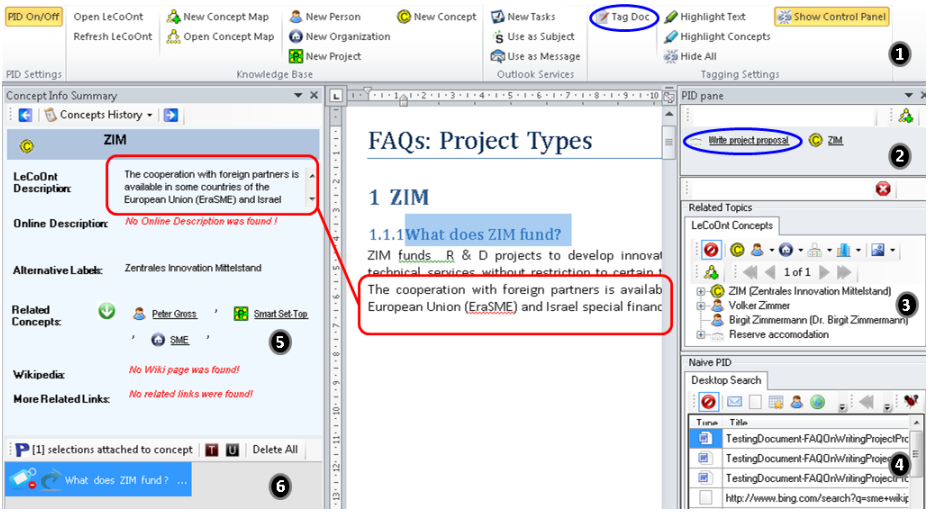


Fig. 1. SWord Overview Screenshot

3.3 Lightweight Knowledge Base

Creation and maintenance of formal ontologies required by semantic web technologies is regarded as a burden by the most users. To realize semantic annotations in MS Word we suggest using a lightweight concept-mapping tool LeCoOnt[8]. LeCoOnt has a simple data model: each concept there consists of a unique URI, a stereotype like "Person" or "Project", and a set of labels corresponding to "rdfs:label" or "skos:altLabel" in respective RDF notations. Concepts can be connected together by arbitrary relations (e.g. "is-a" or "manages") within concept-maps – meaningful views on KB. LeCoOnt supports both personal and collaborative models of KB by providing access rights mechanisms.

3.4 GUI-Components

Fig. 1 shows a screenshot of the SWord GUI consisting of the following logical components:

1. SWord Ribbon (fig. 1-1). Contains GUI-controls for interacting with SWord. These are divided into 2 logical groups: i)working with KB, like creating new concepts or concept maps; and ii)working with semantic text annotations.
2. Interactive Concept Cloud (fig. 1-2). Contains an overview of concepts used for annotating the current Word document or its passages. Clicking on a concept allows to see its details in the embedded concept browser (see below in this section) or in the graphical concept browser LeCoOnt (see section 3.3). The concept cloud contains a search field for manual concept search supported by embedded auto-completion.
3. Concept Suggestions (fig. 1-3). As mentioned in [10], providers of solutions for semantic document annotations have to solve the problem of manual document categorization as it is a tedious and an error-prone process. In SWord we rely on the ontology-based information extraction service iDocument [1] allowing to recognize ontological concepts in the text of documents. The contents of KB is periodically populated as an RDF dump and used by iDocument to extracting concepts. Retrieved concepts are presented in a tree form allowing user to navigate in KB.
4. PID panel(fig. 1-4). Displays documents relevant to the current text selection or to the entire document provided by PID (see section 3.1).
5. Embedded Concept Browser (fig. 1-5). This component displays the concept textual description and its alternative labels. Moreover, it is used to browse to Wikipedia link if available as well as other resources(persons email, contact info, web urls or documents) associated with the concept. Finally it displays other concepts associated with the current one in LeCoOnt concept maps.
6. Semantic Bookmarks Previewer (fig. 1-6). This component shows text passages annotated with the concept displayed in the embedded concept browser. It is mainly used for manipulating annotations and direct navigating to annotated passages in the document.

[8] http://lecoont.opendfki.de/

After this short description of SWord functionalities we illustrate their usage on a real-world example.

4 SWord Use Cases

The scenario takes place in a research institute *Smart R&D*. Researchers – senior (SR) to junior (JR) in *Smart R&D* deal with documents extensively for writing project proposals, reports and scientific papers. To teach a team of JRs, a SR composes a document about FAQ related to different funding programs targeted by *Smart R&D*. The document is publicly shared through the intranet of the company.

Scenario 1: Author The author creates a new MS Word document and annotates it with the concept "Write project proposal" from KB. For this purpose he selects "Tag Doc" option in the SWord Ribbon (fig. 1-1) to switch annotations into the document-level mode. He annotates the document by typing the first letters of the concept name in the "Concept cloud" (fig. 1-2) and using the autocomplete functionality. This annotation will help the SR in the future to easier find the document as the semantic annotation is reflected in document OS-level metadata that can be used later by a desktop search engine such as Windows Desktop Search. The FAQ-document is divided in to four main sections according to types of project funding programs. Each section contains several questions. The author starts writing them. For the first question, some quick brain storming is needed without wasting his time for searching. He consults SWord by selecting the question text with the mouse. Some concepts related to the question appear in the "Concept Suggestions" component of SWord (fig. 1-3). The author clicks on one of them ("ZIM") to see the concept detailed information in the "Concept browser" (fig. 1-5). The concept description from KB seems to be helpful, so he takes it over as an answer for the question. Because the concept is closely related to the question, he annotates the question with it by right clicking on it in "Concept Suggestions" (fig. 1-3) and selecting "attach" in the context menu. While writing the second question, the author mentions Peter Gross as the contact person for "ZIM" project and annotates him with a corresponding concept thus making his contact data directly available to readers. In the third question, he uses the abbreviation "SME"[9]. It is an important concept that the author is asked a lot about, so he decides to annotate it. However, the concept was neither suggested automatically nor found by searching manually. Therefore, he creates a new concept by clicking "New Concept" in the SWord Ribbon (fig. 1-1). The concept label is automatically set to the selected text, and the concept URI is auto generated (fig. 2). He fills further concept details and saves it. Now, when he re-selects the text "SME" in the document, he gets the newly created concept displayed in the "Concept Suggestions" (fig. 1-3) and annotates the text. When the SR is done with the document, he enables the read-only mode and shares it on the intranet.

[9] Stands for Small and Medium Enterprizes.

Scenario 2: Reader When the document is first opened, the JR wants to know what help was provided by the author. Therefore he chooses the "Highlight concepts" option in the SWord Ribbon (fig. 1-1). This highlights occurrences of concept labels in the document texts as bold hyperlinks referencing concepts in KB. Simple clicking on these hyperlinks refreshes the concept data in the "Concept Browser" (fig. 1-5) and updates the list of related bookmarks (fig. 1-6). For example, in (fig. 3-1), "Peter Gross" was changed to a bold hyperlink, because it is an exact textual match to concept "Peter Gross" and it is annotated with it. As the JR is interested in some more details on ZIM projects, he clicks on "Peter Gross" hyperlink. After that he sends him an email directly by clicking on "email" from "more related links" in the "Concept Browser".

Now the JR wants to see all annotated passages in the document, whether they are exact matches of concepts or not. Therefore, he clicks on "Highlight Text" in "Smart Office Ribbon" (fig. 1-1). Next, the reader wants to see an explanation of the concept "SME". He clicks on the hyperlink to read more about the concept (fig. 3-2) and to see related bookmarks attached to multiple parts of the document (fig. 3-3).

He navigates within the annotated passages by clicking the "arrow" button beside each bookmark in the "Semantic Bookmark Previewer" (fig. 3-3). The reader still wants to know more about "SME". Instead of searching for its Wikipedia page, he finds it attached in the concept info. He clicks on "Go to page" from "Wikipedia link" (fig. 3-4). After checking wiki link, he knows the expression is important so he wants to see concept occurrences in the document that were not explicitly annotated by the author. He clicks on the button "U" in the "Semantic Bookmarks Previewer" (fig. 3-3). As a result, he finds a lot of occurrences. To save his time, he right-clicks the concept in the "Concept Browser" (fig. 3-5) and selects "Attach All", to automatically annotate all textual occurrences of concept label and alternative labels. Having a quick look on the document, the reader realizes only a few questions concern him. Therefore, he decides to annotate all the related questions with a concept of his name. This simplifies finding necessary FAQ later.

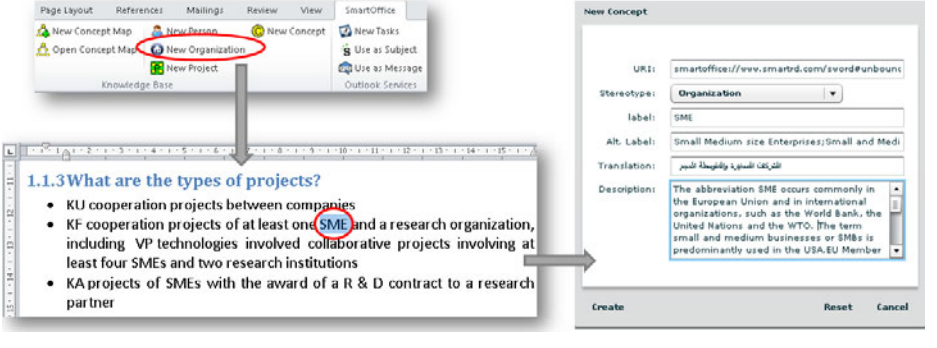


Fig. 2. Creating a new concept in LeCoOnt from SWord

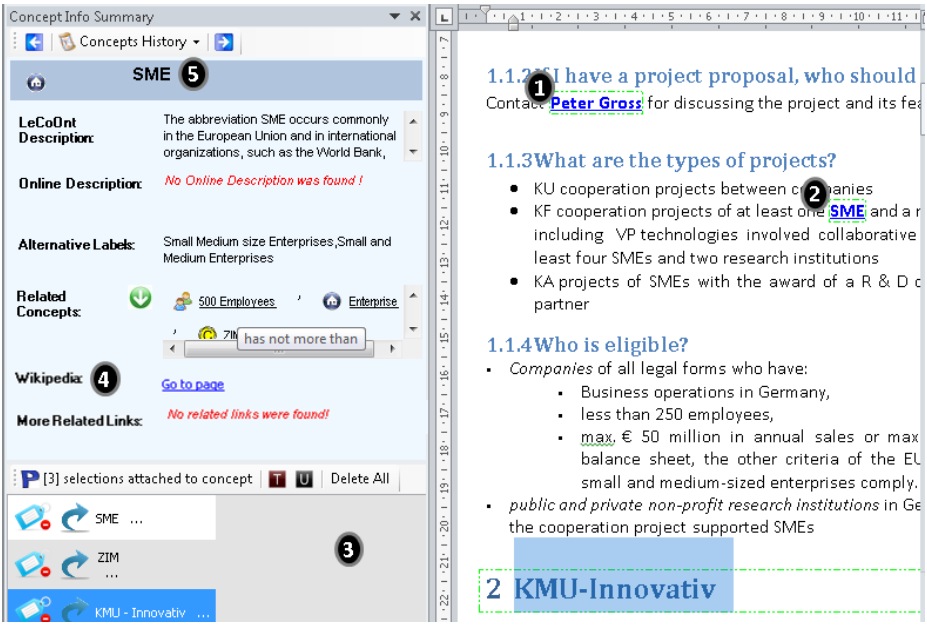


Fig. 3. Annotate all concept occurrences

5 Realizing Semantic Annotations in SWord

5.1 Data Structures and Algorithms to Support Annotations

Concepts in SWord are represented internally by so called *WeightedConceptInfo* objects delivered by LeCoOnt SOAP-based services. It contains concept information such as URI, its primary and alternative labels, informal description, a list of related external information resources and a list of other related concepts. Text passages are presented by their *Microsoft.Office.Interop.Word.Range* objects[10] (fig. 4-2). *Range* object is assigned a *Microsoft.Office.Interop.Word.Bookmark*[11] object (fig. 4-4) to be referenced in SWord. If the passage to be annotated is not assigned to a *Bookmark* object, a new *Bookmark* is created with unique-random-auto-generated name (fig. 4-3). However, if the passage is already bookmarked, then the existing *Bookmark* object is used to represent the range. Consequently, passage can be identified using its unique bookmark name. Therefore, an association between a concept and a document passage can be defined by mapping a concept URI and the bookmark name of the passage (fig. 4-5/6). Annotations are saved in a custom property of the Word document in the following format: *[uri1]:0:[List of Bookmark Names];...*, where *uri1* is the concept URI and *List Of bookmark names* is a comma-separated list of bookmark identifiers. For each

[10] http://msdn.microsoft.com/en-us/library/aa223066(v=office.11).aspx
[11] http://msdn.microsoft.com/en-us/library/aa221387(v=office.11).aspx

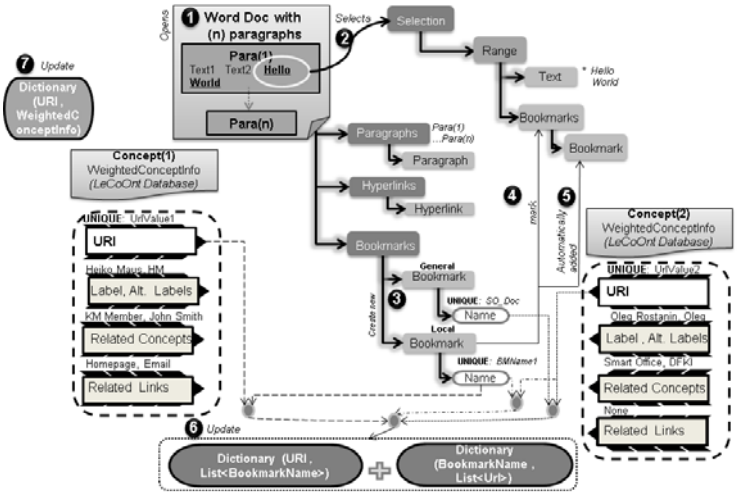


Fig. 4. Storing semantic annotations in SWord

document, there exists one global bookmark used for associating concepts with the entire document in case of document-level tagging. Semantic annotations are also reflected in OS-level file metadata field "Categories" thus allowing to search for documents without specialized search tools.

5.2 Loading and Updating of Semantic Meta-data

SWord internally uses 3 dictionaries to manipulate the semantic annotations in SWord. Dictionary 1 is used for mapping concept URIs in the document to their respective list of attached bookmarks (fig. 4-5). Dictionary 2 is used to map bookmarks in the document with THEIR respective list of attached concepts (fig. 4-5). Dictionary 3 is used to map the concepts URIs with the respective WeightedConceptInfo objects (fig. 4-6). The dictionaries are initialized when the word document is opened. They are updated with the user changes in documents and are saved before the document is closed.

6 Conclusion and Future Work

In this paper we introduced SWord, an extension for MS Word providing functionalities for semantic annotations within documents, supporting semantic bookmarks and a very basic proactive information assistance (PID). SWord is well integrated into the working environment so that we made an important step towards eliminating usage barriers common for knowledge management solutions. We also showed how semantic data are presented inside the word document and described their storage format.

In the future we'd like to extend our approach to support annotating images, tables and embedded objects. Another goal of our work is to proof the feasibility of the approach by conducting real user studies. Furthermore, new ways of using semantic annotations for knowledge work support have to be found to increase user motivation in exploiting them. Our hypothesis is here that PID can be improved significantly by considering users semantic passage- and document-level annotations for generating automatic queries. Another important research question would be the aspect of collaboration between persons working on the same documents.

References

1. Adrian, B., Klinkigt, M., Maus, H., Dengel, A.: Using idocument for document categorization in nepomuk social semantic desktop. In: Pellegrini, A.P.H.W.W.B.K.T.T. (ed.) Proceedings of I-KNOW 2009 and I-SEMANTICS 2009. International Conference on Semantic Systems (I-Semantics 2009), Graz, Austria, September 2-4. J.UCS Conference Proceedings Series, pp. 638–643. Verlag der Technischen Universitt Graz, Graz (September 2009) 5
2. Carr, L., Miles-Board, T., Woukeu, A., Wills, G., Hall, W.: The case for explicit knowledge in documents. In: Proceedings of the 2004 ACM Symposium on Document Engineering, DocEng 2004, pp. 90–98. ACM, New York (2004) 3
3. Eriksson, H.: The semantic-document approach to combining documents and ontologies. International Journal of Human-Computer Studies 65(7), 624–639 (2007); knowledge representation with ontologies: Present challenges - Future possibilities 2
4. Feldman, S., Duhl, J., Marobella, J.R., Crawford, A.: The hidden costs of information work. an idc white paper (2005) 2
5. Fink, J.L., Fernicola, P., Chandran, R., Parastatidis, S., Wade, A., Naim, O., Quinn, G.B., Bourne, P.E.: Word add-in for ontology recognition: semantic enrichment of scientific literature. BMC Bioinformatics (2010) 3
6. Iturrioz, J., Diaz, O., Anzuola, S.F.: Toward the semantic desktop: The semouse approach. IEEE Intelligent Systems 23, 24–31 (2008) 2
7. Rostanin, O., Maus, H., Suzuki, T., Maeda, K.: Using concept maps to improve proactive information delivery in tasknavigator. In: Setchi, R., Jordanov, I., Howlett, R., Jain, L. (eds.) KES 2010. LNCS, vol. 6276, pp. 639–648. Springer, Heidelberg (2010) 3
8. Sauermann, L., Dengel, A., van Elst, L., Lauer, A., Maus, H., Schwarz, S.: Personalization in the epos project. In: Proceedings of the Semantic Web Personalization Workshop at the ESWC 2006 Conference, pp. 42–52 (2006) 2
9. Tallis, M.: Semantic word processing for content authors. In: Handschuh, S., Koivunen, M.R., Dieng, R., Staab, S. (eds.) Proceedings of the Second International Conference on Knowledge Capture K-CAP 2003 on Knowledge Markup and Semantic Annotation, Sanibel, Florida, USA (2003) 3
10. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargasvera, M., Motta, E., Ciravegna, F.: Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semantics: Science, Services and Agents on the World Wide Web 4(1), 14–28 (2006) 2, 5

Using Suffix Arrays for Efficiently Recognition of Named Entities in Large Scale

Benjamin Adrian and Sven Schwarz

Knowledge Management Department, DFKI GmbH, Kaiserslautern, Germany
firstname.lastname@dfki.de

Abstract. In this paper, we present an efficient comparison of text and RDF data for recognizing named entities. Here, a named entity is a text sequence that refers to a URI reference within an RDF graph. We present suffix arrays as representation format for text and a relational database scheme to represent Semantic Web data. Using these representation facilities performs a named entity recognition in linear time complexity and without the requirement to hold names of existing entities in memory. Both is needed to implement a named entity recognition on the scale of for instance the DBpedia database.

1 Introduction

Today, even people without a background in computer science are part of the WWW community as they share their thoughts on Blogs, edit open encyclopedias in Wiki systems, or maintain their own homepages. By now, the common stereotype of a Web resource is most likely a Web page formatted in Hypertext Markup Language (HTML) for structuring and layouting especially its textual but also any other visual information. The incredible and confusing number of billions of published Web resources as well in addition to the necessary need of computers to consume Web resources make people desire more computer assistance for consuming information on the Web.

In the Semantic Web computers can be told about the information content of a Web resource by describing it formally in the Resource Description Framework (RDF) [1]. Web pages of the world's largest information providers such as New York Times, British Broadcasting Corporation, White House, German National Library, and Wikipedia show the great success when adding RDF data as machine-interpretable hints to published information. Actually, by using RDF in attributes (RDFa [2]) markup it is even possible to embed RDF data directly within a Web page's HTML model. In fact, Web content management systems such as Drupal or Wordpress support the creation of RDF and RDFa data, but still prose is of major use in news, education, entertainment, or science. What remains is the automatic creation of such RDFa markup for existing Web pages or at least the extraction of RDF data about its content. Therefore, it is necessary to automatically recognize when sequences of a text correspond to resources that are described within a Semantic Web database.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 420–429, 2011.

In this paper, we present a method for recognizing URI references in RDF graph that contain literal property values which occur as references in text. The overall problem confirms to be a named entity recognition. More precisely, we apply suffix arrays to represent text data and a relational database scheme to represent RDF data. Our approach performs a recognition of named entities in text referring to URI references in the RDF graph in linear time complexity and without the requirement to hold names of existing entities in memory. Both is needed to implement a named entity recognition on the scale of for instance the DBpedia database consisting of hundreds of millions of URI references.

The paper is structured as follows: At first we outline existing approaches and systems that relate to recognizing named entities in text. Then, we describe the task of recognizing named entities to finally present the implementation by using the suffix array and relational database schema.

2 Related Work

The approach and problem domain described in this paper is close to an ontology-based information extraction (OBIE) system presented by Wimalasuriya et al. [3]. The difference to OBIE is, that It we process pure RDF data as input without requiring the existence of ontological structures beyond the semantics of RDF.

The comparison of our approach with existing approaches of the ANNIE system [4] reveals that our system does not depend on any grammatical language models. It just needs a POS tagger and a noun-phrase chunker for a given language.

Zemanta [5] is a web service for building web mashups. Zemanta also spots for labels of DBpedia or Freebase[1] resources (namely instances) in web pages. The API returns results in RDF format and provides ratings about relevance and certainty. Unfortunately, Zemanta's service returns at most ten entities [2].

Open Calais[3] provides services for entity recognition. The focus of Open Calais is set on News content. Entities retrieved by Open Calais are defined in a proprietary ontology hosted by them. The coverage of entities that possess links to other linked databases such as DBpedia is very small.

Ontos' Semantic API[4] is similar to Open Calais' services. Ontos also hosts a proprietary ontology populated with entities that can be retrieved as mentions in text. But compared to Open Calais, the degree of linkage between DBpedia and Ontos' instance base is much higher.

All services provided by Zemanta, Open Calais, and Ontos are freely available but unfortunately closed source. To our knowledge no technical reports nor papers exist that explain details of used algorithms.

[1] http://www.freebase.com
[2] http://developer.zemanta.com/docs/suggest/
[3] http://www.opencalais.com
[4] http://www.ontos.com

3 Separating RDF Graphs

RDF is intended to represent and transfer information on the web. In general, existing RDF data on the web describes a single field of concern such as music artists, geo information, or even Wikipedia articles. We separate these data sets into information units that describe things such as Freddy Mercury, Berlin, or `http://wikipedia.org/wiki/Kaiserslautern` which are referred to as URI references. A URI reference possesses a collection of RDF statements which may be separated into three categories.

1. RDF statements that assign literal values to things:
   `dbp:Konrad_Zuse  foaf:name  ''Konrad Zuse''  .`
2. RDF statements that assign type information to things:
   `dbp:Konrad_Zuse  rdf:type  foaf:Person  .`
3. RDF statements that co-relate two things by using an RDF property:
   `dbp:Konrad_Zuse  dbp-owl:known_for  dbp:Z4_(computer)  .`

The RDF-based named entity recognition system, developed in this paper incorporates existing knowledge about URI references in RDF data in order to identify correlating references in plain text data.

The recognition of a text sequence as entity reference of a semantic entity is a means of semiotics. Here, the text passage "Konrad Zuse" was recognized as a DBpedia resource `dbp:Konrad_Zuse` that provides a data description about the real world person called Konrad Zuse who is the inventor of the Z4 computer. In general, entities can be recognized in a text passage if the semantics behind the text sequence can be resolved as a reference to a real world concept that is or can be described by RDF statements in the RDF graph.

4 Identify Parts of Text Sequences as Entity Candidates

The recognition of entity names (sometimes called proper nouns) is a traditional topic in the field of information extraction. Although properties of entities also comprise other attributes such as height, age, prize, the important step is to recognize the entity's name because in most languages names are used as mentions in text to refer to real world entities. In human languages, names of conceptual as well as physical object are built upon compounds of nouns. The semantic of languages is mostly built upon a syntactic representation comprising of the logical units subject, predicate, and object. A predicate defines a directed relationship between a subject and an object. Let us inspect the following sentence:

> *The Resource Description Framework is a graphical knowledge representation format.*

This sentence statement can be separated into the subject "The Resource Description Framework", the predicate "is a", and finally the object "graphical

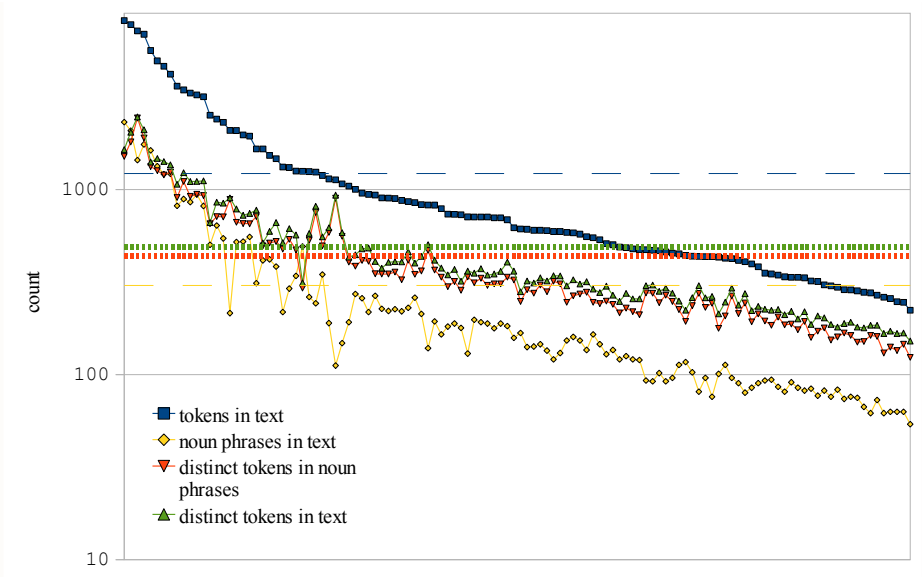


Fig. 1. Number of tokens and noun phrases (x-axis) from 120 randomly chosen Wikipedia articles

knowledge representation format". We base the identification of the proper nouns "Resource Description Framework" and "graphical knowledge representation format" as candidates for possible named entity references on standard NLP approaches, namely POS tagging and noun phrase chunking. The following listings show phrasal units and part of speech tags of this English sentence:

```
RDF is   a  graphical knowledge representation format .
NN  VBZ  DT JJ        NN        NN             NN     .
```

The second line below this sentence contains a sequence of POS tags that describe syntactic language units of this sentence. The tag set is defined by the English brown corpus.

```
RDF  is   a graphical knowledge representation format .
B-NP B-VP O B-NP      I-NP      I-NP           I-NP   O
```

This second example describes an I-O-B notation of phrasal units within the sentence. The tag NP represents a noun phrase, VP stands for noun phrase, and the prefix B marks the beginning of a phrase, I marks the the word to be part of a phrase, and O defines a word to stay outside any phrasal units. In general, we consider noun phrases to be candidates for the recognition of named entities.

POS taggers as well as noun phrase chunkers can be easily built by applying supervised machine learning methods. In this thesis we used POS tagging models

```
SUFFIX_LENGTH = 100      # length of suffixes
suffix_array = []        # the suffix array
for np in noun_phrases :
  for word in np :
     suffix = text[word.start :
(word.start+SUFFIX_LENGTH)]
     suffix_array.append(suffix)
suffix_array.sort()      # sort the suffixes in array
```

Fig. 2. Simple suffix array creation

for the German and English language that are maximum entropy classifiers [6] provided by the Open NLP library[5].

By using these POS taggers, the German Tiger corpus, the English ConLL2000 corpus, and conditional random fields (please refer details in [7]) implemented by the Mallet library [8] as machine learning method for tagging sequences, we trained a noun phrase chunker on the English and German language by using a window about the previous three and succeeding three words and POS tags. We reached on the English ConLL2000 corpus an F-measure of 0.92 (the best listed system from 2001 reached 94.13 [9]). On the German Tiger corpus our model reached an F-measure of 0.90.

In Figure 1 we show a general distribution of noun phrases in relation to single tokens (words and other language units) within 120 randomly chosen articles from Wikipedia. On this logarithmic scale, it can be seen that for reasons of performance it is valuable to reduce the amount of all tokens in text to just those tokens in text that a re part of noun phrases.

5 Resolve URI References

After having identified candidates for proper nouns in text, the next step of the named entity recognition process is to compare these candidates with name properties of URI references in the RDF graph. RDF graphs like DBpedia[6] contain several million literal values. This requires the application of scalable string comparison mechanisms in terms of memory and runtime efficiency. Straight forward approaches like querying a database consisting of these literal values for each entity candidate results in too long response times. The second possibility is hashing all literal values in memory, which is too memory consuming. For example, the DBpedia spotlight service maintains a nine gigabyte hash table of selected literals form DBpedia in memory. Besides the scalability issues, simple

[5] http://incubator.apache.org/opennlp/

[6] The sparql query `select count(?label) where {?s ?p ?label. FILTER (isLiteral(?label))}` counted 105,019,913 literal values in the sparql endpoint of the DBpedia 3.6. The sparql endpoint of linked geo data counted 111,083,681 literal values.

database queries or hash table lookups just support the simplest comparison method that is the exact match. Here, we want to remark that in times of Unicode even this exact match defined by the Unicode collation algorithm is much more complicate than expected. In many cases, a noun phrase does not match exactly with known literal values. If we consider the noun phrase "Peter, Paul and Mary", the DBpedia contains a matching literal that occurs as object in the following RDF triple:

```
dbpedia:Peter,_Paul_and_Mary  rdfs:label  ''Peter, Paul and
Mary''                                                    .
```

But what about matching this phrase with the first names of Peter Yarrow, Noel "Paul" Stookey, und Mary Travers? Substring matches are required to solve this issue. A text index structure that provides such matching operations that scale on large data is the suffix array [10]. A suffix array is an efficient data structure for solving the longest common substring problem. The Python code sample below shows a very simple creation of a suffix array in $O(n * \log(n))$ complexity. [10] propose methods for constructing suffix arrays in even linear time.

The setting of `SUFFIX_LENGTH` depends on the typical length of literal values in the instance base. In the case of DBpedia we made good experiences by setting it to 100 characters.

```
[
'and Mary';
'Mary';
'Paul and Mary';
'Peter, Paul and Mary'
]
```

Fig. 3. The suffix array build upon word suffixes of "Peter Paul and Mary"

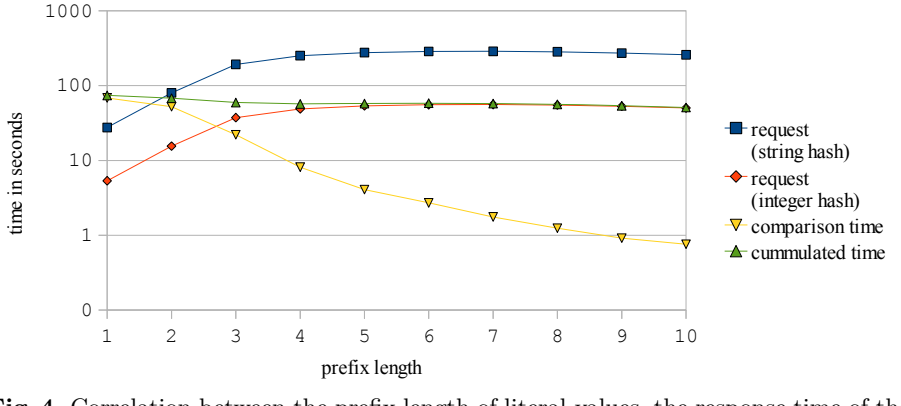


Fig. 4. Correlation between the prefix length of literal values, the response time of the database query and response time of the string comparison

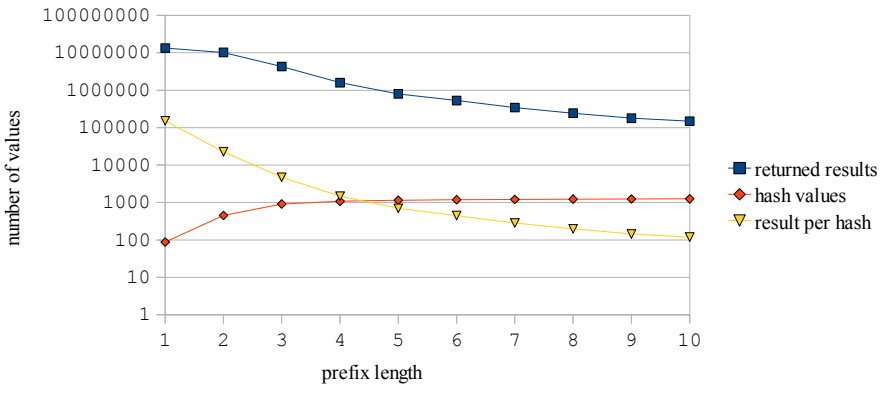


Fig. 5. Correlation between the prefix length of literal values, the number of plain results of the SQL query, and the number of hashed values in the result of the SQL query

```
-- an index table about literal values
CREATE TABLE index_literals (
   index SERIAL PRIMARY KEY,   -- internal key for literals
   literal varchar(256),       -- plain literal value
   prefix int );               -- hash value for prefixes

-- an index table about URI values
CREATE TABLE index_resources (
   index SERIAL PRIMARY KEY,    -- internal key for URIs
   uri varchar(256) UNIQUE);    -- plain URI representation

-- RDF triples between URI and literal
CREATE TABLE symbols (
   subject int,    -- REFERENCES index_resources(index)
   predicate int,  -- REFERENCES index_resources(index)
   object int);    -- REFERENCES index_literals(index)
```

Fig. 6. This relational database schema represents RDF triples that possess literals as object values in tables `index_literals` and `symbols`. Please consider that `index_literals` assigns integer indexes to literal values. Table `index_resource` assigns integer indexes to URI references.

String comparisons can now be performed by searching the lookup string as a prefix value in the sorted list of suffixes in an $O(\log(n))$ complexity. Our implementation uses pointers to text passages instead of physical substring in order to ensure that the text only remains once in memory.

After having the suffix array on the one hand, it is necessary to on the other hand provide an ordered list of literal values of the RDF data that can be compared to the suffix array via prefix matches. This requires the literals of the RDF graph to be indexed with prefix values of a defined length in order to request a list of literal values that possibly match in a prefix comparison with

```
SELECT DISTINCT L.literal, L.index, S.predicate,
FROM index_literals L, symbols S
WHERE (
    symbols.object = L.index AND L.prefix IN (
        -- list of hashed prefix values of suffix array entries
  )) ORDER BY L.literal
```

Fig. 7. This SQL query returns an ordered list of literals that match with the suffix array entries and the RDF properties that possess these values

```
SELECT DISTINCT S.subject, S.predicate, S.object, R.uri
FROM symbols S, index_resources R
WHERE (
  S.subject = R.index AND
    (S.predicate, S.object) IN (
                -- list of property value pairs
    ))
```

Fig. 8. This SQL query returns a list of URI references of recognized subjects, and URIs of the properties that possess the matching literal values with the recognized names entities

entries of the suffix arrays. Within our implementation, we used the open source relational database PostgresQL for storing, indexing, and sorting literal values in UTF-8 encoding.

Figure 4 shows that requesting literal values from the database by using plain literal prefixes as kind of hash description turns out to perform worse for each length of a chosen prefix. The query processing is based on a 32 bit character wise string comparisons. A more efficient approach is to use the following scalar function to hash string prefixes with a single 32 bit integer value.

$$\text{hash}(s) = s[0] * 31^{n-1} + s[1] * 31^{n-2} + ... + s[n-1] \quad (1)$$

This hash function is used by the Java programming language.

Figure 5 shows the correlation between prefix length and number of result or hash values. By taking into account the results from Figure 4 and Figure 5 as well as experiences made while experimenting with the large DBpedia dataset, we currently use a prefix length of 3 in our implementation.

The prefix match between database results and the suffix array produces a list of known literal values form the RDF graph that occur in text. In order to attach information about RDF properties and later on references to RDF subjects to these matches, we modeled a relational schema (see Figure 6) and populated this with data from the RDF graph. The SQL select query in Figure 7 returns for a given list of hashed prefix values of an suffix array an ordered list of pairs consisting of literal values that may are likely to match with any entries in the

suffix array as well as the RDF properties that possess these literals as values. In order to get concrete subjects for recognized property value pairs the following SQL query is used. This query returns complete RDF triples.

6 Conclusion and Outlook

In this paper we presented an efficient implementation of a named entity recognition on the Semantic Web. Here, named entities are text references that link to specific URI references of an RDF graph. We showed how to represent text as suffix arrays, and presented a relational database scheme to represent those parts of RDF data that are necessary for implementing an efficient recognizing of named entities in linear time complexity. This approach does not require to hold all existing literal values of URI references in main memory.

While building suffix arrays, our evaluation proved an increase of efficiency when using only the suffixes of noun phrases as candidates for possible named entity references.

Finally, it can be concluded that the combination of both representation formats provides a scalable software architecture for recognizing named entities.

In future work, we plan to elaborate more on the disambiguation of named entities that were resolved to refer to more than a single URI reference.

This work was financed by the BMBF project Perspecting (Grant 01IW08002) and the research project Semopad funded by the Stiftung Rheinland-Pfalz für Innovation under contract no. 961-386261/1001.

References

1. Manola, F., Miller, E., McBride, B.: RDF Primer. Technical report, World Wide Web Consortium (February 2004)
2. Adida, B., Herman, I., Sporny, M., Birbeck, M.: RDFa 1.1 Primer, rich structured data markup for web documents. Technical report, World Wide Web Consortium (March 2011)
3. Wimalasuriya, D.C., Dou, D.: Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science 36(3), 306–323 (2010)
4. Bontcheva, K., Tablan, V., Maynard, D., Cunningham, H.: Evolving gate to meet new challenges in language engineering. Nat. Lang. Eng. 10(3-4), 349–373 (2004)
5. Tori, A.: Zemanta service. Zemanta (2008)
6. Nigam, K., Lafferty, J., Mccallum, A.: Using Maximum Entropy for Text Classification. In: IJCAI 1999 Workshop on Machine Learning for Information Filtering, pp. 61–67 (1999)
7. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, ICML 2001, pp. 282–289. Morgan Kaufmann Publishers Inc, San Francisco (2001)
8. McCallum, A.K.: Mallet: A machine learning for language toolkit (2002)

9. Zhang, T., Damerau, F., Johnson, D.: Text chunking using regularized winnow. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, ACL 2001, pp. 539–546. Association for Computational Linguistics, Stroudsburg (2001)
10. Kärkkäinen, J., Sanders, P., Burkhardt, S.: Linear work suffix array construction. J. ACM 53, 918–936 (2006)

Extracting Personal Concepts from Users' Emails to Initialize Their Personal Information Models

Sven Schwarz[1], Frank Marmann[2], and Heiko Maus[1]

[1] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Kaiserslautern
[2] Capgemini, Düsseldorf

Abstract. Although the Semantic Desktop paradigm has great potential, new users have to face the cold-start problem. Having to start with empty models is a barrier to any semantic technology and filling them with world-known concepts does not work for *personal* models. We propose to analyze the email database of a user and extract concepts of multiple types to fill the empty PIMO. The paper presents results of the research project Semopad funded by the Stiftung Rheinland-Pfalz für Innovation under contract no. 961-386261/1001.

1 Motivation

Dealing with the immense, constantly increasing information load of contemporary knowledge work is a challenging task. State of the art approaches mostly address this using intelligent storage, indexing, and retrieval of information elements. The Semantic Desktop Paradigm envisions the maintenance of a *Personal Information Model (PIMO)* for each individual user, that is, a semantical model and conceptualization of the most important concepts of an individual worker [11]. These concepts are technical representatives for digital resources (like files or emails), physical entities (like locations or persons), or abstract concepts (like topics or projects). As they are connected with semantic relations, they span a semantic graph connecting particularly information elements like documents with semantic concepts. For the information management this enables a semantic-oriented indexing and retrieval.

Although tools on a semantic model have great potential, they come with a serious tradeoff: As the model is typically an independent data structure with a schema of its own, it has to be filled. In the past, the build-up was typically done by forcing the user himself to enter the data more or less manually. In the days of Web 2.0 researchers have a tendency to shift away from individual PIMOs towards using world knowledge like Wikipedia, Linked Data[1] and web services like DBpedia[2] or OpenCalais[3]. The first approach is a barrier to any kind of semantic technology while the second lacks essential information that only the *individual* worker can know. Phone numbers of team members and internal projects are

[1] http://www.w3.org/DesignIssues/LinkedData.html
[2] http://dbpedia.org/
[3] http://www.opencalais.com/

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 430–439, 2011.

not published online and can not be recognized by services like OpenCalais. Besides, one might not want to pass information containing intellectual property to external web services. Instead, we address the cold start problem directly by analyzing *personal* structures and extracting *personally relevant* concepts for filling an initially empty PIMO.

2 Approach

Our approach towards filling an empty PIMO is to analyze the emails of a user to calculate the most "important" concepts found therein. The analysis runs in the background, and the user may peek at any time at the found concepts so far. If the user is happy with the concepts found after a certain time, they can stop the analysis early and start working right away.

Although we will be including other resources in the future, we are currently focussing on emails as they represent a rich source of personal information. Particularly, the "importance" of concepts can be derived quite well by looking at the amount and balance of the email communication with other persons (i. e., *contacts*). Therefore, our approach analyzes and evaluates the contacts the user is communicating with. For our scenario, filling an empty PIMO, the term "importance" comes down to calculating the *PIMO relevance* for concepts.

Goals

A PIMO contains different dimensions of concepts for relating and annotating resources and concepts of the user's daily business. For initializing an empty PIMO we focus on generating concepts of a realistic subset of these dimensions:

- persons, locations, organizations, topics, projects

The approach focusses on the following conceptional and technical design goals for realizing a tool that is actually felt to be supportive and useful to the user when initializing an empty PIMO.

- require minimal user interaction
- crawl and process emails *iteratively* and allow peeking and working with early, preliminary results
- prefer *quality* not quantity of proposed concepts
- create concepts of *personal* importance
- prioritize up-to-date concepts

Related Work

Elsayed et al [4,3] present an approach that uses email metadata to model identities occurring in an email archive. For modeling and unifying identities, evidence is obtained for names, email addresses, or signatures. The authors are investing in methods to detect identities by analyzing patterns in the communication or

characteristic misspellings in the text to associate, for instance, *Bill* being a common nickname for *William*. However, our focus is on detecting most relevant concepts rather than perfect unification.

The approach presented in [9] analyzes a user's native structures, e. g., his email or file folder hierarchies, and generates concepts for each folder. It attempts to detect relationships (e. g., same-as, subclass-of) between the generated concepts by analyzing their labeling as well as their sub-folder relationships. The approach worked well and required only minimal user interaction. However, it did not detect a concept's *type*, nor its (recent) *relevance*.

The *IRIS Semantic Desktop* [2] is another implementation of the Semantic Desktop idea. IRIS exploits the user's email archive for populating contacts automatically from the sender and recipients metadata. As not every contact which is *visible* in the user's emails is actually of *importance*, this strategy leads to many unnecessary concepts, filling the PIMO with garbage right from the beginning. IRIS also generates concepts for "projects" by merely applying textual clustering to all email contents. While some of these concepts may be relevant, many will not be and the user can not clearly recognize them as singular and known entities. They are type-less concepts without semantics, mere buckets containing/relating textually similar stuff.

General Procedure

As explained, the main goal of our approach is to analyze a user's emails to calculate the *most relevant* concepts to fill an empty PIMO. To do so, we take an IMAP account of the user as the main input and, according to the goals, we crawl the emails *iteratively*. Technically, the implementation retrieves a list of all emails, puts them into a bag and then draws emails *randomly*. Each email picked that way is input into a processing pipeline described below. The processing of a single email will lead to proposals of new concepts as well as to updates of some concepts' relevance values. This iterative processing allows users with large email databases to stop the analysis at any time (meaning very early) and work with the concept proposals calculated so far. The random picking is essential here to guarantee that after a short processing time, the set of emails picked so far will already be similarly distributed as all the emails.

Processing Pipeline

After randomly picking an email from the IMAP store, the email is passed into a processing pipeline applying the following steps:

1. Identify contacts in the email (sender, recipients)
2. Calculate/Update PIMO relevance of these contacts
3. Identify concepts found in the email
4. Classify type of these concepts (e. g., organization versus location)
5. Calculate/Update PIMO relevance of these concepts

2.1 Identification of Contacts in Emails

While one could also argue that the address book already contains contacts, it does not contain all of them, nor is it guaranteed that the information is up to date. Furthermore, the relevance of a contact can not be estimated. On the other hand, the emails themselves contain objective, up-to-date, and complete data enabling a quantitative computation of a relevance value for each specific contact. So, whether or not a contact exists is not the question, rather how "relevant" that contact is for the user.

Many contacts will use multiple email addresses when communicating with the user. As we do not want to represent different contacts of the same person, we need to merge the "identical" ones. Most contact information will be accompanied by the person's name in the sender or recipient field of an email, e. g., `Sven Schwarz <sven.schwarz@dfki.de>`. Besides email addresses as a primary key, the name information (first name, last name) is used to unify contacts, with the assumption that the user does not know two different contacts with the same name. The current implementation also takes into account multiple domains and different ordering of first name and last name as well as optional middle names, titles etc. however a detailed description would exceed the focus of this paper. The main idea is to find a normalized labeling for each contact in the email and then merge contacts with identical normalized labels.

This, of course, may lead to accidental unification of two different contacts, but for our purpose this does no harm. The goal is to fill an initially empty PIMO with *relevant* concepts, and the incorrectly merged contact is still highly relevant if the two individual contacts are highly relevant. The various merged email addresses are stored for each contact and the GUI proposing the concepts to the user also allows to split the contact again later.

2.2 PIMO Relevance of Contacts

We claim that the relevance of a concept found in emails is proportional to the relevance of sender and recipients of these emails. The relevance values for contacts lay a computational basis to estimate the relevance of identified concepts.

To calculate a contact's PIMO relevance, we analyze its *communication characteristics* by evaluating the features for frequency, balancing, and recency. In the following, we explain these features and the corresponding mathematical formulae as well as their computational influence to the contact's relevance value.

Communication Amount Factor ($\mathbf{CAF}_c$): This factor reflects the *amount* or *richness* of communication exchanged with the contact c. It considers not only *how many* emails have been exchanged but also how often the contact was the sender or recipient and how many other recipients were addressed in the same emails. Computationally, it comes down to the following formula:

$$CAF_c = \alpha \cdot \left(\beta \cdot \frac{sent(u,c)}{\max_{o \in C} sent(u,o)} + (1-\beta) \cdot \frac{sent(c,u)}{\max_{o \in C} sent(o,u)} \right) + (1-\alpha) \cdot \left(\frac{occ(c)}{\max_{o \in C} occ(o)} \right) \quad (1)$$

where $sent(c_1, c_2)$, with c_1, c_2 contacts, is the number of emails that c_1 sent to c_2 and $occ(c)$ the number of occurrences (sender or recipient) of contact c in all of the user's emails. The parameters α and β are tuning parameters. Empirical tests showed that $\alpha = 0.85$ and $\beta = 0.9$ delivers good results. As we are striving for conceptional innovation, fine tuning these parameters is not within the scope of our work.

Communication Balance Factor ($\mathbf{CBF_c}$): Highly relevant contacts can not be identified by the *amount* of emails alone. With important contacts the user typically shares a *balanced* communication in which each participant is both active (sender) and passive (recipient). Compare this to a (possibly annoying) contact that only *sends* emails to the user. Also, contacts that only *receive* emails from the user might be "important" officially, but if that contact never/rarely responds, he or she is probably not relevant as a concept in the user's PIMO.

The *communication balance factor* represents this issue in form of a value between 0.0 (totally unbalanced) and 1.0 (well balanced). Technically, we first calculate a value

$$bal_c = \left| \frac{sent(u,c)}{sent(u,c) + sent(c,u)} - 0.5 \right| \quad (2)$$

delivering a value between $[0.0; 0.5]$ — 0.0 in cases where either person has sent equally many emails and 0.5 in cases where emails were only sent from one side. Aiming at treating values near zero as well balanced, we defined a threshold parameter ρ (we set $\rho = 0.1$), and computed the CBF_c as follows:

$$CBF_c = \begin{cases} 1, & \text{if } bal_c < \rho \\ 1 - \frac{bal_c - \rho}{0.5 - \rho}, & \text{otherwise} \end{cases} \quad (3)$$

Recency Factor ($\mathbf{RF_c}$): As stated in the goals in section 2, proposals for filling an empty PIMO should prioritize *recent* concepts. The factor uses a linearly decreasing weight for the occurrences[4] of contact c. The weight decreases linearly from 1.0 to 0.0 antiproportional to the number of months past since today. To speed things up, we ignore old occurrences, defining "old" to be more than four years old. The overall value for the *Recency Factor* is obtained by multiplying the occurrence values with the appropriate linear decreasing weight and dividing the result by the sum of all "not-old" occurrences:

$$RF_c = \frac{\sum_{t=0}^{48} (1 - \frac{t}{48}) occ(c,t)}{\sum_{t=0}^{48} occ(c,t)} \quad (4)$$

[4] A contact *occurs* if he or she is either sender or recipient of an email.

where $occ(c,t)$ the number of occurrences of contact c and t the number of months before today.

PIMO Relevance ($\mathbf{PR_c}$): The PR_c value lies between 0.0 and 1.0, whereas 0.0 means c is not relevant for the PIMO and 1.0 meaning surely relevant. The measures described above will be combined to compute one single relevance value taking into account how often, how "useful" or "interesting" and how recent the contact c is supposedly felt to be by the user:

$$PR_c = \begin{cases} 1, & \textit{if } c \text{ is the user} \\ CAF_c \cdot (0.1 + (0.9 \cdot CBF_c)) \cdot (0.1 + (0.9 \cdot RF_c)), & \textit{otherwise} \end{cases} \tag{5}$$

Note: The *PIMO Relevance* does not have enough expressiveness to be used as an absolute ranking. To imply that contact A with relevance 0.45 is felt more important than contact B with relevance 0.44 is not a valid induction. However, A with relevance 0.45 is probably more relevant than contact C with a relevance 0.2. So, it can be used to rank the concepts and to allow the user skim the top concepts for filling his/her PIMO. Furthermore, the contact's relevance values are used to compute the relevance values for the concepts identified in the emails.

2.3 Identification of Concepts Found in an Email

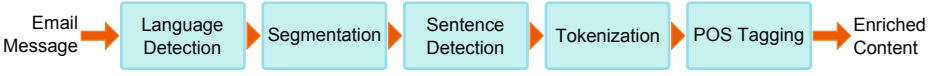
The *persons* dimension is filled with the contacts previously identified. Sorted with decreasing rank they are proposed as concepts to fill the empty PIMO.

Concepts for the other dimensions (e. g., *locations* or *organizations*) are much harder to excavate from the emails because they are encoded textually. An identification requires linguistic approaches. However, natural language processing (NLP) is still a basic research issue and not yet ready for use. Therefore, ongoing research applies "shallow linguistic" approaches like *part-of-speech (POS) tagging* and *information extraction (IE)* without requiring a full understanding of the language [8]. The community works on *named entities recognition* for identifying single entities (e. g., persons, locations) and *relation extraction* for identifying relations (e. g., is-located-in, works-for) between such entities. IE systems mostly rely on local textual context for such identification. This is based on Harris' distributional hypothesis, which states that terms have similar meanings if they occur in the same context [5]. Hearst used text patterns to identify synonyms and hyponyms automatically [6]. Building on that, similar patterns can be specified to spot entities of PIMO-specific concept types.

Technically, the concept identification is realized in two steps: First, the textual email content is *enriched* linguistically. Second, the enriched content is searched for commonly used entities or patterns to detect the concepts relevant for the PIMO.

Enriching Email Content: The patterns for detecting concepts do not work on the pure ASCII text directly. Instead of working on character or word level,

they work on tokens with additional linguistic information. Therefore, an NLP chain segments and tokenizes the text, as well as, classifies *parts of speech* (POS):



Language detection is a minor technical issue that we addressed using the *Java Text Categorizing Library*[5]. ***Segmentation*** is currently done on coarse level, separating normal text, quotations, and email signature(s). The segmentation of the email into the different zones is a non-trivial task discussed in detail in [7]. The authors define various features to train a SVM classifier which segments an email message into zones, however, a tool which includes their trained model is not yet available. We implemented a simple, heuristic classifier using on the first character(s) of the lines of text (`'>'` or `'|'` are treated as quotation, etc.) and achieved a sufficient precision to distinguish text, quotations, and signatures. ***Sentence Detection, Tokenization and POS tagging*** are realized using the *OpenNLP toolkit*[6]. Produced information about sentence boundaries and token positions are retained for further processing.

Detecting Concepts in the Text: The goal is to detect instances of all concept types that are supposed to span a PIMO. Every concept type has its own textual features used by humans to reference instances of that type. We utilize this and apply corresponding pattern matching for detecting the concepts in a text. For some types, a gazetteer supports the detection of known entities, but, generally, we must search for commonly used textual patterns to spot them.

In addition to the concept itself, the pattern matcher also delivers a corresponding evidence value for each found match. As some patterns are more discriminative than others, the evidence value measures not only the quality of the match but also the quality of the pattern itself. As multiple patterns are required to achieve a good recall of concepts, multiple evidences are gathered and stored in an evidence database to be aggregated later for resolving conflicts between multiple detected types for one concept. For computing the PIMO relevance later, this database also holds references to the emails where the concepts have been found.

In the following, we skim over patterns that are actually used for detecting concepts of specific types. Note that the presented patterns are neither complete nor described in detail, as the focus of this paper is on the *conceptional* workout.

Locations are often given in form of a complete address, including a postal code. We use a postal codes gazetteer provided by Geonames[7] containing 16,380 relations (postal code ↔ city). The corresponding pattern matches a *number* followed by a *noun phrase* (POS-tag `NP`). Slightly modified city names,

[5] http://textcat.sourceforge.net/
[6] http://opennlp.sourceforge.net/
[7] http://www.geonames.org/

as in `Frankfurt a.M.` instead of `Frankfurt am Main`, are still detected, however, with a lower evidence value. To increase recall of locations outside address boxes, we use patterns like `‘‘from {NP} to {NP}’’` or `‘‘take[s] place in {NP}’’`. These patterns are highly ambiguous and, hence, deliver much lower evidence values for matches. Nonetheless, they are an important contribution.

Organizations, particularly companies, can be detected following an idea of [10] to "extract company names by looking backward from a company name indicator". We defined a set of company indicators for German and English companies, like `AG`, `GmbH`, `Inc` and defined corresponding patterns like `‘‘{NP} Inc’’`. We also added simple patterns like `‘‘University of {NP}’’`. Actually, available POS-taggers often do a bad job classifying noun phrases. So, we had to do some conceptually irrelevant tweaks to our system.

Topics and Projects are *abstract* concepts, whereas *topic* means the *subject of a discussion or document* and *project encapsulating a series of actions or steps towards some goal*. We tested a gazetteer of Wikipedia categories made available by DBpedia. Experiments revealed that *common* "topics" like *time*, *title*, *milk* or *juice* have been detected more prominently and, hence, with higher PIMO relevance than topics like *Artificial Intelligence* or *LISP*. However, this is against the PIMO's aim towards *personal* and *discriminative* concepts.

As it is impossible to generate a gazetteer containing *personal* (i. e., generally a-priori unknown) topics and projects of *different* users, an alternative approach is required. Current linguistic analysis libraries are not yet capable of providing a technical basis for a reliable detection of abstract concepts. Furthermore, experiments could not detect any shared patterns around textual occurrences of topics and projects in the email contents. However, the labels of a user's email *folders* often contain references to topics and projects prominently. We developed a two-phase approach for their detection: In the first phase, we generate candidates for concepts from the folder names, being either topics or projects. In the second phase, we search for occurrences of this concept in the email contents to classify its type (topic versus project) and deliver a corresponding evidence value. While topics and projects can not be *detected* via text patterns, their type can be *disambiguated* using their textual context. We apply an a-priori trained n-gram classifier [1] working at word-level (3 words left plus 3 words right of the occurrence) for this disambiguation task.

2.4 Classify Type of Found Concepts

It often happens that one concept was found more than once and by using patterns for different concept types. To resolve that conflict, an *evidence database* contains statements for every found concept, including evidence values for (generally) *multiple* matching concept types. The concept type with the highest evidence is chosen to type the found concept, whereas Dempster's rule of combination [12] is applied to combine multiple evidences e_i for the same concept

type. However, in our case, there are only evidence values *for* and no evidence *against* some concept hypothesis, leading to a simple, iterative calculation for combining these evidence values to one e:

$$\begin{aligned} e = e_1 \oplus \ldots \oplus e_n = (\ldots(e_1 \oplus e_2) \oplus \ldots) \oplus e_n \\ with \quad e_i \oplus e_j = e_i + (1 - e_i) \cdot e_j \end{aligned} \tag{6}$$

2.5 PIMO Relevance of Concepts

As described in section 2.2, the PR_c of contacts are the basis for computing the PR of concepts of other dimensions. Therefore, the source email is always stored with each evidence. Given a given concept x, let E_x be the set of all emails where x was found. As the *sender* of an email is giving the main influence in terms of PR, we currently (as a simplification) ignore other aspects to calculate the PR of an email:

$$PR_e = PR_c \quad where\ c = sender\ of\ email\ e \in E_x \tag{7}$$

Having only positive belief values for PR_e and none against, we have the same special case of Dempster-Shafer gain and can compute the PR_x for concept x:

$$PR_x = p_1 \oplus \ldots \oplus p_n \quad and \quad p_k \in \{PR_e | e \in E_x\} \tag{8}$$

3 User Interface and Evaluation

The described approach generates concepts with estimated relevance values (PR). A user interface presents these concepts in columns of different dimensions (persons, locations, organizations, topics, and projects), in each column ordering the concepts by decreasing PR. The user can select arbitrary concepts (particularly top concepts) for initializing his or her PIMO. While the approach can not be tested against some ground truth, we checked the *utility* of our tool with the following measures:

- How many user-expected concepts have been actually found by the tool?
- How many concepts were imported into the PIMO in what amount of time?

The participants were asked to write down concepts they would expect before starting the tool. On average around 66% of the expected concepts were eventually imported. While this looks promising, keeping in mind the goal is laying a suitable foundation for incremental extension, this recall should be increased in the future. On average, the participants imported 183 concepts in 17,5 minutes of their human time, implicating a speed of 11 concepts per minute, or 6 seconds per concept. Being definitely faster than manual creation the participants were able to initialize their PIMO quickly and with minimal user interaction.

4 Conclusion

We present an approach which analyzes the email database of a user and extracts concepts for initializing an empty PIMO. Aiming at prioritizing the most "relevant" concepts when filling the PIMO, our approach estimates a *PIMO relevance (PR)* value for each identified concept. This PR is based on the communication statistics of the contacts in the emails. It is used to rank the proposed concepts in the GUI presented to the user. The evaluation has shown that concepts with high PR are found to be among the most relevant concepts. Participants found the tool required minimal user interaction and time to initialise an empty PIMO.

References

1. Cavnar, W.B., Trenkle, J.M.: N-gram-based text categorization. In: Proceedings of SDAIR 1994, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US, pp. 161–175 (1994)
2. Cheyer, A., Park, J., Giuli, R.: Iris: Integrate. relate. infer. share. In: Decker, S., Park, J., Quan, D., Sauermann, L. (eds.) Proc. of Semantic Desktop Workshop at the ISWC, Galway, Ireland, vol. 175 (November 2005)
3. Elsayed, T., Namata, G., Getoor, L., Oard, D.W.: Personal name resolution in email: A heuristic approach (2008)
4. Elsayed, T., Oard, D.W.: Modeling identity in archival collections of email: A preliminary study. In: CEAS (2006)
5. Harris, Z.: Distributional structure. In: Fodor, J.A., Katz, J.J. (eds.) The Structure of Language, pp. 33–49. Prentice-Hall, Englewood Cliffs (1954)
6. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational linguistics, COLING 1992, vol. 2, pp. 539–545. Association for Computational Linguistics, Stroudsburg (1992)
7. Lampert, A., Dale, R., Paris, C.: Segmenting email message text into zones. In: EMNLP 2009: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 919–928. Association for Computational Linguistics, Morristown (2009)
8. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
9. Osterfeld, F.: Ein lernfähiges System zur Akquisition und Wartung von persönlichen Informationsmodellen. Master's thesis, Fachbereich Informatik, University of Kaiserslautern (2006)
10. Rau, L.F.: Extracting company names from text. In: Proc. of the Seventh Conference on Artificial Intelligence Applications CAIA 1991, Miami Beach, FL. vol. II: Visuals, pp. 189–194 (1991)
11. Sauermann, L., van Elst, L., Dengel, A.: PIMO - A Framework for Representing Personal Information Models. In: Pellegrini, T., Schaffert, S. (eds.) Proceedings of I-MEDIA 2007 and I-SEMANTICS 2007 International Conferenceson New Media Technology and Semantic Systems as part of TRIPLE-I 2007, J.UCS, pp. 270–277. Know-Center, Austria (2007)
12. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)

A Diagrammatic Approach to Discovering Chances in Team Relationships

Ruediger Oehlmann and Balpreet Gill

Kingston University London,
Faculty of Computing, Information Systems and Mathematics
Penrhyn Road, Kingston upon Thames, KT1 2EE, UK
R.Oehlmann@Kingston.ac.uk

Abstract. Typically interpersonal relationships in teams are investigated with questionnaires, interviews or observational studies, methods which lack either depth or are very costly. This affects the chances for improving such relationships that can be discovered. The first part of the paper describes a system that takes a diagrammatic approach to acquiring data on interpersonal relationships in teams. It will be argued that this approach is suitable to acquire data with less cost than conventional methods. The second part describes the analysis of relationships in a security team that utilizes the approach. The results of the analysis clearly identify strength and weaknesses of the intra-team relationships, and indicate that the approach is useful in organizational settings but also as research tool for investigating intra-group relationships in social psychology.

Keywords: Chance Discovery in Teamwork, Diagrammatic Reasoning, Interpersonal Relationships, Relationship Descriptors.

1 Introduction

The use of diagrams has a considerable tradition in computer science and cognitive science. For instance, Larkin and Simon (1987) have argued that a diagram can communicate a large amount of information. Information can be localized by linking it with a particular component of the diagram. Various research efforts have focused on the question how useful diagrams are for communication and reasoning (Cheng, et al., 2001). It should however be noted that most of the previous work has considered diagrams in the domains of engineering and science rather than in social science.

Investigations beyond these domains have considered diagrams in decision making, For instance, Howard and Matheson (2005) have proposed influence diagrams, which focus on the sequential and the informational aspect of decision processes. Chance discovery is concerned with the identification and the management of rare, but significant events, such as potential risks or opportunities, in some domain or application (Ohsawa, 2003). It is noteworthy that such chances are contextualised and presented in diagrams, which are necessary for a team of experts to identify the most promising chances.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 440–446, 2011.

Equally close to the objectives of the work described below is the concept of the family life space diagram (FLSD). This was originally developed by Mostwin (1980, cited in Barker et al. 1997). An FLSD consists of a large circle that contains several smaller symbols, often labeled circles or rectangles. The large circle represents the lifespace and the symbols represent living entities within that life space. The concept of a life space was introduced by Lewin (1951, cited in Barker et al., 1997). It conceptualized the relationships between living entities.

Whereas the FLSD characterizes relationships just in terms of the distance between icons in a semi-quantitative way without qualifying the type of relationship, Oehlmann (2006) proposed a Social Diagrammatic Language (SDL) that used typed relationship characteristics, such as *trusting* relationship, or *supporting* relationship.

Diagrams have also been used to describe characteristics of an individual rather than a relationship. For instance, Ortony et al. (1998) have proposed an ontology of emotions. The representation involved opposites, such as love and hate as well as a strength value that was taken from a Likert scale from 1 to 5, where 1 indicated a strong negative emotion, such as hate, and 5 indicated a strong positive emotion such as love. This approach has been incorporated into SDL (Oehlmann, 2006).

The references mentioned above suggest that an agreed set of descriptors is essential for any diagrammatic approach to intra-team relationships. Hinde (1997) has equally emphasized the need for an agreed set of descriptors and the difficulty to arrive at such a set. Kiesler (1983, cited in Soguet et al., 2001) has developed an interpersonal circle that contains opposite relationship characteristics, such as mistrust – trust. The characteristics are positioned at geometrically opposite parts of the circle. Hinde (1997) has argued that such descriptions are necessarily incomplete, because the nature of interpersonal relationships is influenced by too many factors. Moreover relationships are changing. So any description of relationships can only be a snapshot.

Given the large body of previous work, the brief description given above has necessarily to be incomplete. However even this brief description reveals a number of limitations. Firstly, there are hardly any examples of diagrams that represent social relationships. Moreover, there are no computerized tools to support the generation of such diagrams in a methodical way. Secondly the descriptors for characterizing social relationships that are described in the literature are often too general and make to a large extent use of psychological terminology, which is not always understandable by laypeople. In addition, there is often confusion between descriptors characterizing behavioral features attributed to an individual and descriptors characterizing relationships between individuals.

Hinde's (1997) concerns about the limitations of relationship descriptions can be addressed in two ways. Firstly, it is agreed that such descriptions are incomplete by nature. However, this incompleteness is in part set off by the possibility of team descriptions, where different team members describe the same relationship. Secondly, as all relationship diagrams are stored in a database, it will be possible to analyze the development of team relations over time. The next section will explain how the design and implementation of the TaROT[1] system has addressed these issues.

[1] TaROT is the acronym for **T**hinking **a**bout **R**elationships in **O**rganisational **T**eamwork.

2 TaROT – A System for the Diagrammatic Analysis of Team Relationships

The TaROT system implements a methodology of diagram-based reasoning about intra-team relationships. The overall procedure is shown in Figure 1. The system is fully implemented in Java as a stand-alone system using the MySQL database system for data storage. Important design principles include the reduction of reasoning about team relationships to binary relationships, the integration of qualitative and quantitative data, and data visualization.

The data stored in the system are sensitive by nature. Therefore, unlike most social networking systems, TaROT puts considerable emphasis on privacy. For instance, user name and contact data are hold in a database that is separate from the database that stores the relationship data. The user is only identified by a set of aliases. Users interact with the system through a sequence of different screens.

The first screen allows storing images of up to six team members together with their aliases. The next screen shows the images at the upper border of the screen and the user's image in the middle of the screen. All images of team members can be selected and moved around (Figure 2). In this way the user is encouraged to think about his relationships in general terms. A short distance between the image of a team member and the image of the user indicates a close relationship, whereas a long distance represents a distant relationship.

The image of the team member that is selected when the next screen is activated appears in the middle of that screen. Now the user can select nine buttons to characterize that team member. The number of buttons is consistent with numerous result from studies on memory span, which that people find it difficult to deal with higher number so items at the same time. For instance, pressing the button labeled "lazy-ambitious" results in positioning that text near the image. A slider can be used to assign a number between -3 and +3 to the text. Depending on the number the text will change the color and the font size. In this way the numerical value is presented optically. There are nine pairs of opposite characteristics the user can choose from. These are commonly used in social psychology, for instance in several of the diagrammatic structures mentioned in Section 1. In addition the user has the option to create a new tenth pair. It is planned to use this feature to update the list in a systematical fashion. This may be the appropriate course of action, if, for instance, one of the prepared features is rarely used, but another feature is often proposed by users. Besides the generation of quantitative date, the user can also generate qualitative data in the form of brief event descriptions or stories that explain the quantitative decision made in relation to the arrow.

Having characterized the other team member, the user can now focus on the detailed characteristic of the relationship with that member. The design of the next screen is similar to the previous screen. However it displays images of the user and of that team member the user used on the previous screen. The user then selects a button

on the left side. If a button at the left-hand side is pressed, a relationship descriptor appears in the middle part of the screen. The user then draws an arrow from the user image to the team member image under that descriptor (Figure 3). A slider is again used to assign values between -3 and +3 to the arrow. This also changes the color and the stroke of the arrow. The color scheme is the same as on the previous screen. Finally the user again writes a brief story on the right side of the screen justifying the button choice and the value assigned to the arrow.

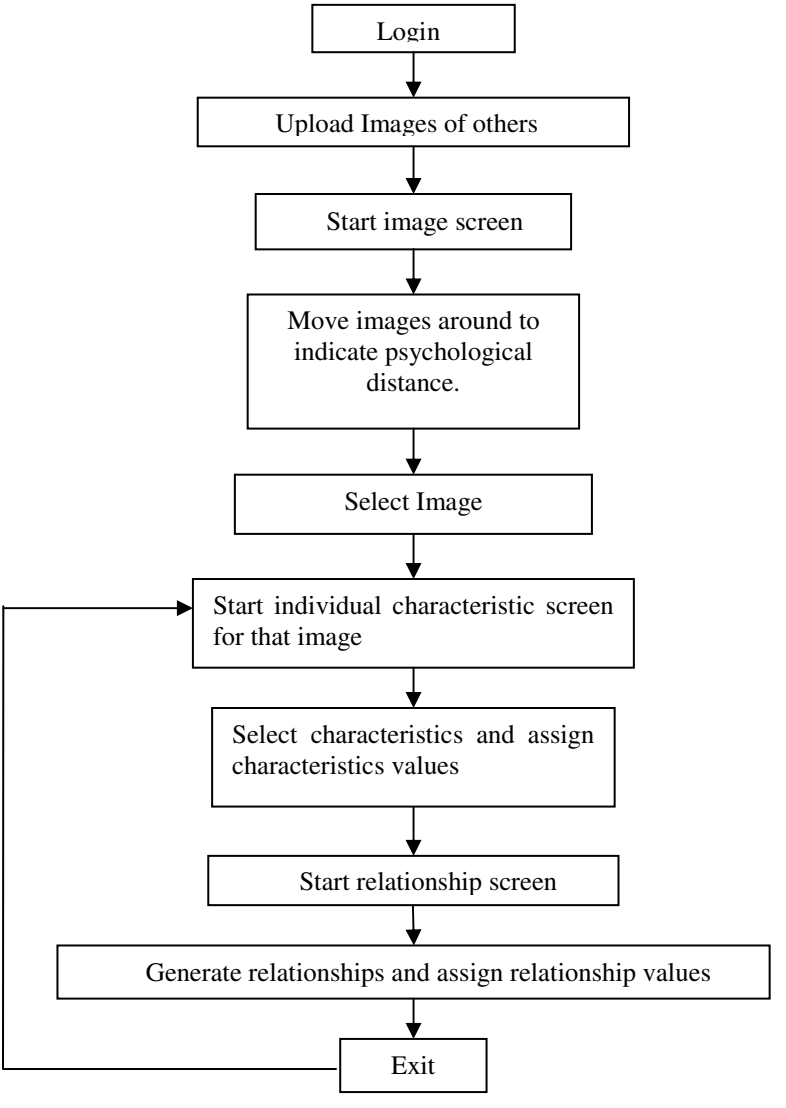


Fig. 1. Overall Control Structure

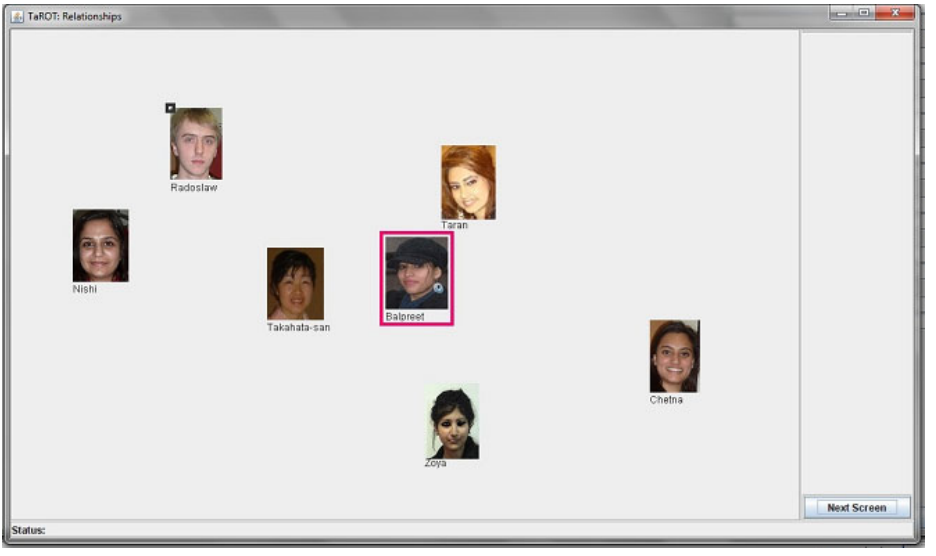


Fig. 2. Team Relationships

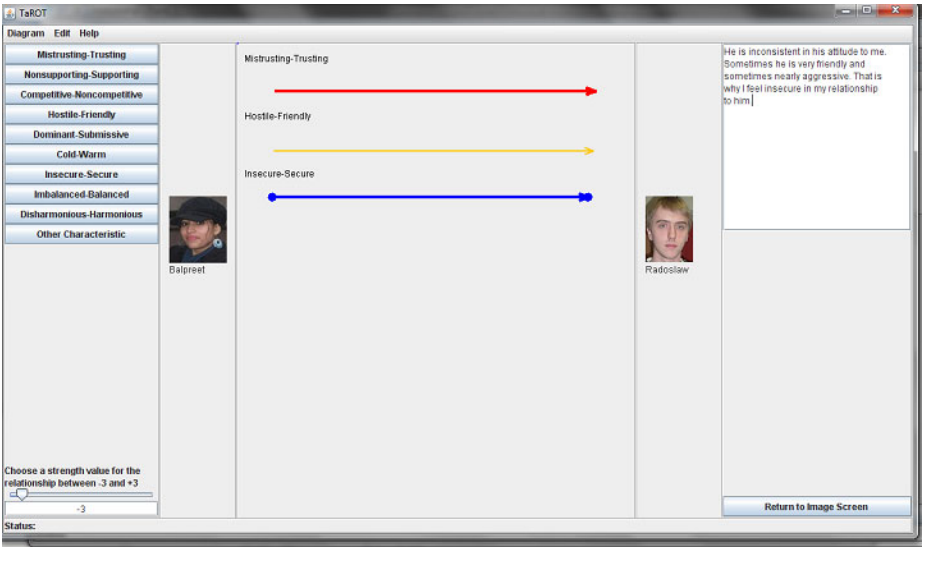


Fig. 3. Relationship between two team members

3 Study of Conflict and Harmony in a Security Team

The study aimed at identifying strength and weaknesses in the intra-team relations of a security team in a North London Shopping Center[2].

[2] We thank Linda Sakiwah-Adutwum for the collection of these data.

3.1 Method

Design. The study was designed as a exploratory case study with a security team of a London shopping centre.

Participants. The team consisted of 10 participants with an age range between 26 and 32 years. Participants are anonymized and denoted with letters from A through K.

Procedure. All participants were instructed to individually characterize their relationships within the team using the diagrams mentioned in Section 2. From the set of provided descriptors, they were free to choose descriptors they felt were appropriate. They also knew how to define new descriptors.

3.2 Results and Analysis

Each team member used the relationship characteristics of mistrusting-trusting and disharmonious-harmonious. Other characteristics were only used by subgroups of the team and are not reported here. None of the participants used the feature of creating new descriptors. The means of the numerical judgments made about team member by their fellow team members for the commonly used characteristics are shown in Table 1.

Table 1. Means for two relationship characteristics

Relationship Characteristics	A	B	C	D	E	F	G	H	J	K
mistrusting-trusting	0.22	0.33	-0.11	-0.11	0.11	0.00	0.11	0.11	0.67	0.11
disharmonious-harmonious	0.56	0.00	-0.22	0.11	-0.22	0.22	0.89	0.67	1.11	0.22

The table shows that participant J is both most trusted and the main contributor to team harmony. Participants B and G are complementary in that B still shows the second highest score on the trust scale, but has a neutral value on the harmony scale. In contrast participant G shows the second highest value on the harmony scale but is nearly neutral on the trust scale.

Participant C is the only participant who scores with negative values on both scales. With respect to negative assessments, participants D and E are complementary. Participant D scores negatively on the trust scale but positively on the harmony scale, whereas participant E shows the inverse pattern. In most cases, harmony and trust were poorly correlated. However, Participant A showed a correlation with relatively high positive harmony and trust values.

Due to the small amount of data, it was not possible to calculate the significance of these differences. However they are sufficient to localize problems within a team as well as positive influences. Both can then be a focus of further analysis.

4 Conclusion

This paper has described TaROT, a tool for acquiring data on intra-team relationships. TaROT is a stand-alone system and as such not a social networking site, although its

design was influenced by the Mixi site in that it also supports the use of aliases. Currently a Web-based version is under development, which is focusing on the communication of diagrams and will be more comparable with social networking approaches. As a data acquisition tool, it bears little relationship to various analysis systems that use graph-theoretical or network approaches. Note that the only graphs used are independent arrows. However, a follow-up project to TaROT currently investigates the visualization of relationships in entire teams. Here graph theory and related areas may have more of an impact. Currently, as in the study described above, chances for improving relationships and therefore team performance are identified by comparing characteristics of binary relationships. This allows identifying weak and strong relationships in a team. It is hoped that visualizations of an entire team will also lead to improved chance discovery.

References

1. Ohsawa, Y.: Modeling the Process of Chance Discovery. In: Ohsawa, Y., McBurney, P. (eds.) Chance Discovery. Springer, Berlin (2003)
2. Barker, S., Barker, R., Dawson, K., Knisely, J.: The use of the family life space diagram in establishing interconnectedness: A preliminary study of sexual abuse survivors, their significant others, and pets. Individual Psychology 53(4), 435–450 (1997)
3. Cheng, P., Lowe, R., Scaife, M.: Cognitive Science Approaches to Understanding Diagrammatic Representations. Artificial Intelligence Review 15, 79–94 (2001)
4. Hinde, R.: Relationships, A Dialectical Perspective. Psychology Press, Hove (1997)
5. Howard, R., Matheson, J.: Influence Diagram Retrospective. Decision Analysis 2(3), 144–147 (2005)
6. Larkin, J., Simon, H.: Why a Diagram is (Sometimes) Worth Ten Thousand Words. Cognitive Science 11, 65–99 (1987)
7. Oehlmann, R.: The Function of Harmony and Trust in Collaborative Chance Discovery. New Mathematics and Natural Computation 2(1), 69–84 (2006)
8. Ortony, A., Clore, G., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge (1998)
9. Soguet, G., Nelson, L., Safran, J.: The relationship between interpersonal schemas and personality characteristics. Journal of Cognitive Psychotherapy: An International Quarterly 15(2), 99–108 (2001)

Ontology-Based Knowledge Navigation Platform for Intelligent Manufacturing

Reiko Fujiwara, Akira Kitamura, and Kouji Mutoh

Tottori University, 4-101 Koyama-Minami, Tottori 680-8550, Japan
{reifw}@ybb.ne.jp, {kitamura}@ike.tottori-u.ac.jp,
{Kouji.Mutou}@jp.yokogawa.com

Abstract. This paper describes a knowledge-based manufacturing technology. We propose a knowledge navigation platform with ontology-based modeling for a user interface. This platform has a terminology connection search and can automatically support vocabulary collection and class definition for automatic generation of an ontology. The interface of the process model based on the ontology is expected to help users to recognize their standard procedure. Moreover, generation of an ontology from text resources may be an effective use of unstructured data. Both aspects exemplify the examination of use of ontologies with a user-centric approach for collaborative manufacturing.

Keywords: Collaborative Manufacturing, Ontology, Process Modeling, Text Mining, Connection Search.

1 Introduction

Due to the "Year 2007 problem" (2007 is the year the first baby-boomers reached retirement age), the need to plan for skill succession is now widely recognized in Japan [1], [2]. In addition, the shift to overseas manufacturing has caused a growing need for domestic technological capability in product development. It is crucial for international competitiveness to transfer experience of individuals to the organizational intellectual capital. To address these research topics, the SKIP (Self Knowledge and Information Expansion) [3] group at Tottori University investigated the issue of handling immense volumes of technical information through daily product development activities, such as specification sheets, design charts, analysis reports, and nonconformance reports. This issue has to be considered not only from an engineering viewpoint but also in the context of dynamic processes of production conducted by human beings [4].

In the research field of knowledge sharing and knowledge processing, the application of ontologies has proven to be an advantageous paradigm in recent years. Some knowledge management systems have been proposed based on ontology engineering for the field of manufacturing [5]. Although the study has made a great contribution to

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 447–456, 2011.

knowledge modeling or ontology building from the perspective of a knowledge expert who constructs the ontology, there seems to be little study from the user's viewpoint. In other words, no studies have considered how to represent an ontology for effective knowledge acquisition and how to deal with the increasing information in actual daily design or manufacturing settings. These two issues also indicate the user-centric approach of this study.

First, we introduce the concept of intelligent manufacturing, named "SKIP" and propose the ontology-based business modeling methodology to implement the navigation system in practice. Then, we discuss the dynamic ontology generation system. It is automation of new vocabulary extraction for the ontologies of further linked knowledge, which is tagged by a domain-dependent and in-house specific ontology.

The study also covers the following research items:

- A modeling tool with an ontology-based modeling framework for activity in manufacturing settings
- Approaches to utilize the business model as an interface of the knowledge navigation system
- Examination of an ontology-circulation system for continuous knowledge structuring.

2 Concept of Intelligent Manufacturing SKIP

We study intelligent manufacturing on a concept named "SKIP". The SKIP concept originated in the conflict between theory and practice in manufacturing technology. With the advances in technology, quality is not always as designers expect. Some of the core manufacturing techniques still rely on the experience and intuition of human experts.

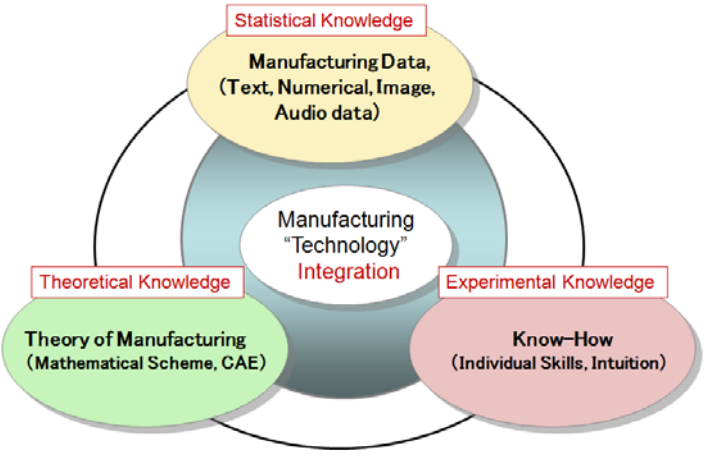


Fig. 1. Intelligent manufacturing as achieved by circulation and integration of three knowledge elements

To clarify this issue, we considered knowledge related to manufacturing technology. As a result, we defined manufacturing technology as integration and circulation of three elements of knowledge, namely "theoretical," "statistical," and "experimental". Figure 1 shows the concept of intelligent manufacturing of SKIP.

The most important knowledge is "theoretical knowledge" of manufacturing. "Theory" is a form of expression that has been proven correct mathematically and has universality of physical understanding. Possible symptoms can be predicted by the theory. For example, when considering new product design and manufacturing, simulation and theoretical predictions would be performed based on the technical theory.

However, the theory is only a part of the truth and may not represent all of the facts; in other words, it is an idealized structure. For instance, consider bending of metals: although bending of a material has been elucidated in the theory of continuum mechanics and plastic processing, the surface area of contacting surfaces and lubricants with tools and materials are not yet fully understood theoretically.

The next significant knowledge is probably "statistical knowledge," which includes text, numerical data, and other various forms of information related to production. These data include the facts to a limited extent, provided sufficient data are aligned. It is crucially different from "theory" in that we cannot create a product only by statistical information, especially for design or production of an innovative product.

The third element "experimental knowledge" is the so-called "know-how," which includes the sense of production and expertise of humans. Know-how is practical knowledge of how to get something done. It depends on individual skills and may not be readily explicable. Although some part of know-how can be explained theoretically and is proven to be experimentally reproducible, thus making it a universal fact, even experts cannot easily explain why their know-how works.

3 Manufacturing Knowledge Platform

Circulation and integration of the three elements of knowledge are achieved by organizational collaboration. Therefore, an interface must be available to people in a design or a manufacturing setting. We have prepared a manufacturing knowledge platform based on intelligent manufacturing. Theoretical knowledge is built into the "structured" information part, and experimental and statistical knowledge into the "unstructured" information part. Figure 2 shows the overall structure of SKIP and the relationship of the three elements of knowledge. The SKIP knowledge platform also has a retrieval part that enables a keyword cross-cutting search for structured and unstructured information. Figure 3 is the screen shot of the index page to access structured and unstructured information.

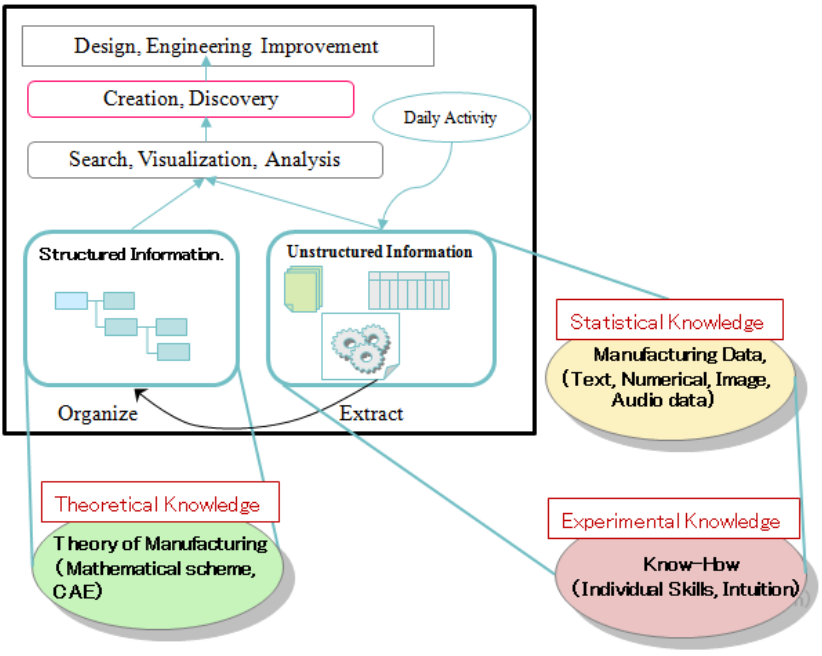


Fig. 2. Overall structure of SKIP and the relationship of three elements of knowledge

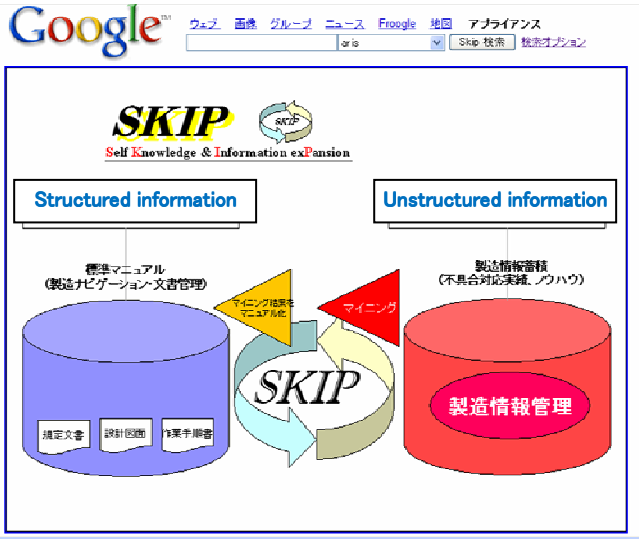


Fig. 3. Screen shot of the index page of the SKIP knowledge platform to access structured and unstructured information

3.1 Ontology-Based Modeling

Manufacturing is defined as a "process" including design and development to convert tangible and useful goods into tangible material [6]. Thus, we focus on "process" as a core element for "experimental knowledge." A process represents the tangible human activity and behavior performed by enterprise resources such as humans, facilities, and data. The function of the process and the resources can be statistically described as a structured hierarchy, and the process can be described as a dynamic relationship between the function and resources. An ontology is effective for organizing knowledge of manufacturing.

The approach of our ontology modeling provides not only a generic and methodological framework but also a well-documented, web-based business process model, thus enabling the practical application of our self-developed ontological modeling tool. Apart from showing the operational process, the tool semantically and hierarchically describes resources, data, activity, and products in the specific graphical notation. The task ontology is transformed into process model. Figure 4 shows the task ontology and the process model constructed by our tool.

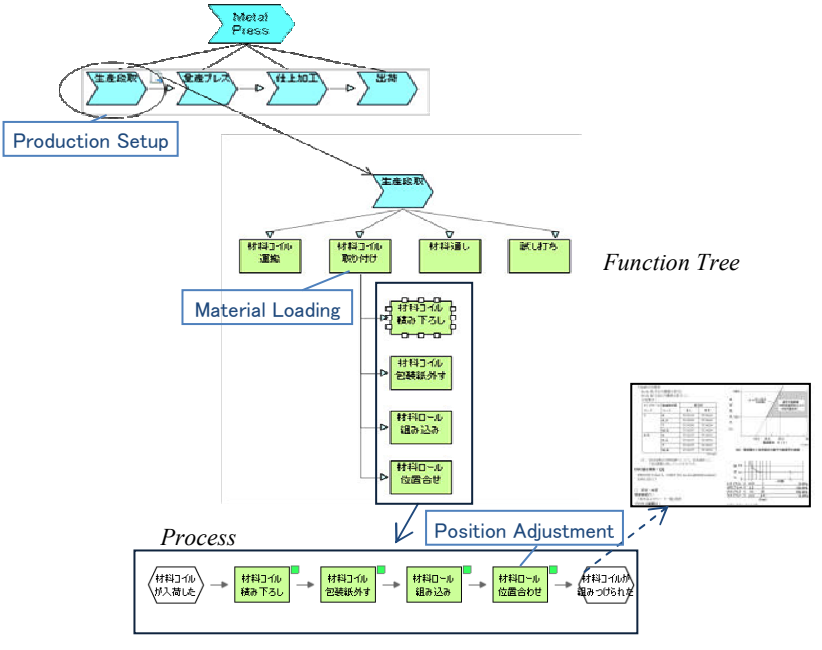


Fig. 4. Hierarchical task ontology and process model. The related information and documents are linked to each symbol of activity.

3.2 Knowledge Navigation

Users can access the same interface of the modeling view and navigate from one model to another by following links at each symbol. The related information and doc-

uments, both internal and external resources, are linked to each symbol of activity. Users may access information to verify their work step. This interface is also related to the "ba" concept [7]. Users can get accustomed to their own task since it is easy to access and acquire knowledge and then share it with others.

In this study, we considered the case of a metal press factory, where the graphical process model is equivalent to a "QC process sheet." The employees can access process-related information with text documents, moving images, and sounds in the form of internal or external links. Related information contains experts' tips, which are retrieved by interviewing them or from working memos. The business task ontology is the common communication backbone for sharing process-related knowledge.

3.3 Connection Search by Linked Data

The SKIP platform archives a variety of manufacturing terminology connections in the form of ontology, and the navigation can be traced from the links of process models.

Figure 5 shows examples of a "fault tree," which usually describes causal connections of undesirable events for the fault tree analysis. The primary objective here is to analyze from different viewpoints: product view and installing circumstances view. The two fault trees from different viewpoints have a common cause of failure; that is, we can infer that FailureX is caused by Installation EnvironmentB.

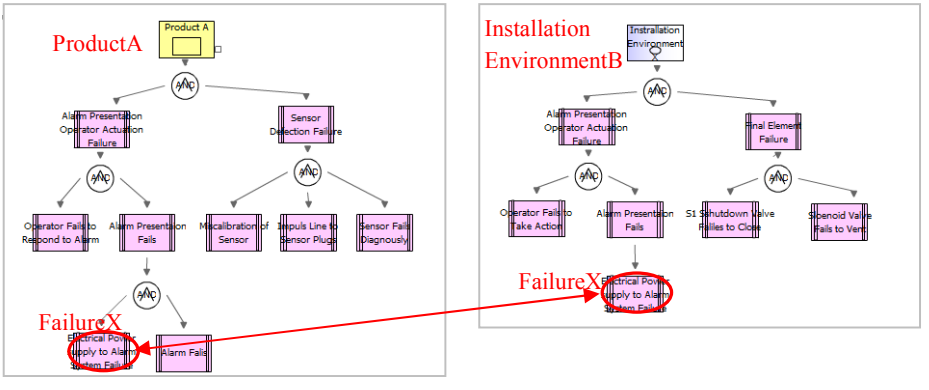


Fig. 5. Failure ontologies from different viewpoints, providing a common cause of failure

Furthermore, each class or instance of the ontology, which is described by our self-developed modeling tool, can have linked internal and external resources and also can activate a program. It may be possible to discover any internal and external related contents of each class or instance of the ontology by periodic activation of Web mining to find the related contents. In fact, during our trial of these multiple ontologies, we unexpectedly discovered the related paper "Hydrogen Effect against Hydrogen Embrittlement" [8] through the Google alert service using the designated keywords "metal, hydrogen embrittlement." This result may get a chance to find a solution for the principal undesirable effect of the diaphragm pressure sensor in our trial case.

4 Ontology Generation Support from Text Resources

The right side of the SKIP knowledge platform is the interface to collect and access various unstructured data created in daily manufacturing activities. Most of these are in the form of text, numbers, and images that require human interpretation and processing. However, there is a large amount of data for many years, so there must be a treasury of collective knowledge.

The purpose of this study is to extract "experimental knowledge" belonging to expert workers and transfer it to a machine readable structural format. Such unstructured data are important resources to find tacit knowledge to be transformed into a human- and machine-usable format. We extracted a keyword from the nonconformance report of a plant control device manufacturer by using text mining [9] and attempted to construct an ontology that represents knowledge for dealing with nonconformance. Figure 6 is part of the nonconformance data accumulated over the past few years. There is a textual description of each item, including the "description of nonconformance," "defect factor," and "corrective action."

現象名	不適合要素名	部位名	不適合要旨	不適合詳細	原因要約
その他の機能、動作、不良	部品の加工、寸法、位置、仕上げ不良	ｶﾊﾞｰﾌﾗﾝｼﾞ、ﾒｰﾀｰｶﾊﾞｰ、ｹｰｽ、端子盤、ゼロ調部、ｺﾈｸﾌﾟﾛ(計器外装周辺部)		状況: 伝送器不適合 作業内容: 1.ベントプラグと本体がかじっている。2.ｽﾀｰﾄｱｯﾌﾟ時にベントプラグと本体がかじっている状況を確認しました。(このSRは、QSRを発行するためのものです。)	出荷後に現地にてﾍﾞﾝﾄﾌﾟﾗｸﾞを締めた際にネジ先端部で"かじり"が発生したものと推定。QIR-EJA-5L12
その他の機能、動作、不良	かしめ、ねじ締め、かじり、インサート等	ｶﾊﾞｰﾌﾗﾝｼﾞ、ﾒｰﾀｰｶﾊﾞｰ、ｹｰｽ、端子盤、ゼロ調部、ｺﾈｸﾌﾟﾛ(計器外装周辺部)	端子部ｶﾊﾞｰが動かない	<状況> 端子部カバーを外していら途中に動かなくなった。(開閉どちらにも動かない)	保証SVにて処理
その他の機能、動作、不良	組立、取付、添付作業の不適合	ｶﾊﾞｰﾌﾗﾝｼﾞ、ﾒｰﾀｰｶﾊﾞｰ、ｹｰｽ、端子盤、ゼロ調部、ｺﾈｸﾌﾟﾛ(計器外装周辺部)	指定された仕様の右高圧・プロセスコネクタ上部取付に対し、納入品は仕様の異なる左高圧・プロセスコネクタ下部取付となってた。	貴社から指定されました仕様の右高圧・プロセスコネクタ上部取付に対し、納入品は仕様の異なる左高圧・プロセスコネクタ下部取付となっていました。	アンプケース組み付け時に向きを誤り、以後の目視確認において発見されず流出したと考えられる。CQR-JP-05015
浸水、冠水	組立、取付、添付作業の不適合	ｶﾊﾞｰﾌﾗﾝｼﾞ、ﾒｰﾀｰｶﾊﾞｰ、ｹｰｽ、端子盤、ゼロ調部、ｺﾈｸﾌﾟﾛ(計器外装周辺部)	指示振り切れ Err1.3.4.7発生。アンプケースに水が浸入していた。	状況: 指示振り切れ Err1.3.4.7発生 作業内容: ｱﾝﾌﾟｹｰｽに水が浸入していた。客先予備品と交換、引取原因調査いたします。	ゼロ調ねじ部Oリングに亀裂が入った為、水が浸入し電気回路が絶縁低下を起こし出力異常に至ったと推測する。CQR-JP-04085
その他	部品の加工、寸法、位置、仕上げ不良	ｶﾊﾞｰﾌﾗﾝｼﾞ、ﾒｰﾀｰｶﾊﾞｰ、ｹｰｽ、端子盤、ゼロ調部、ｺﾈｸﾌﾟﾛ(計器外装周辺部)	ゼロ調及びボリューム調査	EJA差圧伝送器ゼロ調及びボリューム調査を実施致しました。指示振り切れ(エラー7)となり、外部調整ボリュームがきかない(4台/8実施)	ゼロ点が大幅にずれておりエラーが発生していました。
	その他	ｶﾌﾟｾﾙ(ﾄﾞﾗｲﾊﾞASSYを除く)、ｾﾝｻ、ｴﾚﾒﾝﾄ	鹿島共同火力㈱鹿島共同発電所殿に昨年10月納入したトランスミッタ不具合に対する検討依頼	ゼロ点がずれる	指摘のありました 0.2%0.5%の変動は、周囲温度変化によって発生したものと判断されます。CQR-JP-06034
	異常なし、良品	ｶﾌﾟｾﾙ(ﾄﾞﾗｲﾊﾞASSYを除く)、ｾﾝｻ、ｴﾚﾒﾝﾄ	ダイアフラム膜(カプセル内部)表面が異形なため、調査をお願いします。客先殿は水素透過も懸念されていま	ダイアフラム膜(カプセル内部)表面が異形なため、調査をお願いします。客先殿は水素透過も懸念されています。	カプセル内部に気体が混入していることを示す状態は認められませんでした。QIR-EJA-8M11

Fig. 6. Part of the image of nonconformance reports indicating the "description of nonconformance," "defect factor," and "corrective action"

4.1 Vocabulary Extraction and Automatic Ontology Generation

When building such a domain-dependent ontology, it is unrealistic to build from scratch simply through listening, even though we need to extract experience and knowledge from experts by this method because of lack of systematized documentation. Therefore, we try to automatically generate an ontology by utilizing accumulated nonconformance reports as the information source for extracting a vocabulary of nonconformity.

We performed text mining on the indications of the items of "nonconformance summary," "details of nonconformance," and "summary of causes" and generate a base ontology using the lexical co-occurrence relations with the vocabulary extracted as a result. As a preliminary preparation, we extract vocabularies that mean cause or phenomenon in word frequency analysis and register the notational variants

as a thesaurus in the user's dictionary. In addition, we register the vocabulary identified as the same category as a group dictionary in phenomena and causes, respectively. These dictionaries can be reused as text mining parameters.

Next, we perform mining by using these dictionaries and generate a base nonconformance ontology from the vocabularies of failure event, causes, and their co-occurrence relations and generate an ontology from keyword graph data of text mining. The generated ontology is reused as grouping dictionary for subsequent keyword extraction.

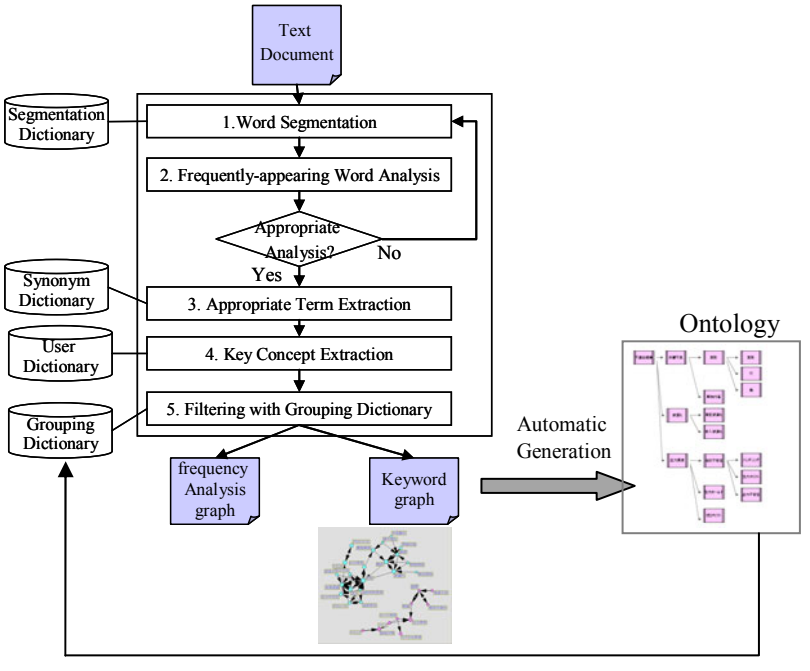


Fig. 7. Flow chart of keyword extraction by text mining and automatic generation of ontology. The generated ontology is reused as a grouping dictionary for subsequent text mining.

4.2 Ontology Maintenance Support

The ontology, which is daily increasing information such as a nonconformance reports, is not static. Maintenance can be reduced by applying semi-automatic generation of ontologies using text mining.

We examined mining of the nonconformance text data and evaluated the difference between the generated and previous ontology of our modeling tool. With the exception of the equivalent variations, new vocabularies were extracted. We checked the original text, consulted the co-occurrence with other words, and then added an instance of the ontology as needed. We verified better results in less time than that required in manual processing.

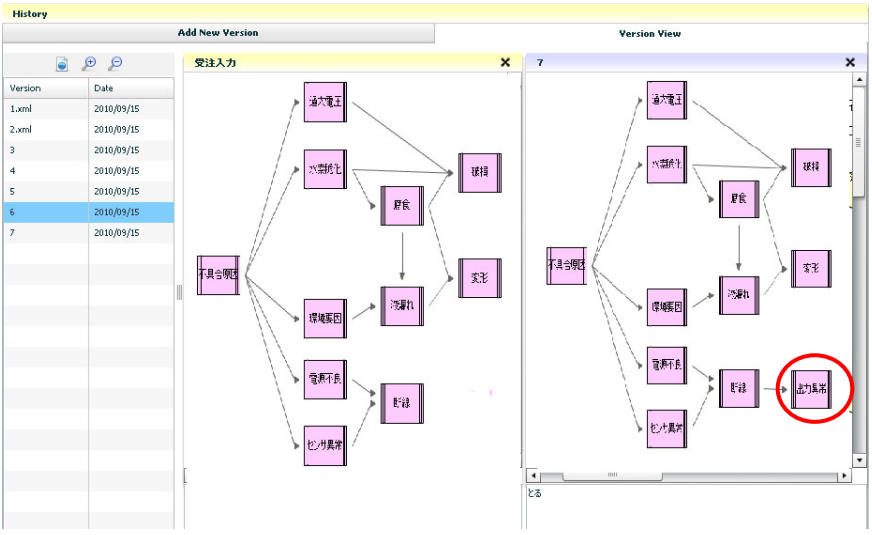


Fig. 8. Difference between the previous stored ontology and newly generated ontology of the failure event. The new item (circled in red) was found.

5 Conclusion and Future Work

In this paper, we proposed a manufacturing navigation platform with two aspects of use of an ontology. One is the ontology-based process modeling and the other is the ontology generation support by text mining.

The user interface of the ontology model is in trial use; it has been verified that users can recognize the standard process as enterprise knowledge faster with aggressive use of ontology presentation.

Moreover, the automatic ontology generator was effective to organize the unstructured data for domain-dependent knowledge. In the past, the process was conducted in conjunction with technical experts and knowledge analysts. This method has also proved useful for maintenance of a dynamic ontology. In the course of the trial, we realized that the ontology could be classified into two categories: static concepts independent so far, and a dynamic ontology for expanding concepts in a specific area.

In the ontology of the manufacturing space, increasing vocabulary is expected to be widely available for not only standardization of terms and concepts in the enterprise but also diagnosis and inference of information systems. We plan to continue trials of our method and update the manufacturing ontology by leveraging in practice.

References

1. Sano, S.: "Year 2007 problem" and manufacturing industry. - Imagine the short/medium/long term "Retirement-bubble" countermeasures – CIAJ. J. Comun. Inf. Netw. Assoc. Japan (2005)
2. Yamamoto, T.: Skill Transfer and the Role of Techno-workers Association for the Study of Industrial Management. Academic Journal 111, 92–95 (1997)

3. Fujiwara. R. et al.: Semantic Model-based Repository of Production Knowledge (2008)
4. Taura, T.: Topologies of Technological Knowledge (Series New Approaches to the Engineering Mind 1). University of Tokyo Press (1997)
5. Takafuji, S., Kitamura, Y., Mizoguchi, R.: Technical Knowledge Management Platform Based on Ontology Engineering and XML Technology. Journal of the Japanese Society for Artificial Intelligence 23(6), 424–436 (2008)
6. Takahiro. T.: Management Text, Technology and Operations Management, 4, Nikkei Inc. (2005)
7. Nonaka, I., Konno, N.: The concept of "ba": Building a foundation for knowledge creation. California Management Review 40(3), 40–45 (1998)
8. Murakami, Y., Kanezaki, T., Mine, M.: Hydrogen Effect against Hydrogen Embrittlement, open access at Springerlink.com. Metallurgical and Materials Transactions A 41(10), 2548–2562 (2010), doi:10.1007/s11661-010-0275-6
9. Ueda, T.: Text Mining: Learning through Case Studies. Kyoritsu Printing Company (2008)

Finding Top-N Chance Patterns with KeyGraph®-Based Importance

Yoshiaki Okubo[1], Makoto Haraguchi[1], and Sachio Hirokawa[2]

[1] Graduate School of Information Science and Technology
Hokkaido University
N-14 W-9, Sapporo 060-0814, Japan
{yoshiaki,mh}@ist.hokudai.ac.jp
[2] Research Institute for Information Technology
Kyushu University
Fukuoka 812-8581, Japan
hirokawa@cc.kyushu-u.ac.jp

Abstract. In this paper, as our first proposal, we discuss a method for finding a rare pattern, called a *chance pattern*, which connects a pair of more frequent patterns. Particularly, our chance pattern is defined with a *KeyGraph*®-based importance of patterns. More concretely speaking, a chance pattern is a pattern C which often appears in a part of documents containing a frequent pattern X_L as well as in those containing another pattern X_R, that is, confidence values of association rules, $X_L \Rightarrow C$ and $X_R \Rightarrow C$, are relatively high. It would be expected that such a chance pattern C reveals a hidden and implicit relationships between X_L and X_R. We design clique-search-based algorithms for finding chance patterns with Top-N confidence values.

1 Introduction

In this paper, we are concerned with a problem of mining *rare patterns* from a collection of documents, where a pattern is a set of words appearing in the collection. We say that a pattern is *rare* if it appears in a relatively small number of documents in the collection. Conversely, a pattern is said to be *frequent* or *general* if it is contained in many documents.

In traditional frameworks of *pattern mining* originated from [2], *higher frequency* has been regarded as the most important factor of interestingness of patterns. The idea is certainly true as has been widely accepted. It is, however, also true that patterns *not frequent* would be valuable in some situations and several methods for finding such rare patterns have been investigated (e. g. [10]).

From this point of view, the authors have been also investigating methods of finding *rare concepts* [11,12,13] based on *Formal Concept Analysis* [1], where a concept is an equivalent notion of *closed pattern* [9]. Particularly, the literature [11] has discussed an algorithm for extracting rare concepts each of which connects a pair of more frequent concepts, expecting that such a rare concepts

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 457–468, 2011.

reveals a *hidden* and *implicit relationship* between the connected concepts. We would find concrete examples of such rare concepts in *Ig Nobel Prizes*.

In 2009, the winner has been a research on creating *diamond film from tequila* in the field of *Chemistry*. *Diamond* would belong to a well-known concept "*precious stone*" and *tequila* to "*liquor*". This interesting and unexpected idea of the research might be triggered by *something* which is concerned with both "*precious stone*" and "*liquor*". We expect our rare concept to play such a valuable trigger.

In the framework in [11], we have to basically find *triples of concepts*, a rare concept and a pair of connected ones, in a concept space which is in general quite huge. In order to define a problem of finding *meaningful* triples of concepts which can be solved *exactly* (not *heuristically*), therefore, we are required to impose several constraints. As the result, the problem in [11] seems a little bit complex.

Rareness is a key notion also in the field of *Chance Discovery* [4] where *KeyGraph*® [1] is a well-known useful algorithm. *KeyGraph*® has been originally proposed in the literature [5] as an algorithm for indexing a document. The underlying idea has been applied to several actual domains (e.g. [3,6]). In a word, given a document to be indexed, *KeyGraph*® identifies a set of key words each of which co-occurs with some frequent terms. It should be noted here that each of the extracted key words is *not necessarily* frequent. They are rather *rare* but significant in the document.

As our first proposal, in this paper, we reformulate our problem of finding a rare pattern which connects more frequent patterns by taking a *KeyGraph*®-based importance of patterns. In previous *KeyGraph*®-based frameworks, each target to be extracted is supposed to be an individual word or event. In our framework, however, our target could be a combination of words or a compound event.

Our rare pattern is defined as a pattern C which is supported by a pair of frequent patterns X_L and X_R in the following sense: *"The pattern C often appears in a part of documents containing X_L, and also in those containing X_R"*. More technically speaking, this means that the confidence values of *association rules*, $X_L \Rightarrow C$ and $X_R \Rightarrow C$, are relatively high. In this sense, a relationship between X_L and X_R is implied by C. Moreover, by imposing a constraint for *uncorrelatedness* of words in C based on *Bond* measure [7], we can also expect C to be rare. Another measure based on *Bond* is also used for evaluating similarity among documents containing C. We call such a rare pattern a *chance pattern* and present algorithms for finding them with Top-N confidence values.

It should be emphasized here that as opposed to the framework in [11], we do not require that C must be a concept in our *concept lattice*. That is, our chance pattern could be just a pattern, not a closed pattern. In this sense, our chance pattern might be invisible in the concept lattice. Therefore, our chance of finding hidden and implicit relationships between X_L and X_R would be enhanced.

[1] *KeyGraph*® is a registered trademark owned by Yukio OHSAWA, The University of Tokyo, JAPAN.

2 Preliminaries

Let $\mathcal{W}$ be a set of *vocabulary words*. A *document* is written with words in $\mathcal{W}$.

Let $\mathcal{D}$ be a collection of documents. For a document $d \in \mathcal{D}$, the set of words occurring in d is referred to as W_d. For a set of documents $D \subseteq \mathcal{D}$, we denote the set of words occurring in every document in D by $Word(D)$, that is,

$$Word(D) = \{w \in \mathcal{W} \mid \forall d \in D,\ w \in W_d\}.$$

For a word $w \in \mathcal{W}$, the set of documents in $\mathcal{D}$ including w is denoted by $Doc(w)$, that is,

$$Doc(w) = \{d \in \mathcal{D} \mid w \in W_d\}.$$

Similarity, for a set of words $W \subseteq \mathcal{W}$, often called a *pattern*, we denote the set of documents in $\mathcal{D}$ including every word in W is denoted by $Doc(W)$, that is,

$$Doc(W) = \{d \in \mathcal{D} \mid W \subseteq W_d\} = \bigcap_{w \in W} Doc(w).$$

For a set of documents $D \in \mathcal{D}$, if $Doc(Word(D)) = D$, we say D is *closed*. Similarly, for a set of words $W \in \mathcal{W}$, W is said to be closed if $Word(Doc(W)) = W$. It should be noted here that for a set of documents $D \subseteq \mathcal{D}$, the pattern $Word(D)$ is always closed, since we have $Word(Doc(Word(D))) = Word(D)$. Dually, for a set of words $W \subseteq \mathcal{W}$, the set of documents $Doc(W)$ is also closed since $Doc(Word(Doc(W))) = Doc(W)$.

An *undirected graph* is denoted by $G = (V, E)$, where V is a set of *vertices*, $E \subseteq V \times V$ a set of *edges*. That is, any edge $(x, y) \in E$ is identified with (y, x).

For any vertices $x, y \in V$, if $(x, y) \in E$, x is said to be *adjacent to* y. For a vertex $x \in V$, the set of vertices adjacent to x is denoted by $N_G(x)$, that is,

$$N_G(x) = \{y \in V \mid (y, x) \in E\}.$$

If it is clear from the context, we simply denote it by $N(x)$.

For a subset X of V, a graph $G(X)$ defined by $G(X) = (X, E \cap X \times X, w)$ is called a *subgraph* of G and is said to be *induced by* X. If a subgraph of G is complete, then it is called a *clique* in G. A clique is simply referred to as the set of vertices by which it is induced.

For a pair of cliques in G, X and Y, if $X \subset Y$, then Y is called an *expansion* of X. Particularly, if there exists no clique Z such that $X \subset Z \subset Y$, Y is called an *immediate expansion* of X. If a clique has no immediate expansion, then it is said to be *maximal*.

3 Chance Patterns with KeyGraph®-Based Importance

We formalize in this section our notion of *chance patterns* and define a problem of mining them.

Roughly speaking, a chance pattern is defined as a rare pattern which can be *supported* by a pair of frequent patterns, where for patterns X and Y, "X *supports* Y" means that Y often appears in the documents containing X.

Such a chance pattern C supported by X_L and X_R would reveal a hidden connection between X_L and X_R because the pattern C is shared with both a part of documents with X_L and a part of those with X_R. Particularly, the connection would be more interesting if X_L and X_R are conceptually different patterns.

3.1 Rare Pattern as Weakly-Correlated Words

The underlying idea of rare patterns in this paper is found in [12]. In a word, we require a rare pattern must consist of *weakly-correlated general words*.

Let w_i and w_j be a pair of words in $\mathcal{W}$. A *correlation between* w_i *and* w_j *in* $\mathcal{D}$, denoted by $correl(w_i, w_j)$, is simply define as

$$correl(w_i, w_j) = \frac{|Doc(w_i) \cap Doc(w_j)|}{|Doc(w_i) \cup Doc(w_j)|}.$$

This is based on the same idea of *Bond* in [7] which is regarded as an extension of *Jaccard Coefficient*.

Based on the correlation, for any pair of words w_i and w_j in a pattern X, if $correl(w_i, w_j)$ is lower than a given threshold, then we consider the pattern to be *weakly-correlated*. From the definition of *correl*, a weakly-correlated pattern tends to be rare.

As will be mentioned later, we also take *generality* of words into account, where a word is considered general if it often appears in $\mathcal{D}$. From the generality of the words, it is expected that such a rare pattern can be concisely interpreted.

3.2 Conceptual Similarity of Documents

Given a pattern X, the set of documents containing X, $Doc(X)$, is uniquely determined. In order to make the pattern meaningful, it would be reasonable to require that $Doc(X)$ consists of documents which are *conceptually similar*. Therefore, any pattern to be extracted is imposed a constraint for this requirement.

Let d_i and d_j be a pair of documents in $\mathcal{D}$. A *similarity between* d_i *and* d_j is defined as

$$sim(d_i, d_j) = \frac{\left|W_{d_i} \cap W_{d_j}\right|}{\left|W_{d_i} \cup W_{d_j}\right|}.$$

The numerator is the number of words appeared in both d_i and d_j and the dominator the number of those included in d_i or d_j (or both). Thus, a higher value of $sim(d_i, d_j)$ shows that most of the words in d_i or d_j are shared in both documents. That is, they are expected to be *conceptually similar*.

As we see, the underlying idea of our similarity is the same as that for the correlation between words. Thus, the idea of *Bond* in [7] is used for both words and documents in our framework.

Based on the similarity, we can easily define a similarity among more than two documents. For a set of documents D, a *similarity among the documents in* D, denoted by $Sim(D)$, is defined as the minimum of the similarity between any pair of documents in D. That is,

$$Sim(D) = \min_{d_i, d_j \in D} \{sim(d_i, d_j)\}.$$

If D gives a higher value of Sim, we can expect D consists of documents which are conceptually similar.

3.3 Supporting Power of Base Patterns

Let X be a pattern in $\mathcal{W}$. Given a pattern $W \subseteq \mathcal{W}$, a *power of* X *for supporting* W is denoted by $power(X, W)$ and defined as

$$power(X, W) = \frac{|Doc(W) \cap Doc(X)|}{|Doc(X)|},$$

where X is called a *base pattern supporting* W. Roughly speaking, $power(X, W)$ corresponds to the conditional probability of W given X and is also known as the *confidence* of an association rule $X \Rightarrow W$ [2].

Let X_i and X_j be a pair of patterns in $\mathcal{W}$. Given a set of word $W \subseteq \mathcal{W}$, we simply define a *supporting power of* X_i *and* X_j *for* W, denoted by $SP(X_i, X_j, W)$, as the minimum of *power* of each base pattern, that is,

$$SP(X_i, X_j, W) = \min\{power(X_i, W), power(X_i, W)\}.$$

3.4 Problem of Mining Top-N Chance Patterns

Based on the above discussion, we can now define our problem of mining rare patterns, called *chance patterns*, each of which can be strongly supported by a pair of frequent base patterns with a Top-N value of support power.

Definition 1. (Top-N Chance Pattern Mining)
Let $\mathcal{D}$ be a set of documents, max_correl an upper bound of *correl* for weak-correlatedness of pattern and min_sim a lower bound of Sim for conceptual similarity within pattern. Then a problem of *mining Top-N chance patterns* is to extract triples of patterns (C, B_L, B_R) such that
(C1) Conceptual Similarity within Pattern:
$Sim(Doc(C)) \geq min\_sim$, $Sim(Doc(B_L)) \geq min\_sim$ and $Sim(Doc(B_R)) \geq min\_sim$.
(C2) Closedness of Base Patterns:
B_L and B_R are closed, that is, $Word(Doc(B_L)) = B_L$ and $Word(Doc(B_R)) = B_R$.
(C3) Minimality of Base Patterns:
B_L and B_R are minimal among patterns satisfying (C1) and (C2).

(C4) Disjointness of Base Patterns:
$Doc(B_L) \cap Doc(B_R) = \emptyset$.
(C5) Weak-Correlatedness within Chance Pattern:
$\max_{w_i, w_j \in C}\{correl(w_i, w_j)\} \leq max\_correl$.
Supporting Power of Base Patterns as Objective Function:
The value of $SP(B_L, B_R, C)$ is in the Top-N among triples of patterns satisfying the above five constraints. ∎

We impose five constraints on our patterns to be extracted. By the constraint C1, all of the documents with our pattern are expected to be conceptually similar. In this sense, our patterns are meaningful. The constraint C2 requires that our base pattern is the *intent* of a *formal concept* [1]. In other words, we can observe an exact (1-to-1) correspondence between our base pattern B and its document set $Doc(B)$ because any document except those in $Doc(B)$ never contains B. That is, our chance pattern intensionally connects conceptual groups of documents. The constraint C3 means that our base patterns must be as frequent as possible so that our chance pattern suggests a hidden relationship between general concepts we can easily observe. Interestingness of such a relationship would be further gained with the constraint C4 which requires the connected concepts to be extensionally disjoint, that is, non-overlapping. By the constraint C5, our chance patterns tend to be rare because the number of documents containing a combination of weakly-correlated words is generally expected to be small. Moreover, it should be noted that maximizing our objective function implicitly means we prefer a rare concept which consists of general (frequent) words. It is, therefore, expected that our chance pattern can be interpreted concisely and clearly.

4 Algorithms for Finding Top-N Chance Patterns

In this section, we discuss algorithms for finding Top-N chance patterns.

4.1 Basic Search Strategy

Our target triples (C, X_L, X_R) is extracted according to the following stages:

Enumerating Candidates of Base Patterns: We first enumerate the patterns which satisfy the constraints (C1), (C2) and (C3). The set of such patterns is denoted by $Bases$ and used as candidates of base patterns. Actually, each candidate in $Bases$ can be extracted as a minimal closed patterns which are also cliques in an undirected graph.

Identifying Candidate Pairs of Base Patterns: Then we identify the candidate pairs of base patterns in $Bases$ each of which satisfies the constraint (C4). The set of candidates is denoted by $Pairs$.

Finding Chance Patterns for Each Pair of Base Patterns: For each candidate pair $(X_L, X_R) \in Pairs$, we try to find patterns which can be strongly supported by X_L and X_R. Such a pattern can be also extracted as a clique in another undirected graph constructed based on the constraint (C5).

4.2 Enumerating Candidates of Base Patterns

Let $\mathcal{D}$ be a collection of documents and $\mathcal{W}$ the set of words appearing in $\mathcal{D}$. We here consider an undirected graph $G_{\mathrm{sim}} = (\mathcal{D}, E_{\mathrm{sim}})$, where the set of edges E_{sim} is defined as

$$E_{\mathrm{sim}} = \{(d_i, d_j) \mid d_i, d_j \in \mathcal{D}(i \neq j) \wedge sim(d_i, d_j) \geq min\_sim\}.$$

That is, an edge is created for any pair of documents in $\mathcal{D}$ if the documents are conceptually similar under the threshold of min_sim.

From the definition of our problem, every base pattern $X \subseteq \mathcal{W}$ to be extracted must satisfy the constraint (C1) on conceptual similarity within X. More concretely speaking, for any pair of documents d_i and d_j in $Doc(X)$, we have to observe $sim(d_i, d_j) \geq min\_sim$. It is easy to see that for such a pattern X, $Doc(X)$ forms a clique in G_{sim}.

X must also satisfy the constraints (C2) and (C3) for closedness and minimality, respectively. By (C2), it is required for X that $Word(Doc(X)) = X$. By (C3), furthermore, there must exist no pattern Y such that $Y \subset X$ and Y satisfies (C1) and (C2). Since the minimality of X corresponds to the maximality of $Doc(X)$, such a pattern X can be obtained based on a maximal clique in G_{sim}.

Let a set of documents Q be a maximal clique in G_{sim}. If Q is closed, that is, $Doc(Word(Q)) = Q$, then $Word(Q)$ is a minimal closed pattern satisfying (C1). On the other hand, if Q is not closed, that is, $Doc(Word(Q)) \supset Q$, $Word(Q)$ cannot satisfy (C1) even though $Word(Q)$ is closed. In this case, we can find some word $w \in \mathcal{W} \backslash Word(Q)$ such that the set of words $Word(Doc(Word(Q) \cup \{w\}))$ is closed and minimal among those satisfying (C1). Therefore, we can obtain all candidates of base patterns satisfying the constraints by enumerating each maximal clique Q in G_{sim} and then computing

$$B = \{Word(Doc(Word(Q) \cup \{w\})) | w \in \mathcal{W} \backslash Word(Q)\}.$$

Any pattern in B which is minimal under set-inclusion relation becomes a candidate of base patterns.

An algorithm for this task is presented in Figure 1. In the algorithm, for a set S, a function "$min(S)$" returns the set of minimal elements in S. Furthermore, the set of maximal cliques can be efficiently enumerated by several state-of-the-art algorithms (e.g. [8]).

4.3 Identifying Candidate Pairs of Base Patterns

Let $Bases = \{X_1, \ldots, X_\ell\}$ be the set of candidate base patterns obtained at the previous stage. For each pair of candidates X_i and X_j such that $i < j$, if they satisfy the constraint (C4) for disjointness, (X_i, X_j) is regarded as a candidate pair of base patterns which might be able to support some chance pattern.

More concretely speaking, for each candidate $X_i \in Bases$, we first compute $Doc(X_i)$, then for each $X_j \in Bases$ such that $i < j$, we check whether $Doc(X_i) \cap Doc(X_j) = \emptyset$ or not. If it is true, we regard (X_i, X_j) as a candidate pair of base patterns. The set of such pairs is denoted by $Pairs$.

procedure FindBasePatternCand($\mathcal{D}$, $\mathcal{W}$, min_sim):
[Input] $\mathcal{D}$: a set of documents.
$\mathcal{W}$: a set of words appearing in $\mathcal{D}$.
min_sim: a lower bound of Sim for conceptual similarity.
[Output] a set of candidate base patterns.

$Base \leftarrow \emptyset$;
construct $G_{\rm sim} = (\mathcal{D}, E_{\rm sim})$ under min_sim;
$\mathcal{M} \leftarrow$ the set of maximal cliques in $G_{\rm sim}$;
for each $Q \in \mathcal{M}$ **do**
begin
if Q is closed **then**
$Bases \leftarrow min(Bases \cup \{Word(Q)\})$;
else
$Bases \leftarrow min(Bases \cup \{Word(Doc(Word(Q) \cup \{w\}))|w \in \mathcal{W} \backslash Word(Q)\})$;
endif
end
return $Bases$;

Fig. 1. Algorithm for Enumerating Candidates of Base Patterns

4.4 Finding Chance Patterns for Each Pair of Base Patterns

At this stage, for each candidate pair of base patterns in $Pairs$, (X_L, X_R), we try to find chance patterns each of which, denoted by C, can be supported by X_L and X_R with strong supporting power.

Such a pattern C must satisfy the constraints (C1) and (C5) for conceptual similarity and weak-correlatedness, respectively. Since we already have the set of maximal cliques in $G_{\rm sim}$ computed at the first stage, we can simply find C satisfies (C1) by checking whether there exists a maximal clique M such that $Doc(C) \subseteq M$. If it is true, C satisfies (C1). In the following discussion, we focus on the constraint (C5).

Let us consider here another undirected graph $G_{\rm weak} = (\mathcal{W}, E_{\rm weak})$, where $E_{\rm weak}$ is defined as

$$E_{\rm weak} = \{(w_i, w_j) \mid w_i, w_j \in \mathcal{W} \wedge correl(w_i, w_j) \leq max\_correl\}.$$

We can easily observe that any pattern satisfying (C5) always forms a clique in the graph $G_{\rm weak}$. By examining any clique in $G_{\rm weak}$, therefore, we can extract chance patterns with strong supporting power.

Basically speaking, given a pair of base patterns (X_L, X_R), we examine cliques in $G_{\rm weak}$ one-by-one. For each clique C, we compute $SP(X_L, X_R, C)$. If the supporting power is in the top N among those already examined, then we keep (C, X_L, X_R) as a candidate of our targets. That is, during this process, we maintain a list which stores triples (C, X_L, X_R) with *tentative* Top-N supporting power. On completion of the process for every pair in $Pairs$, the list provides us the targets, that is, the triples with *actually* Top-N supporting power.

In general, since there are a large number of cliques in a given graph, it would be impractical to examine all of the cliques. The following observation is useful for excluding useless cliques which can never be our targets.

Observation 1
Let C and C' be patterns such that $C \subseteq C'$. For a pattern W, then

$$power(W, C) \geq power(W, C').$$ ■

Let us assume we already have a list of triples with tentative Top-N supporting powers, where the N-th power is γ. Given a pair of base patterns $(X_L, X_R) \in Pairs$, if for a clique C in G_{weak}, $power(X_L, C) < \gamma$ or $power(X_R, C) < \gamma$ holds, then any expansion of C, C', can never be supported by X_L and X_R with Top-N supporting power. Therefore, we can exclude such C's as useless cliques.

In order to enhance this pruning effect as much as possible, we explore all cliques in G_{weak} by expanding a clique step-by-step. More precisely speaking, for a clique Q, we can obtain an immediate expansion of Q by adding a vertex $\alpha \in N_{G_{\text{weak}}}(Q)$, where

$$N_{G_{\text{weak}}}(Q) = \bigcap_{v \in Q} N_{G_{\text{weak}}}(v),$$

that is, $N_{G_{\text{weak}}}(Q)$ is the set of vertices each of which is adjacent to every vertex in Q. Starting with the initial $Q = \emptyset$ and $N_{G_{\text{weak}}}(\emptyset) = \mathcal{W}$, we explore every clique in G_{weak} in *depth-first manner*. During our search, for a clique Q, if we find $power(X_L, Q) < \gamma$ or $power(X_R, Q) < \gamma$, we do not have to expand Q and can backtrack immediately. If $power(X_L, Q) \geq \gamma$ and $power(X_R, Q) \geq \gamma$, the tentative Top-$N$ list is adequately updated for the triple (Q, X_L, X_R). Then an immediate expansion of Q is generated and the same procedure is recursively performed for the expansion. The procedure is iterated until no Q remains to be examined. Figure 2 shows an algorithm for the task at this stage.

5 Preliminary Experimental Results

In this section, we present our preliminary experimental results.

A system based on our framework for finding Top-N chance patterns has been implemented in C and compiled with gcc-4.4.5. Our experimentation has been carried out on a PC with Intel® Core™2 Duo processor (L7700:1.80 GHz) and 2 GB main memory.

Our dataset is a collection of articles appeared in a Japanese Newspaper "*Mainichi*" in 1994. Especially, it consists of articles in the category of "Economy". The number of articles (documents) is 9,810. From the dataset, we have extracted only nouns. Then, too frequent and too infrequent nouns have been removed. The remaining 1,415-nouns are used as vocabulary words, that is, each article is represented as a set of those nouns appeared in the article.

Under the parameter setting, $max\_correl = 0.03$ and $min\_sim = 0.4$, we have tried to extract Top-3 chance patterns. One of the obtained chance patterns is as follows:

procedure FindPatternTriples($\mathcal{D}$, $\mathcal{W}$, $Pairs$, max_correl, N):
[Input] $\mathcal{D}$: a set of documents.
max_correl: an upper bound for weak-correlatedness.
N: an integer for Top-N.
[Global Variable] $\mathcal{M}$: the set of max. cliques found by FindBasePatternCand.
[Output] the set of pattern triples with Top-N supporting power.

$Triples \leftarrow \emptyset$; /* used as a list of tentative Top-N triples */
construct $G_{\text{weak}} = (\mathcal{W}, E_{\text{weak}})$ under max_correl;
for each $(X_L, X_R) \in Pairs$ **do**
begin
CliqueCheck($\emptyset$, X_L, X_R);
end
return $Triples$;

---

procedure CliqueCheck(Q, X_L, X_R):
[Input] Q: a clique in G_{weak} (a subset of $\mathcal{W}$).
X_L, X_R: base pattern candidates.
[G. Vars.] $\mathcal{M}$: the set of maximal cliques in G_{sim}
obtained by FindBasePatternCand.
$Triples$: the tentative Top-N list in FindPatternTriples.
G_{weak}: the graph constructed in FindPatternTriples.
[Side Effects] $Triples$ is updated.

---

if $\exists M \in \mathcal{M}$ such that $Doc(Q) \subseteq M$ **then**
if $Triples$ already contains a tentative N-th triple **then**
$\gamma \leftarrow N$-th supporting power in $Triples$;
if $power(X_L, Q) < \gamma$ or $power(X_R, Q) < \gamma$ **then**
return; /* based on Observation 1 */
else
$Triples$ is adequately updated for the triple (Q, X_L, X_R);
endif
endif
endif
for each $\alpha \in N(Q)$ **do**
begin
CliqueCheck($Q \cup \{\alpha\}$, X_L, X_R);
end
return;

Fig. 2. Algorithm for Finding Chance Patterns for Pair of Base Patterns

Chance Pattern: {`Business`, `Special`, `Improvement`} (Frequency : 2)

Base Pattern 1: {`Business`, `Release`, `Picture`, `Special`, `Compact`, `Equipment`, `Recreation`, `RV`, `Multipurpose`} (Frequency : 3)

Base Pattern 2: {`Price`, `Business`, `Release`, `Tokyo`, `Standard`, `Limited`, `Equipment`, `Nagoya`} (Frequency : 4)

Each document with the chance pattern is about a release information for a limited edition compact car with some particular improvement. The number of them (i.e. the frequency of the chance pattern) is just 2. On the other hand, each document with Base Pattern 1 is about a release information for a multipurpose compact RV with a picture and that with Base Pattern 2 about a release information for a limited edition car sold in some limited areas including Tokyo and Nagoya. The numbers of those are 3 and 4, respectively. The supporting power is 0.25.

Thus, our chance pattern is rare and shows some relationship between more frequent base patterns. However, the documents with Base Pattern 1 seem to be conceptually close to those with Base Pattern 2 while they are extensionally disjoint. Therefore, interestingness of the chance pattern would not be so high. We are required to introduce some additional constraints on conceptual difference between base patterns. Although we need to further examine and verify usefulness of our method, the authors expect that our chance patterns have potential ability to reveal hidden implicit relationships among frequent patterns.

The computation time for finding Top-3 chance patterns with the above parameter setting is 125 seconds, including times for constructing graphs. We have first obtained 34 base patterns satisfying the constraints. From them, 55 pairs of base patterns have been extracted. Then, for each candidate pair, it has been checked whether chance patterns can be supported by the pair of base patterns with Top-N supporting powers. Although some pruning rules are available in our search, computational performance of our algorithm must be improved for much larger datasets. Particularly, we might need some additional constraints concerned with meaningfulness of patterns.

6 Concluding Remarks

As our first proposal, we discussed a method for finding Top-N chance patterns with *KeyGraph*®-based importance. Our computation process for the targets was described and a clique-search-based algorithm for each computation stage was presented. We conducted some preliminary experimentations to verify usefulness of our method and got a feeling that our chance patterns would be expected to have potential ability of suggesting hidden relationships among more frequent patterns.

In order to make our chance patterns more valuable, we need to further improve the current framework from several points of views. In the current framework, we only focus on co-occurrence of a chance pattern and a base pattern as a supporting power. However, we might be required to take a *causal relationship* between a chance pattern and a base pattern into account so that our chance pattern can actually work as a valuable trigger for chance discovery. It is quite important to progress our future work in this direction.

If a chance pattern suggests a relationship between base patterns conceptually different, its interestingness and unexpectedness will be gained. It would be, therefore, also worth investigating conceptual distance between base patterns.

Furthermore, it is necessary to improve efficiency of our computation algorithms for larger scale datasets.

References

1. Ganter, B., Wille, R.: Formal Concept Analysis - Mathematical Foundations, 284 pages. Springer, Heidelberg (1999)
2. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. of the 20th Int'l Conf. on Very Large Databases - VLDB 1994, pp. 487–499 (1994)
3. Maeno, Y., Ohsawa, Y.: Human-Computer Interactive Annealing for Discovering Invisible Dark Events. IEEE Transactions on Industrial Electronics 54(2), 1184–1192 (2007)
4. Ohsawa, Y.: Discovery of Chances Underlying Real Data. In: Arikawa, S., Shinohara, A. (eds.) Progress in Discovery Science. LNCS (LNAI), vol. 2281, pp. 168–177. Springer, Heidelberg (2002)
5. Ohsawa, Y., Benson, N.E., Yachida, M.: KeyGraph: Automatic Indexing by Segmenting and Unifing Co-occurence Graphs. The IEICE Transactions on Information and Systems (Japanese Edition) J82-D-I(2), 391–400 (1999) (in Japanese)
6. Ohsawa, Y., Yachida, M.: Discover Risky Active Faults by Indexing an Earthquake Sequence. In: Arikawa, S., Nakata, I. (eds.) DS 1999. LNCS (LNAI), vol. 1721, pp. 208–219. Springer, Heidelberg (1999)
7. Omiecinski, E.R.: Alternative Interest Measures for Mining Associations in Databases. IEEE Transactions on Knowledge and Data Engineering 15(1), 57–69 (2003)
8. Tomita, E., Akutsu, T., Matsunaga, T.: Efficient Algorithms for Finding Maximum and Maximal Cliques: Effective Tools for Bioinformatics. In: Laskovski, A.N. (ed.) Biomedical Engineering, Trends in Electronics, Communications and Software, pp. 625–640. InTech, Vienna (2011)
9. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Efficient Mining of Association Rules Using Closed Itemset Lattices. Information Systems 24(1), 25–46 (1999)
10. Szathmary, L., Napoli, A., Valtchev, P.: Towards Rare Itemset Mining. In: Proc. of the 19th IEEE Int'l Conf. on Tools with Artificial Intelligence - ICTAI 2007, pp. 305–312 (2007)
11. Okubo, Y., Haraguchi, M.: An Algorithm for Finding Indicative Concepts Connecting Larger Concepts Based on Structural Constraints. In: The 9th Int'l Conf. on Formal Concept Analysis - ICFCA 2011, Contributions to ICFCA 2011, pp. 53–68 (2011)
12. Okubo, Y., Haraguchi, M.: An Algorithm for Extracting Rare Concepts with Concise Intents. In: Kwuida, L., Sertkaya, B. (eds.) ICFCA 2010. LNCS (LNAI), vol. 5986, pp. 145–160. Springer, Heidelberg (2010)
13. Okubo, Y., Haraguchi, M., Nakajima, T.: Finding Rare Patterns with Weak Correlation Constraint. In: Proc. of the 2010 IEEE Int'l Conf. on Data Mining Workshops - ICDMW 2010, pp. 822–829 (2010)

Quantitative Evaluation Method of Criticism in Value Creating Conversation

Yoko Nishihara and Yukio Ohsawa

Department of Systems Innovation, School of Engineering, The University of Tokyo
7-3-1, Hongo, Bunkyo, 113-8656 Tokyo, Japan
{nishihara,ohsawa}@sys.t.u-tokyo.ac.jp
http://www.panda.sys.t.u-tokyo.ac.jp

Abstract. Creative activity performed by a group of people needs human communications in which people discuss and think up new ideas for solving their problems. The communication is useful for finding defects and improvements of new idea. Many business companies tend to record communications of idea creation and make transactions of communication. The transactions contain valuable information for idea creation. However, the transactions are merely read because of the low amount of contained information. New methods for extracting valuable information from the transactions are required.

This paper proposes a new method for extracting valuable information from conversation records. We verified that extracted information by our method was valuable information for idea creation.

Keywords: Question and Criticism, Value Creation, Human Communication, Creative Activity, Business Chance Discovery.

1 Introduction

Creative activity is a process of making new and important products and items [2]. We see many creative activities in business companies making new commercial items and new industrial products. Research activities in academic universities and colleges are also the creative activities.

People doing creative activity have purposes to be achieved. People think up new ideas to achieve their purposes (action 1), then create new products and items based on thought up ideas (action 2). The process of creative activity is run by running over the two actions. If people obtain valuable ideas, they can run the process smoothly. The most important thing is to obtain valuable ideas.

When thinking up new ideas, group conversations often work well. The proverb *four eyes see more than two* represents the goodness. A group conversation gives hints to people which are obtained from individuals' viewpoints. The more hints are obtained, the more ideas are improved into new ideas less defects. Opinion exchanges in a conversation are made naturally by people for obtaining good ideas.

In group conversations, useful information is provided to rise the value of idea. Business companies have aggressively conversation records to utilize in the

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 469–477, 2011.

future conversations [8]. The recorded transactions, unfortunately, will rarely be utilized. Because information obtained from reading the recorded transactions is fewer though it takes much time to read the recorded transactions. It is necessary for rising up the value of idea to utilize conversation records in the future conversations. It is also necessary for utilizing the recorded transactions to organize the recorded transactions by extracting useful information from the transactions.

This paper proposes a new method for extracting utterances including useful information from conversation records. We focus on conversations in which new ideas for creative activity are thought up. The authors think that ideas are not useful just after being thought up. It is necessary to be reconsidered by referring other's comments for obtaining useful ideas. For example, good research themes can not be obtained without any discussions by members of research team. When a researcher thinks up new a new research theme, he/she have to discuss the new research theme to be improved by other's comments and criticisms. If he/she gets many comments and criticisms, the comments and criticisms should be ordered according to importance to be solved. The motivation of our research is to improve thought up ideas by extracting useful information from conversational records.

2 Related Work

Our work is related to the previous works about structuration and visualization of discussion, and communication in creative activity. This section introduces the previous works and shows the differences between the previous works and our work.

2.1 Structuration and Visualization of Conversation

Conversations have to be held over and over again. Reflecting the recorded transactions about past conversations is necessary to make future conversations fruitful ones.

To support for reflecting past conversations, the simplest way is to write and publish the past conversation records. However, the amount of words included in a conversation record is too much to read. The important information which should be reflected is few in a conversation record. The conversation records have to be organized.

Many researchers have proposed methods of structuration and visualization of conversation [5,4,6]. The previous methods can automatically extract related utterances from transactions to represent the relationships between utterances in a conversation. The methods can also detect separating points of conversation. These functions of structuration and visualization assist people in understanding conversation flows and conversation perspectives. Then, people can revolve new conversations based on findings and knowledge obtained from reflections of the past conversations.

On the other hand, the previous works have been focused on understanding conversation perspectives, not on understanding conversation details such as important utterances. When supposing conversations for thinking up new ideas for creative activity, not only conversation perspectives, but also conversation details such as what opinions have been given to ideas are quite important. If people can grasp the detailed information easily, reflections of conversation records are more supported, and idea thinking is also supported. Our research proposes a new method for extracting utterances including information which will be useful for rising up the value of new idea. The method matches extracted utterances with the thought up ideas for organizing structured conversation records.

2.2 Communication in Creative Activity

As one of the communication methods for thinking up new ideas, brainstorming has been used for a long time[9]. Brainstorming is a group creativity technique designed to generate a large number of ideas for the solution of a problem. There are four basic rules in brainstorming. These are intended to reduce social inhibitions among group members, stimulate idea generation, and increase overall creativity of the group: (1) withhold criticism, (2) welcome unusual ideas, (3) focus on quantity, and (4) combine and improve ideas. Brainstorming emphasis on the quantity of ideas as shown by the rule (3). Therefore, the rule (1) and the rule (2) for making no limitation in people's thoughts have been set. Even if people obtain many ideas, only a few ideas can be utilized in creative activity. Moreover, people have to consider the quality of idea for achieving the purpose of creative activity. To obtain useful ideas utilized in creative activity, it is considered that people need to make conversations without the rule (1) and the rule (2).

What can people do to rise up the value of idea? The theory advocated by Popper who was a researcher in the area of scientific philosophy should be referred. Popper said that the value of theory was risen by obtaining criticisms and defeating the criticisms[10]. The theory by Popper has been limited to the value of scientific theory. It is considered that his theory can be applied to the value of idea in creative activity. The value of idea thought up in creative activity may be risen up by obtaining and defeating the criticisms.

Criticisms are sometimes given by asking questions. For example, a question *your system can be used in this situation?* is considered as a criticism *your system is not considered to be used in this situation.* If the criticisms and questions are used effectively in conversations thinking up new ideas, it has been discovered that the value of idea is risen[7]. Eris has been discovered that there were two questions which were good for industrial design: Deep Reasoning Question and Generative Design Question[1]. When obtaining requirements from customers, positive/negative questions to intention, purpose, and technique to apply requirements are possibly able to reveal constraints having by customers[3]. These researches intended to follow Popper's research.

It has been discovered that utterance styles of criticism and question are useful for rising up the value of idea thought up by people. However, few researches have tried to find linguistic features of the utterance styles. Therefore, this paper finds the linguistic features and proposes a new method for extracting effective utterances of criticism and question. We will accelerate to understand the details of conversation by structuring, and to make reflections of conversation.

3 New Conversation Model: Bitters Do Good to the Stomach

Figure 1 shows our conversation model. The model consists of four factors: conversation topic, sub-topics relating a conversation topic, attendees to a conversation, and utterance styles given to a conversation. The attendees give sub-topics respectively and converse each other.

A various types of utterance styles are given in a partial conversation related to a sub-topic. Let us suppose that a criticism is given by an attendee in a partial conversation. When criticisms are given, attendees have usually found defects of ideas. That is to say, criticisms can make attendees notice lacked information from a conversation topic. If they can improve defects, a conversation topic will be risen. The other attendee replies against a criticism. Defects are improved by replying in a partial conversation. If a partial conversation ends shortly, a criticism does not give an important information for attendees. However, if a partial conversation continues for a long time, a criticism gives a quite important information which should be considered deeply. Moreover, if a criticism is objective, a deep consideration is meaningful. If a criticism is objective, and a partial conversation continues for a long time, the conversation model evaluates a criticism as important information which is useful for rising up efficiency of a topic[1].

3.1 Algorithm of Utterance Extraction Based on a Proposed Conversation Model

This section explains our algorithm of utterance extraction based on the proposed conversation model. When a conversation text is given, the algorithm gives weights of rising up efficiency of a topic to each utterance, then extracts utterances of high weights.

The algorithm gives weights to utterances as follows. Firstly, a given conversation text T is separated into partial texts related to each sub-topic. The algorithm defines an utterance as a set of words given by an attendee until the other attendee starts to utter. If two attendee exchange utterances continuously, it is considered that they converse about a sub-topic. Therefore, when separating

[1] We name the conversation model *Bitters Do Good to the Stomach*. The reason is that strict criticisms and questions which can not be agreed easily exactly contribute to rise up the value of idea.

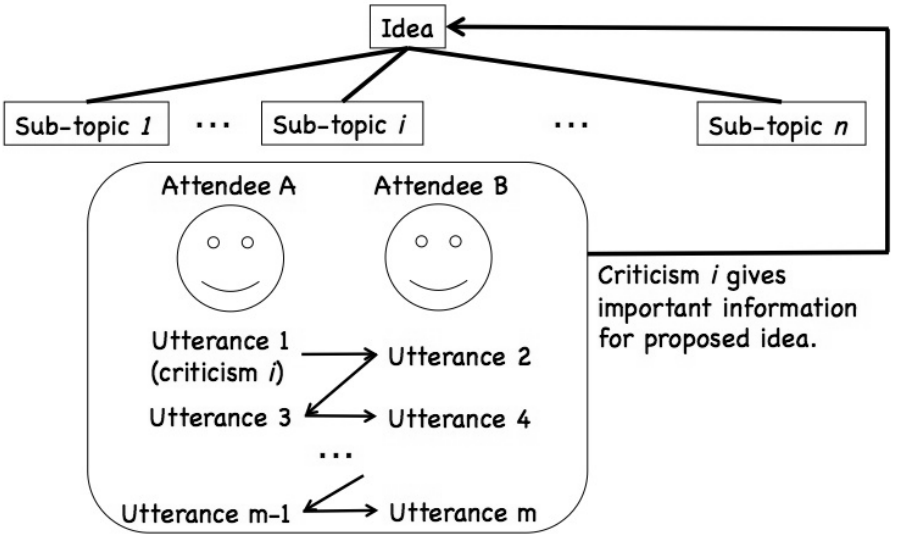


Fig. 1. Proposed conversation model

a given conversation text into partial texts, the algorithm focuses on uttering attendees, and detects a set of continuous utterances by two attendees as a partial text.

Next, the algorithm judges whether or not each utterance style is a criticism. Words are extracted by a parser (we used a Japanese parser called ChaSen in the following experiment) from each utterance $s_i (i = 1, 2, ..., n)$. If words represent a style of criticism, the algorithm judges an utterance style is a criticism[2].

Finally, the algorithm gives utterances weights $evaluation_p(s_i)$ by using Equation (1). The depth of conversation made by a criticism is evaluated by the number of utterances until a next criticism, and the objectiveness of criticism is evaluated by the number of kinds of words included in an utterance.

$$evaluation_p(s_i) = len(s_i) \times words(s_i), \tag{1}$$

where $len(s_i)$ denotes the number of utterances until a next criticism, and $words(s_i)$ denotes the number of kinds of words included in an utterance s_{i+1}.

4 Experiment

We experimented with the proposed method. The purpose of experiment is to verify that the proposed method can extract utterances which are useful for rising up the value of topic.

[2] Our algorithm can work only on Japanese texts. Particles and auxiliaries which are used for describing criticisms in Japanese have been chosen for judging criticism.

4.1 Experimental Procedure

The procedure of experiment was as follows:

exp1) Transcribe a conversation records about idea thinking up into a text.
exp2) Extract utterances from a text which are useful for rising up the value of idea.
exp3) Compare extracted utterances with utterances which are judged as useful by participants.

Table1 shows texts used in the experiment. When inputting a text into the proposed method, an idea was regarded as a topic. We made small texts including utterances relating to a topic, and then input small texts into the proposed method.

To evaluate the efficiency of extraction by the proposed method, we prepared a baseline method which extracts utterances by evaluating word weights by tfidf[11]. The baseline method extracted equal number of utterances of the proposed method.

To compare the efficiencies between the proposed method and the baseline method, we calculated precisions and recalls of utterance extraction. The precision and the recall were calculated by using Equation (2) and Equation (3).

$$precision = \frac{\#\ of\ utterances\ common\ with\ extracted\ and\ chosen}{\#\ of\ extracted\ utterances} \tag{2}$$

$$redall = \frac{\#\ of\ utterances\ common\ with\ extracted\ and\ chosen}{\#\ of\ chosen\ utterances} \tag{3}$$

The chosen utterances were defined by questionnaires to participants. The number of participants was 25 who were graduate/undergraduate students majoring engineering. We asked the participants to read texts and extract utterances which were judged as triggers for rising up the value of idea. We defined chosen utterances which were extracted by more than three participants.

Table 1. Conversation records used in a experiment. The number of attendees, the number of ideas, the number of utterances and the number of chosen utterances judged by participants are shown.

Data	# of attendees	# of ideas	# of utterances	# of chosen utterances
#1	6	16	596	70
#2	7	20	1212	103
#3	12	13	380	47
#4	11	12	321	52
#5	11	17	730	98

4.2 Experimental Result

Figure 2 and Figure 3 show experimental results about the precisions and the recalls. The precisions of the proposed method were higher than that of the baseline method ($t = 12.7, p < 0.01$). The recalls of the proposed method were higher than that of the baseline method ($t = -3.5, p < 0.01$). Next section discusses the reasons of obtaining high precisions and high recalls referring to the extracted utterances.

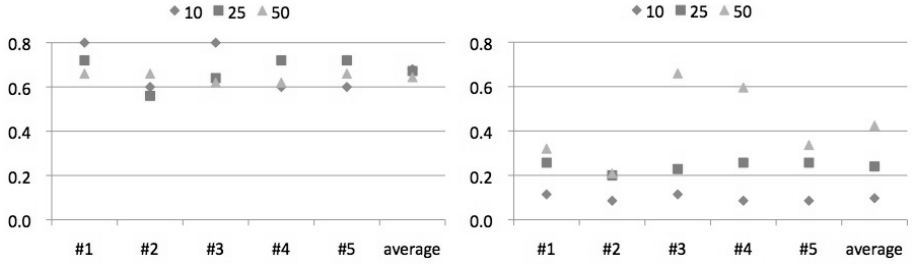


Fig. 2. Precisions and recalls obtained by a proposed method. The horizontal axis denotes conversational records. The vertical axis denotes values of precision and recall. The numbers 10, 25, and 50 denote the numbers of extracted utterances by a proposed method.

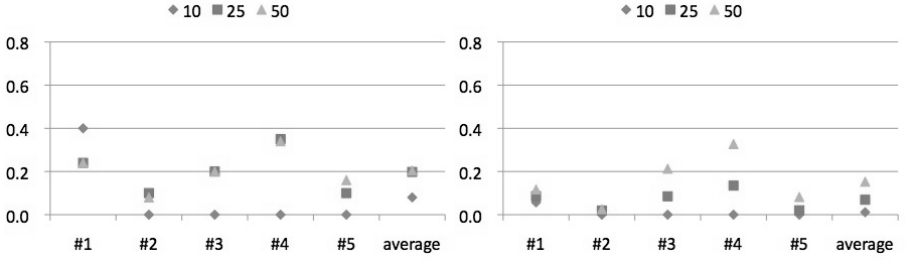


Fig. 3. Precisions and recalls obtained by a baseline method. The horizontal axis denotes conversational records. The vertical axis denotes values of precision and recall. The numbers 10, 25, and 50 denote the numbers of extracted utterances by a baseline method.

5 Discussion

Table 2 shows an example of a separated text from #2 text. The sub topic in Table 2 is about idea of new machine putting bath agent into bathtub automatically to assist people's comfortable bath time. The first utterance *I would like to put the bath salt in by myself and not automatically* in Table 2 was the chosen utterance. This utterance criticized that putting bath salt into bathtub by

Table 2. Example of a separated text from #2 text. Italic font sentence is the chosen and extracted sentence by the proposed method.

Attendee	Extracted utterance by the proposed method
A	*I would like to put the bath salt in by myself and not automatically.*
B	This system can take a solid bath salt, not a liquid one.
A	I would like to watch the broadening of the bath salt by adding myself to a bath.
B	You would like to watch it, wouldnft you?

himself/herself was the very enjoyable action. After this utterance, three utterances were given, but the problem was not solved. An attendee of this utterance might think the value of idea was risen if the defect was improved. However, it was difficult to improve the defect because other attendees did not accept the sales point of proposed idea. If the sales point is removed from the proposed idea, the idea does not have any features for sales. It is considered that the difficulty for improvement is shown in the length of exchanges of utterances.

The result indicates that utterances contributing to rise up the value of idea are usually given by criticisms and questions. If it is difficult to solve problems pointed by the questions and criticisms, the length of utterance exchanges becomes long. This phenomenon matches with our conversation model. Therefore, the proposed method could extract utterances with high precisions and high recalls.

Table 3 shows an example of a extracted utterance from #2 text by the baseline method. The idea shown in Table 3 is about a system which can move with measuring the present location. The system may be useful for discovering mines under the ground and arresting strangers barricading themselves inside. For this idea, the baseline method extracted an utterance *the idea can work in a time of disaster like earthquake. An attendee understood the reason of attaching GPS to the system.* The baseline method evaluated utterances including words with high weights of tfidf. The tfidf method evaluates words appearing frequently in an utterance and words not appearing in other utterances. If the length of utterance is long, the utterance has been prepared by the attendee for uttering. The utterance included no new information at least for the attendee. The utterance did not have any potential power for rising up the value of idea. Therefore, the baseline method could not extract the chosen utterances.

Table 3. Example of an extracted utterance by the baseline method from #2 text. Italic font sentence is the extracted sentence by the baseline method.

Attendee	Extracted utterance by the baseline method
E	I see. Your system can be used for saving living people.
F	Exactly.
E	*Your system could work to help children in Sisen (a city in China) earthquake. To achieve the mission, your system needs GPS sensors. It is quite natural.*

6 Conclusion

This paper proposes a extracting method of utterances from a conversation record. The method has developed based on our conversation model in which criticisms and questions give a topic power to rise up the value of topic.

We experimented with the proposed method, and verified that the proposed method could extract utterances which were useful for rising up the value of idea. The precisions and recalls in extraction by the proposed method were higher than that by the baseline method using word weights of tfidf.

As the future work, we will try to support reflections of conversation by using the proposed method. We will also try to make a framework for assisting communications for value creation.

References

1. Eris, O.: Effective Inquiry for Innovative Engineering Design. Kluwer Academic Publishers, Dordrecht (2004)
2. Hori, K.: A System for Aiding Creative Concept Formation. IEEE Transactions on Systems, Man, and Cybernetics 24(6), 882–894 (1994)
3. Kushiro, N., Ohsawa, Y.: A Requirement Acquisition Process as an Evolved Chance Discovery, Chance Discoveries in Real World Decision Making. SCI, vol. 30, pp. 315–328 (2006)
4. MacLean, A., Young, T.M., Bellotti, V., Moran, T.: Questions, Options, and Criteria: Elements of design space analysis. In: HCI (1991)
5. Matsumura, N., Ohsawa, Y., Ishizuka, M.: Mining and Characterizing Opinion Leaders from Threaded Online Discussions. In: Proceedings of the 6th International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies, pp. 1267–1270 (2002)
6. Moran, T.P., Carroll, J.M.: Design Rationale: Concepts, Techniques, and Use. Lawrence (1996)
7. Nishihara, Y., Ohsawa, Y.: Communication Analysis focusing Negative Utterances in Combinatorial Thinking Games. The Review of Socionetwork Strategies 4(2), 31–46 (2010)
8. Nonaka, I., Takeuchi, H.: The Knowledge-Creating Company. Oxford University Press, Oxford (1995)
9. Osborn, A.F.: Applied imagination: principles and procedures of creative problem-solving. Creative Education Foundation (1953)
10. Popper, K.R.: The Logic of Scientific Discovery. Routledge, New York (1959)
11. Salton, G.: Developments in Automatic Text Retrieval. Science 253, 974–979 (1991)

Chance Discovery and Unification in Linear Modal Logic*

Vladimir V. Rybakov

School of Computing, Mathematics and IT,
Manchester Metropolitan University,
Manchester M1 5GD, UK, and
Institute of Mathematics, Siberian Federal University, Krasnoyarsk, Russia
V.Rybakov@mmu.ac.uk

Abstract. This paper studies an interpretation of the Chance Discovery (CD) as the unification problem in the linear modal logic $S4.3$. We prove that any formula unifiable in $S4.3$ has a most general unifier, and we give an algorithm which, for any unifiable formula, constructs a most general unifier. More precisely, we merely write out, explicitly, the formulas giving substitutions for a most-general unifier. From computational viewpoint, we find an algorithm for the construction of sets of 'best unifiers' (complete collections of 'most' general unifiers) in the linear modal logic $S4.3$.

Keywords: chance discovery, CD, modal logics, linear modal logic, Kripke models, logical unification, unifiers.

1 Introduction, Background

This paper investigates applications of logical methods to innovative research area in Knowledge Engineering: Chance Discovery, via study logical unification in linear modal logics from the research area. Chance Discovery (CD in the sequel, cf. Ohsawa and McBurney [16], Abe and Ohsawa [1]) is a modern direction in Artificial Intelligence (AI) which analyzes important events with uncertain information, incomplete past data, so to say, *chance* events, where *a chance* is defined as some event which is significant for decision-making in a specified domain. The prime aim of CD is determination of methods for discovering various chance events. Research in AI for CD is mostly focused on practical applications, this direction formed a solid branch in knowledge representation (KR in the sequel) and information sciences (cf. e.g. [1,2,3,4,17,18,19,14,15,19]). From theoretical viewpoint, properties of CD in logical systems originating in AI (as well as some interpretations of CD-like operations via existence, without direct reference to CD) were also studied (cf. Rybakov [26] – [31,24,25]) in some degree (though, interpretation of CD in these papers was primarily modeled via an

* Supported by Engineering and Physical Sciences Research Council (EPSRC), U.K., grant EP/F014406/1.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 478–485, 2011.

introduction of new logical operations based on the existence question). Thus, these papers aimed to model CD as only various kind of logical operations.

In current paper we would like to investigate CD for internal logical problems connected with unifiability in linear modal logic. Logical unification is the problem to answer whether, for two given formulas, there is a substitution, which makes these two formulas equivalent w.r.t. a given logic L, and, if yes (and if possible), to describe all such substitutions. It is relevant to say that logical methods showed to be very effective and useful for CS (cf., e.g. [7,8]).

Unification problematic initially was originated in computer science. There it was the problem of making two given terms semantically equal via a substitution of terms in place of variable-letters. That is, if there is a substitution, for two given terms, which makes them semantically equal: such that the values of these two resulting terms are the same w.r.t. all models and all interpretation of variables of these terms. Even, at the very beginning, unification in computer science started as the task: whether two given terms may be turn to syntactically equal, by replacing their variables by terms; if yes they were said to be unifiable; this term was independently introduced in automated deduction by Robinson [32] and in term rewriting by Knuth et al [14]. Than it was suggested that instead of making terms syntactically equal, it is relevant to consider the semantic equivalence: then the all possible real values the terms would be the same. Since then, all instruments of mathematical logic and universal algebra might be involved in the research (cf. the survey Baader and Ghilardi [6]) Since, for the most part, non-classical logics from AI and CS areas are duals of equational theories of special algebraic systems (as modal, pseudo-boolean, temporal, etc. algebras) this problematic directly coincide with its logical-unification counterparts.

From the viewpoint in this paper, the unification is the problem: if there is a chance to make two formulas to be equivalent via a substitution some terms in place of atomic letters (and when this chance exists, the terms are said to be unifiable). From applications viewpoint, logical unification (being, in particular, a base for logic programming, e.g. Prolog language) was situated at the focus of interest from logical community and computer science for a reasonably long time. Unification problem (weather a formula can be unified in a given logic) is, in fact, a particular case of more complicated problem: the substitution problem: whether a formula can be made a theorem after replacing a part of variables (keeping the same value for coefficients — parameters), which was studied and solved, Rybakov [21,22,23]), for intuitionistic logic and modal logics $S4$ and Grz (but only to determine if a solution exits and to compute a particular one if yes).

S. Ghilardi [9,10,11,12,13] studied extensively the unification in intuitionistic logic and propositional modal logics over K4 (using ideas from projective algebras and applying a technique of special projective formulas) with aim to describe all possible unifiers. In these works the problem of construction best unifiers, in logics under consideration, was solved and algorithms for computation best unifiers were constructed. This approach showed also to be very useful in dealing with admissibility and bases of admissible rules, actually in presence computable finite sets of best unifiers, a solution of admissibility problem immediately follows.

Innovative, impressive results of S. Ghilardi ([10] – [12]) on unification, in particular, complete description of best unifiers in modal logic $S4$ (cf. [12]) were giving computational ground to describe (construct, compute) all unifiers. Such methods turned out to be very flexible, bringing effective computational methods to work with problems. This impressive beak down in unification in non-classical logics set up a numerous problems concerning unification in relevant, close non-classical logics.

In our paper we study the unification problem (as a logical interpretation for CD) for the modal linear logic $S4.3$. We show that any formula unifiable in $S4.3$ is projective; and we explicitly construct a projective substitution (the proof and construction are very short and occupy hardly a half of page). This immediately gives a description of a set of best unifiers for any unifiable formula: any this set consists of only one unifier (so, any unifiable formula has a most general unifier). From computational background, we find an algorithm for the construction of 'best' unifiers (complete collections of 'most' general unifiers) in the linear modal logic $S4.3$.

2 Definitions and Preliminaries

We begin with providing information necessary for reading this paper. Main object for our research will be linear modal logics, at which we will interpret the CD notion. The language of modal logics consists of the language of Boolean logic extended by two unary modal operations: $\Diamond$ (to be read – possible) and $\Box$ (necessary). Extension of the definition for standard Boolean formulas is: if φ is a formula then $\Diamond\varphi$ is a formula also (to be read φ *is possible*), and, if φ is a formula, then $\Box\varphi$ is a formula also (to be read φ *is necessary*).

Semantics for modal logics which we will use here is Kripke-Hintikka models. To recall, a Kripke/Hintikka frame is a pair $\mathcal{F} := \langle F, R\rangle$, where F is the base of $\mathcal{F}$ – a non-empty set, and R is a binary (accessibility) relation on F. $|\mathcal{F}| := F$, in the sequel, $a \in \mathcal{F}$ is a denotation for $a \in |\mathcal{F}|$. In this paper we consider only reflexive, transitive and linear frames, i.e. R, in the sequel, is always reflexive, transitive and linear relation, that is

$$\forall a(aRa);$$

$$\forall a,b,c[(aRb) \wedge (bRc) \Rightarrow (aRc)];$$

$$\forall a,b,c[(aRb) \wedge (aRc) \Rightarrow [(bRc) \vee (cRb)].$$

If, for a set of propositional letters P, a valuation V of P in $|\mathcal{F}|$ is defined, i.e. $V : P \to 2^F$, in other words, $\forall p \in P(V(p) \subseteq F)$, then the tuple $\mathcal{M} := \langle \mathcal{F}, V\rangle$ is called a Kripke-Hintikka model (structure). The truth values of formulas are defined at elements of $\mathcal{F}$ by the following rules:

$$\forall p \in Prop,\ \forall a \in \mathcal{F},\quad (\mathcal{F}, a) \Vdash_V p \Leftrightarrow a \in V(p);$$

$$(\mathcal{F}, a) \Vdash_V \varphi \wedge \psi \Leftrightarrow (\mathcal{F}, a) \Vdash_V \varphi \text{ and } (\mathcal{F}, a) \Vdash_V \psi;$$

$$(\mathcal{F}, a) \Vdash_V \varphi \vee \psi \Leftrightarrow (\mathcal{F}, a) \Vdash_V \varphi \text{ or } (\mathcal{F}, a) \Vdash_V \psi;$$

$(\mathcal{F}, a) \Vdash \varphi \to \psi \Leftrightarrow \neg[(\mathcal{F}, a) \Vdash_V \varphi]$ or $(\mathcal{F}, a) \Vdash_V \psi$;

$(\mathcal{F}, a) \Vdash_V \neg\varphi \Leftrightarrow \neg[(\mathcal{F}, a) \Vdash_V \varphi]$;

$(\mathcal{F}, a) \Vdash_V \Diamond\varphi \Leftrightarrow \exists b \in \mathcal{F}((aRb) \wedge (\mathcal{F}, b) \Vdash_V \varphi)$;

$(\mathcal{F}, a) \Vdash_V \Box\varphi \Leftrightarrow \forall b \in \mathcal{F}((bRa) \wedge (\mathcal{F}, b) \Vdash_V \varphi)$.

For any $a \in \mathcal{F}$, $Val_V(a) := \{p_i \mid p_i \in P, (\mathcal{F}, a) \Vdash_V p_i\}$. For any formula φ, $V(\varphi) := \{a \mid a \in \mathcal{F}, (\mathcal{F}, a) \Vdash_V \varphi\}$. For $a \in \mathcal{F}$, $C(a) := \{b \mid (aRb) \wedge (bRa)\}$, i.e. $C(a)$ is the cluster containing a.

Definition 1. *For a Kripke-Hintikka structure $\mathcal{M} := \langle \mathcal{F}, V \rangle$ and a formula φ, φ is* true in $\mathcal{M}$ *(denotation – $\mathcal{M} \Vdash \varphi$) if $\forall a \in \mathcal{F}$ $(\mathcal{F}, a) \Vdash_V \varphi$. $\mathcal{F} \Vdash_V \varphi \Leftrightarrow \forall w \in \mathcal{F}((\mathcal{F}, w) \Vdash_V \varphi)$.*

Definition 2. *For a Kripke-Hintikka frame $F := \langle \mathcal{F}, R \rangle$ and a formula φ, φ is* true in F *(denotation – $F \Vdash \varphi$) if $\forall V$, $\forall a \in \mathcal{F}$ $[(\mathcal{F}, a) \Vdash_V \varphi.]$*

Definition 3. *For a set of frames SF, $L(SF) := \{\varphi \mid \forall F \in SF, F \Vdash \varphi\}$. $L(SF)$ is said to be the modal logic generated by SF.*

For example the modal logic $S4.3$ may be represented as $S4.3 = L(K)$, where K is the set of all finite reflexive transitive, linear and routed frames (i.e. $\forall F \in K$ there exists $a \in F$, such that $\forall b \in F$, (aRb)).

Representation of chance discovery (CD in the sequel) in formalism of modal logics follows from interpretation of possibility. Being at a world w of a Kripke model $M = \langle F, R, V \rangle$, we say a statement φ represented by a formula φ is possible if $(M, w) \Vdash_V \Diamond\varphi$. That is we mean to say that the chance for a formula to be true from viewpoint of w holds. So, this sketchy interpretation suggests $CD = \Diamond$, though it would not model CD enough precisely. Several variations of this interpretation were studies before in Rybakov [31]. We may reflect the essence of CD (as an event which is difficult to discover) within suggested formalism in many more ways.

Example 1. For a natural number $k \in N$, and any formula φ consider the formula $M_{CD(k)}\varphi := \Diamond\varphi \wedge \bigwedge_{1 \le i \le k}[\Diamond\neg\varphi \wedge p_i \wedge [\bigwedge_{1 \le j \le k, j \ne i} \neg p_j]]$. This formula would say that there are at least k different evidences against φ while yet φ may be true.

Example 2. Take the formula $\Diamond\varphi \wedge \Box(\Diamond p \to \neg\varphi)$. This formula says that there is a chance for the formula φ to be true, but only if the sate satisfying φ is close enough to the terminating state of the run: if there is no chance for the statement p to be true since this state.

Example 3. Consider the formula $\Diamond\varphi \wedge \Box(\Diamond p \to \neg p) \wedge \Diamond\Box p$. This formula tells us that, since current state, there is a chance for φ to be true, but this chance may happen only in the middle part of the run because at the terminating cluster the statement p will be true and it is inconsistent with φ.

In this paper we will study CD for internal logical problems connected with unifiability. In general, CD is a special sort of the existence problem for the case if the solution exists and the solution is rare, non-trivial. The logical unification, generally speaking, considers the problem: if two (different) statements may be made equivalent modulo to some replacement of atomic letters by terms. That is, if there is a chance to make this statements equivalent by a substitution some terms in atomic letters. In a any case, when this chance exists, the terms are said to be unifiable. Consider as example the terms $x \wedge \Box y$ and $x \vee z$. The question is if they are unifiable in modal logic is simple. The question if $x \wedge \Box y \equiv x \vee z$ may be solved has (for instance) an immediate answer: the substitution $x \to \Box p$, $y \to \Box p$, $z \to \Box p$ unifies these terms. But we do not see immediately what could be a set of all solutions. Unification problem deals with the question 'if the chance to make terms equal via substitutions may be satisfied', and if yes to describe all possible substitutions. Below we start with precise definitions and description the problem.

Let For be the set of all formulas and ε be a mapping (we will refer to ε as substitution) of a set of letters $Dom(\varepsilon)$ in For. Any such mapping ε can be extended to the set of all formulas in letters form $Dom(\varepsilon)$, by

$$\varepsilon(\varphi(x_1, \ldots, x_n)) := \varphi(\varepsilon(x_1), \ldots, \varepsilon(x_n)).$$

Definition 4. *A formula φ is unifiable in a logic L if there is a substitution ε (which is called a unifier for φ) such that $\varepsilon(\varphi) \in L$ (sometimes, in the sequel, notation $\varphi^{\varepsilon} := \varepsilon(\varphi)$, if it is more readable in complex expressions, may be used).*

Definition 5. *A unifier ε (for a formula φ in a logic L) is more general than a unifier ε_1 iff there is a substitution δ such that for any letter x, $[\varepsilon_1(x) \equiv \delta(\varepsilon(x))] \in L$.*

If a logic L is decidable, to check the unifiability a formula in L is (theoretically, not computationally) an easy task: it is sufficient to use only ground substitutions: mappings of variable-letters in the set $\{\bot, \top\}$. But the problem is how to find all unifiers - all solving substitutions.

Definition 6. *A set of unifiers SU for a given formula φ in a logic L is a set of* best *unifiers, if the following holds. For any unifier σ for φ in L, there is a unifier σ_b from SU, where σ_b is more general than σ (i.e. σ is a substitutional example of σ_b).*

Definition 7. *A formula φ is said to be projective in a logic L if the following holds. There is a substitution σ (which is called projective substitution) such that $\Box\varphi \to [x_i \equiv \sigma(x_i)] \in L$ for any letter x_i from φ.*

Theorem 1. *Any unifiable in $S4.3$ formula is projective.*

Proof. Let a formula $\varphi(x_1, ..., x_n)$ built out of letters x_i be unifiable in $S4.3$ (it is, easy to verify by ground substitutions). Let, for any subset V of the set of all subformulas $Sub(\varphi)$ of the formula φ (symbolically – $X \subseteq Sub(\varphi)$),

$$\Psi(X) := [\bigwedge_{\psi \in X} \Diamond[\psi \wedge \Box\varphi] \wedge [\bigwedge_{\psi \in Sub(\varphi) \setminus X} \neg\Diamond[\psi \wedge \Box\varphi].$$

If $\Psi(X)$ is satisfiable in $S4.3$ (in the sequel we denote it by $\Psi(X) \in Sat$), we may compute a rooted finite model M of bounded size for $\Psi(X)$ (using decidability of $S4.3$ and, say, filtration), satisfying $\Psi(X)$ at a world of the minimal cluster. Take w to be a world from next (up) cluster from M to the maximal one, where $\Box\varphi$ is false; if $\Box\varphi$ is true at all worlds, take w from the minimal cluster of M.

Let $Char(\Psi(X)) := \{x_i \mid (M, w) \Vdash x_i\}$, and, for any x_i, $T(\Psi(X), x_i) := \top$ if $x_i \in Char(\Psi(X))$, otherwise $T(\Psi(X), x_i) := \bot$.

Since φ is unifiable in $S4.3$, φ is true at the single world model with a valuation V. If $V(x_i)$ is empty, we set $T(x_i) := \bot$ and we set $T(x_i) := \top$ otherwise.

For any letter x_i, occurring in φ, we define the following substitution:

$$\sigma(x_i) := (\Box\varphi(x_1, ..., x_n) \wedge x_i) \vee$$

$$\left(\neg\Box\varphi \wedge \Diamond\Box\varphi \wedge \bigvee_{\Psi(X)\in Sat} [\Psi(X) \wedge T(\Psi(X), x_i)]\right) \vee$$

.

$$(\neg\Diamond\Box\varphi \wedge T(x_i)) .$$

It is immediate to see that σ is a projective substitution for φ. In fact, σ unifiers φ, which is easy to verify. Indeed, take any finite $S4.3$-model M_1 with a valuation of letters of φ.

If (1) $\Box\varphi$ is true at all worlds of M_1, σ does not change the truth values of φ, so $\sigma(\varphi)$ is true at all worlds of M_1.

If (2) $\Diamond\Box\varphi$ is not true at the minimal cluster of M_1, truth values of $\sigma(x_i)$ are constant and coincides with truth values of $T(x_i)$ and hence $\sigma(\varphi)$ is true all all worlds of M_1.

(3) Assume $\Diamond\Box\varphi$ is true at the minimal cluster of M_1, bit $\Box\varphi$ – not, and C is the minimal cluster of M_1 where $\Box\varphi$ is true. At all worlds accessible form C, $\sigma(\varphi)$ is true as in case (1) above. If a is any world from M_1, which is not accessible from C, some unique $\Psi(X)$ is true at all such a. Then the truth values of all $\sigma(x_i)$ coincide at all such a with truth values of x_i at the world w in M. And then, easy verifiable by induction on length of φ, $\sigma(\varphi)$ has the same truth value at all such a as φ at w in M. This means $\sigma(\varphi)$ is true at all such a.

Thus, $\sigma(\varphi)$ is true at all finite $S4.3$ models. Hence σ is projective. Q.E.D.

Lemma 1. *If a substitution σ_p is projective for a formula φ in a logic L, then $\{\sigma_p\}$ is a set of best unifiers for φ (i.e. σ_p is most general unifier).*

Proof. Indeed, let σ be a unifier for φ in L. Since we assume σ_p is projective for φ in L, we have $\Box\varphi \to [x_i \equiv \sigma_p(x_i)] \in L$ for any letter x_i from φ. Acting by σ on the formula above we get $\sigma(\Box\varphi) \to [\sigma(x_i) \equiv \sigma_p\sigma(x_i)] \in L$, that is $\sigma(x_i) \equiv \sigma_p\sigma(x_i) \in L$. Q.E.D.

Thus, Theorem 1 gives us a construction of a most general unifier for any formula φ unifiable in $S4.3$. The construction is given in the proof explicitly, – we just write out directly the formulas $\sigma(x_i)$ giving most general unifier.

3 Conclusion

We investigated here only a part of questions connected with unification and possible treatments of the CD. An interesting open question is to investigate problems of unifiability in $S4.3$ (and other modal logics) for formulas with coefficients, those where we consider a part of letters as variables and the remaining ones as coefficients (which have to be intact w.r.t. any substation). It is more difficult and interesting problem. Another portion of open questions concern the unification with parameters in non-linear modal logics, as $K4$, $S4$ and relevant ones. Besides the open question of modeling CD in modal logics or temporal logics with more preciseness would be essential, because until now it is modeled primarily as only various versions of the existence problem.

We feel that the technique developed in this paper may help to answer these questions as well as to clarify more the essence of CD and applications of CD problematic to AI and CS.

References

1. Abe, A., Ohsawa, Y. (eds.): Readings in Chance Discovery. International Series on Advanced Intelligence (2005)
2. Abe, A., Kogure, K.: E-Nightingale: Crisis Detection in Nursing Activities. In: Chance Discoveries in Real World Decision Making, pp. 357–371 (2006)
3. Abe, A., Ohsawa, Y.: Special issue on chance discovery. KES Journal 11(5), 255–257 (2007)
4. Abe, A., Hagita, N., Furutani, M., Furutani, Y., Matsuoka, R.: Exceptions as Chance for Computational Chance Discovery. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 750–757. Springer, Heidelberg (2008)
5. Barringer, H., Fisher, M., Gabbay, D., Gough, G.: Advances in Temporal Logic. Applied logic series, vol. 16. Kluwer Academic Publishers, Dordrecht (1999)
6. Baader, F., Ghilardy, S.: Unification in modal and description logics. Logic J. of IGPL, (2010) doi: 10.1093/jigpal/jzq008 (First published online: April 29, 2010)
7. Gabbay, D.M., Hodkinson, I.M., Reynolds, M.A.: Temporal Logic: - Mathematical Foundations and Computational Aspects, vol. 1. Clarendon Press, Oxford (1994)
8. Gabbay, D.M., Hodkinson, I.M.: An axiomatisation of the temporal logic with Until and Since over the real numbers. Journal of Logic and Computation 1, 229–260 (1990)
9. Ghilardi, S.: Unification Through Projectivity. J. Log. Comput. 7(6), 733–752 (1997)
10. Ghilardi, S.: Unification, finite duality and projectivity in varieties of Heyting algebras. Ann. Pure Appl. Logic 127(1-3), 99–115 (2004)
11. Ghilardi, S.: Unification in Intuitionistic logic. Journal of Symbolic Logic 64(2), 859–880 (1999)
12. Ghilardi, S.: Best solving modal equations. Annals of Pure and Applied Logic 102, 183–198 (2000)
13. Ghilardi, S., Sacchetti, L.: Filtering Unification and Most General Unifiers in Modal Logic. Journal of Symbolic Logic 69(3), 879–906 (2004)

14. Hahum, K.S.: The Window of Opportunity: Logic and Chance in Becquerel's Discovery of Radioactivity. In: Physics in Perspective (PIP), vol. 2(1), pp. 63–99. Birkhäuser, Basel (2000)
15. Magnani, L.: Abduction and chance discovery in science. International J. of Knowledge-Based and Intelligent Engineering Systems 12 (2008)
16. Ohsawa, Y., McBurney, P. (eds.): Chance Discovery (Advanced Information Processing). Springer, Heidelberg (2003)
17. Ohsawa, Y.: Chance Discovery with Emergence of Future Scenarios. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3213, pp. 11–12. Springer, Heidelberg (2004)
18. Ohsawa, Y.: Chance Discovery, Data-based Decision for Systems Design. ISDA (2006)
19. Ohsawa, Y., Ishii, M.: Gap between advertisers and designers: Results of visualizing messages. International J. of Knowledge-Based and Intelligent Engineering Systems 12 (2008)
20. Rybakov, V.V.: A Criterion for Admissibility of Rules in the Modal System $S4$ and the Intuitionistic Logic. Algebra and Logic 23(5), 369–384 (1984), (Engl. Translation)
21. Rybakov, V.V.: Problems of Substitution and Admissibility in the Modal System Grz and in Intuitionistic Propositional Calculus. Ann. Pure Appl. Logic 50(1), 71–106 (1990)
22. Rybakov, V.V.: Rules of Inference with Parameters for Intuitionistic logic. J. of Symbolic Logic 57(3), 912–923 (1992)
23. Rybakov, V.V.: Admissible Logical Inference Rules. Series: Studies in Logic and the Foundations of Mathematics, vol. 136. Elsevier Sci. Publ., North-Holland (1997)
24. Rybakov, V.V.: Logical Consecutions in Discrete Linear Temporal Logic. J. of Symbolic Logic 70(4), 1137–1149 (2005)
25. Rybakov, V.V.: Linear Temporal Logic with Until and Before on Integer Numbers, Deciding Algorithms. In: Grigoriev, D., Harrison, J., Hirsch, E.A. (eds.) CSR 2006. LNCS, vol. 3967, pp. 322–333. Springer, Heidelberg (2006)
26. Rybakov, V.: Until-Since Temporal Logic Baed on Parallel Time with Common Past. Deciding Algorithms. In: Artemov, S., Nerode, A. (eds.) LFCS 2007. LNCS, vol. 4514, pp. 486–497. Springer, Heidelberg (2007)
27. Rybakov, V.: Logics with Universal Modality and Admissible Consecutions. Journal of Applied Non-Classical Logics 17(3), 381–394 (2007)
28. Babenyshev, S., Rybakov, V.V.: Describing Evolutions of Multi-Agent Systems. In: Velásquez, J.D., Ríos, S.A., Howlett, R.J., Jain, L.C. (eds.) KES 2009. LNCS, vol. 5711, pp. 38–45. Springer, Heidelberg (2009)
29. Rybakov, V.V.: Linear Temporal Logic LTK_K extended by Multi-Agent Logic K_n with Interacting Agents. J. Log. Comput. 19(6), 989–1017 (2009)
30. Rybakov, V.V.: Algorithm for Decision Procedure in Temporal Logic Treating Uncertainty, Plausibility, Knowledge and Interacting Agents. IJIIT 6(1), 31–45 (2010)
31. Rybakov, V.V.: Interpretation of Chance Discovery in Temporal Logic, Admissible Inference Rules. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010. LNCS, vol. 6278, pp. 323–330. Springer, Heidelberg (2010)
32. Robinson, J.A.: A machine oriented logic based on the resolution principle. J. of the ACM 12(1), 23–41 (1965)

From Epistemic Luck to Chance-Seeking
The Role of Cognitive Niche Construction

Lorenzo Magnani and Emanuele Bardone

Department of Philosophy & Computational Philosophy Laboratory,
University of Pavia, Italy
{lmagnani,bardone}@unipv.it

Abstract. In this paper we will focus on how humans try to get lucky by smart eco-cognitive manipulations and how that has an impact on our evolution. We will present a brief account of epistemic luck, which will rely on abduction. Our main claim is that epistemic luck is cognitively relevant insofar as it contributes to affording human beings to generate or select the correct hypothesis solving a problem. We will then present the main idea of the paper: by definition luck can be neither predicted nor planned. However, it can actively be *sought* and *domesticated* by seeking those chances maximizing *abducibility*, which will be described as the opportunity of being afforded by lucky events – events that are out of one's control.

1 Introduction

A growing literature has questioned the idea that luck should be kept outside of rational investigation. Among the many contributions appeared in the last decade, Chance discovery (see for instance [14,16,1,12,2]) and *epistemic* luck [17] are by far the most interesting contributions on the matter. In this paper we will focus on the relationship between epistemic luck and chance-seeking. More precisely, we will focus on how humans try to get lucky by smart eco-cognitive manipulations and how that has an impact on our biological evolution. Our paper will proceed as follows: in the first part we will illustrate our interpretation of epistemic luck. We will present a brief account, which will rely on abduction. Our main claim is that epistemic luck is cognitively relevant insofar as it contributes to affording human beings to generate or select the correct hypothesis solving a problem one is facing with. We will then present the main idea of the paper: by definition luck can be neither predicted nor planned, but it can actively be *sought* and *domesticated* by seeking those chances maximizing *abducibility*, which will be described as the opportunity of being afforded by lucky events – events that are out of one's control.

We will root this ability of seeking chances in evolution, more precisely, in the ability – not entirely unique to our species – of creating powerful cognitive niches, in which the environment serves the purpose of maximizing abducibility. Indeed, evolution – in its Darwinian sense – does not display any purposefulness, as it is characterized by what we will call *evolutionary luck*. However the evolution

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 486–494, 2011.

of our species as powerful eco-cognitive engineers constituted a turning point. Even though evolution still remains *blind*, through the construction of powerful cognitive niches, our species has introduced a second inheritance system – an eco-cognitive one – in which potentially benefiting chances for taming or at least lessening the negative impact of evolutionary luck have been uncovered.

2 From Epistemic Luck to Chance-Seeking

An interesting example – though fictional – illustrates the role that epistemic luck can play in human affairs. In one of the episodes of the American television medical drama House MD, the main character – the diagnostician Dr. Gregory House – is dealing with the case of a 70-year-old scientist who collapsed in his laboratory while doing some experiment on rats. After the usual trial and error process, Dr. House successfully solves the case arriving at the correct diagnosis – amyloidosis, which consists in the abnormal deposition of amyloid, a particular protein, in various tissues of the body. What is interesting here is the way Dr. House came up with the diagnosis. The story develops as follows. Dr. House is hanging around in the hospital when he stumbles upon an underaged girl he previously met. Her father was cured some days before and now she comes back to the hospital where Dr. House works claiming that his dad lost his medicine and she has to refill. After a short conversation, the girl leaves turning her back to Dr. House who eventually checks out her Congo red thong-covered ass. Some second later Dr. House gets his insight: it might be amyloidosis. How did he come up with the explanation? Amyloidosis can be confirmed by performing a test called Congo Red Dye Test. Congo red is used to stain microscopic preparates. That is, it is added to a sample of patient's tissue and then put under the microscope. Under polarized light it indicates the presence of amyloid fibrils, as the amyloid tissue turns a dark red. Let us discuss the example with relation to epistemic luck.

First of all, the event – Dr. House checking out the girl's dark red thong – was *out of his control*. It was not him asking the girl to put her thong on display. However, the event was not only out of his control. It was also *accidental*. For instance, sunset is beyond our power. We cannot control it. But it is not accidental. Every day we can expect that the sun will go down. Whereas the case of the girl's thong was not predictable anyhow. It has just happened unexpectedly. It is accidental, but also *unique* or *singular*. And this is the third element. Dr. House could have been exposed to the same event some other time. But that would not have necessarily had the same impact. The lucky event also has a *positive* impact on the abductive process Dr. House and his team were involved in. This is the fourth element. More generally, epistemic luck is *biasing* in the sense that it affords solving a problem. From an abductive point of view, we may argue that epistemic luck makes visible to the abducer the clue (or the set of clues) enabling him or her to infer the correct hypothesis. In our example, the girl's dark red thong made visible one of the crucial symptoms for amyloidosis. This is what can be called *abducibility* (cf. [10]). The last aspect related to epistemic

luck and worth mentioning here is irrelevance. Medically speaking, the Congo red thong is completely *irrelevant* with relation to amyloidosis. It might help explain Dr. House's expression of enjoyment, but not the reason why the patient collapsed in his laboratory, and all other symptoms he had come up with. It is relevant insofar as it affords the abducer to formulate the correct hypothesis as we have just seen.

In sum, epistemic luck is an event that is described by five main elements: 1) *uncontrollability*: the event is just out of one's control; 2) *accidentality*: the event is accidental; 3) *singularity*: the event enters the scene as a single and unique one; 4) *abducibility*: the event is not neutral, but it enables a person to abduce the correct solution for the problem he or she is facing; 5) *irrelevance*: the event is apparently irrelevant, meaning that it would not bear any connection with the solution to abduce.

As already mentioned, luck cannot be planned anyhow. Indeed, there are certain situations that may be described referring to *pure* luck. To some extent this is the case approximated by what we will call *evolutionary* luck. Even thought it represents more an ideal condition than an actual fact, pure luck *just* affects the cognitive agent. We will come back to this in the next section.

Pure luck apart, in all other situations the cognitive agent – the human one in our case – is not entirely passive. Humans are afforded in different ways by external circumstances. This point is captured by the fourth of the five conditions listed above, abducibility. We have posited that epistemic luck is not neutral in the sense that it prompts a certain reaction. Interestingly, in our example the dark red thong helps Dr. House select the correct disease among all those plausible, because he could be afforded by the event. That is, his medical knowledge made possible for him to grab the chance delivered by luck. Clearly, a person who did not know about the Congo red would not have been afforded by the thong. Therefore, epistemic luck could have not brought about any substantial effect – whether negative or positive.

More than a century ago Louis Pasteur observed that luck favors the *prepared* mind. It seems that knowledge plays a crucial role for enhancing our ability as chance seekers. Even though epistemic luck cannot be controlled or planned anyhow, we posit that it can be domesticated so as to maximize abducibility [10]. That is, the opportunity that an accidental, singular, out of our control, and apparently irrelevant event may afford/suggest us to select and/or generate the correct hypothesis. Let us make a second example.

In 1990 in Vietnam malnutrition was affecting the majority of children. Dr. Jerry Sternin was sent there on behalf of a NGO called "Save the Children" to try to figure out what to do in order to mitigate such plague.[1] Indeed, he had no chance to tackle down the problem, which was dependent on other broader issues like poverty, ignorance, and poor sanitation. His budget could allow him to do nothing, but one thing: he recruited local mothers to weigh the children in the villages. By doing that he could find out those children who were not underweight

[1] The case is reported in [6].

and consequently analyze their family background. What he discovered after that was quite interesting. He identified a group of mothers – not belonging to any of the rich and influential families – who used to give their children a bowl of plain rice like any other mother, but adding shrimp, crabs, and sweet-patato greens. In doing so, their children actually got a daily portion of proteins and so they could stay healthy.

This case is quite different from the previous one. The case could be solved after Sternin identified the group of mothers adding shrimp and crabs to the bowl of plain rice. It was not just plain luck. Actually, he acted on the environment by performing a *manipulative abduction* [9,10]. A manipulative abduction is the process in which a hypothesis is generated and/or selected by resorting to a extra-theoretical and extra-sentential behavior unfolded by the skillful manipulation of external resources. In our case, weighing the children can be considered a manipulative abduction, which allowed Sternin to generate the hypothesis about the local wisdom of that group of mothers, which represented what the congo red did for Dr. House. That is, the missing clue enabling him to formulate the correct hypothesis: underweight children's mothers do not add shrimps and crabs to plain rice when they could, if appropriately instructed. Interestingly, Sternin's case is still a case of epistemic luck, because he could not predict that weighing children would allow him to spot the group of wise mothers, and so solve the puzzle. But what he did was to perform a smart manipulation of the situation in order to maximize abducibility. We may now derive some interesting implications. First of all, luck can be sought by seeking all those situations in which we are afforded to generate the correct hypothesis by luck. Secondly, the ability of seeking chances rests on our knowledge. In fact, chances are those situations in which abducibility is maximized. But that cannot be happened, if one lacks knowledge in terms of abductive skills required to be afforded [12]. Thirdly the kind of knowledge we are talking about is eco-cognitive in its essence, that is, it is related to our ability of specifically manipulating the environment.

More generally, we claim that maximizing abducibility is carried out at the *eco-cognitive level*. That is, humans maximize their chances to be lucky by constructing cognitive niches so as to be better afforded as problem-solvers. In the following sections we will illustrate how our account about epistemic luck and chance-seeking might be fruitfully applied to evolution in order to shed light on a quite controversial and hotly debated topic, namely, the role of *purposefulness* in evolution. Neo-darwinism states that there cannot be any role for purposefulness in evolution. Living organisms as part of a species evolve without following a pre-determined path, which makes sense of the various adaptations. We might say that what drives evolution is pure or, better, *evolutionary* luck. That is, organisms happen to develop adaptive solutions for their survival and reproduction simply by (evolutionary) luck. In the following we will show how purposefulness may emerge in evolution in terms of chance-seeking, which will be defined as that activity in which humans create and maintain cognitive niches to be better afforded by epistemic luck.

3 The Notion of Evolutionary Luck

According to the traditional view [13], there are four major features characterizing evolution: *multiplication*, *variation*, *heredity*, and *competition*. Multiplication refers to the fact that an entity can reproduce and in doing so it can give two, three or more others. Variation accounts for the fact that not all entities are identical. Heredity means that different entities will produce different entities. So, for instance, entities of type A will produce entities of type A, whereas entities of type B will produce entities of type B. The last feature is competition. Competition refers to the fact that a given variation has different consequences in terms of survival and multiplication for the entities that inherited it.

One of the most important issues concerning natural selection deals with variation. In an ideal world the entity with the greatest ability of surviving and reproducing will sooner or later outnumber all others. So, it will be the only existing entity. That is not what actually happens in the real world. In fact, from generation to generation some variation may occur so as to produce a complex functional system. According to the traditionally accepted view of Darwin's theory, variations are random in origin [13]. Although some biologists have recently started to challenge some of the main assumptions behind the idea that variations are random [5], it is worth nothing here the role played by evolutionary luck is somehow analogous to the one played by epistemic luck. Let us see how.

As we have already mentioned, an ideal world, in which variations do not occur, would end up after a number of generations with one entity outnumbering all the others. However, that might put life at a great risk. What if after some generation the environmental conditions changed so as to make the only species left in our ideal world unfit to survive? Indeed, life would end. As already mentioned, this is not what actually happens in our real world, and evolutionary luck plays an important role here. Since environmental conditions, namely, selection pressures, do change from time to time, in order life to persist there should be a mechanism, which favors mutation and thus variation, *when it is needed*. We maintain that evolutionary luck is what does this job.

Arguably, when the rate of environmental change is quite low, mutation is not particularly benefiting. In fact, entities have a disposal what Lablonka and Lamb [5] called DNA-care-taker genes. Basically, such genes – present in all organisms – observe and direct the execution of DNA reproduction. When some errors occur in the copying process, they are promptly fixed. However, when selection pressures dramatically change, these DNA care-taker genes are turned off and evolutionary luck takes over. Indeed, as the rate of mutation increases, so does the possibility for an entity to bear a maladaptive mutation and thus being selected out. In a way, to use an analogy introduced by Lablonka and Lamb [5], it is like for poor people to buy a lottery ticket with the last coin left in their pocket. They might get billionaire or, going to the other extreme, they might not win and so starve and then die. Indeed, this last case is extreme, but it makes more visible how mutations and variations are not designed or planned. Meaning that the survival of a given entity does not respond to any particular design – intelligent or

not it does not matter. It is just the product of evolutionary luck. More generally, evolutionary luck is responsible for the emergence of a particular adaptation, which actually makes the bearer survive and reproduce at a higher rate. As already stressed above, adaptations are not chosen by the evolving organism, but they just appear.

In sum, evolution is not anyhow planned or designed to achieve a particular outcome. Foresights are not possible to make. In the last part of the paper we will show how humans try to domesticate it by means of cognitive niche construction. That is, the construction of more and more sophisticated cognitive niches, which may enhance humans' chance of being afforded by luck. The illustration of cognitive niche construction will also shed light on the issue of purposefulness in evolution.

4 Cognitive Niche Construction and Chance-Seeking

It is well-known that one of the main forces that shape the process of adaptation is natural selection. That is, the evolution of organisms can be viewed as the result of a selective pressure that renders them well-suited to their environments. Adaptation is therefore considered as a sort of *top-down process* that goes from the environment to the living creature [3]. In contrast to that, a small fraction of evolutionary biologists have tried to provide an alternative theoretical framework by emphasizing the role of niche construction [8,7,15].[2]

According to this view, all organisms try to modify their surroundings in order to better exploit those elements that suit them and eliminate or mitigate the effect of the negative ones. So, in any ecological niche, the selective pressure of the *local* environment is drastically modified by organisms in order to lessen the negative impacts of all those elements which they are not suited to. This new perspective constitutes a radical departure from traditional theory of evolution introducing a second inheritance system called *eco-cognitive inheritance system* [10]. According to this view, acquired characters – discarded for such a long time – can enter evolutionary theories as far as they cause a modification to the environment that can persist and thus can modify the local selective pressure. Eco-cognitive inheritance system is different from the genetic one in the following way [15]: 1) genetic materials can be inherited only from parents or relatives. Conversely, modifications on the environment can affect everyone, no matter who he/she is; 2) genes transmission is a one way transmission flow, from parents to offspring, whereas environmental information can travel backward affecting several generations; 3) genetic inheritance can happen once during one's life, at the time of reproductive phase. In contrast, ecological information can be transferred during the entire duration of life. Indeed, it depends on the eco-engineering capacities at play; 4) genetic inheritance system leans on the presence of *replicators*, whereas the ecological inheritance system leans on the *persistence* of whatsoever changes made upon the environment.

[2] This view is also supported by J. Scott Turner who claims that organisms take an active part in evolution, as they are agents of *homeostasis* [19].

Indeed, a large part of the niche construction process is intrinsic to the Neo-Darwinian framework. The information that basically drives niche construction is of course at the level of semantic information encoded in DNA and provided by the evolutionary process as the result of natural selection. However, niche construction is also *active* and not reactive, *profitable* and not goalless, like natural selection and, moreover, it is always an informed selective process.

So, niche construction introduces "a sense of purposefulness" in evolution. What is worth noting here is that the modifications made on the local environment are not random. And, more importantly, insofar as those modifications persist, they enter the scene as *potential chances* so that evolutionary luck is not the only factor affecting evolution.[3] In this sense, as argued by Turner, "evolution becomes less a province of one class of arbiters of future function – genes – and more the result of a nuanced interplay between the multifarious specifiers of future function" [18, pp. 348–349].

Now, that a sense of purposefulness emerges in evolution means that in a way human beings are in a better position for having an impact on their fitness and on what evolutionary path they may take. This is implicit in the definition of niche construction. Human beings are not merely passive, but they actively sort out chances for reproduction and survival by constructing *cognitive* niches. Crucial to this is the high level of *plasticity* exhibited by humans. In a nutshell, developing plastic response is a fundamental condition for the emergence of chance-seeking activities. In fact, plasticity helps humans exploit latent chances and enhance abducibility. Let us see how.

Plasticity of response to an ever-changing environment is connected to the necessity of having other means for acquiring information, more readily and quickly of the genetic one [4]. We posit that (cognitive) niche construction plays a fundamental role to meet this requirement. Plasticity depends on niche construction as far as various organisms may alter local selective pressure via niche construction, and thus increase their chances for surviving. More specifically, cognitive niches are crucial in developing more and more sophisticated forms of plasticity – and thus chance-seeking activities – because they constitute an additional source of information favoring abducibility.

Let us now go back to the example of the group of Vietnamese mothers mentioned in the first section. In the light of what we have just argued, we may say that Sternin maximized his chances to be lucky, as he could rely on a piece of equipment available in the local cognitive niche, namely, a scale. In fact, a scale is not just over there in the natural world. But it is something made, designed, and used by humans for a specific purpose. It provides a set of data and information that, first of all, one would not be able to gather otherwise. Secondly, interacting with one's medical knowledge, it may contribute to uncover specific symptoms. The scale – as part of Sternin's cognitive niche – gave him access to clues, which exposed him to favorable chances to formulate the correct abduction. As a matter of fact, without the scale that abduction would be simply out

[3] For a discussion of the moral implications of cognitive niche construction, see Magnani [11].

of his reach. Thirdly, both the tool design and the knowledge required to use it belong to the eco-cognitive inheritance system, as they are the result of many generations, which managed to develop, accumulate, maintain, and passed them on to the next ones.

5 Conclusions

In this paper we have tried to analyze the role of epistemic luck in problem-solving situations. First of all, we have outlined the main characteristics of epistemic luck. We have argued that epistemic luck can be predicted, since it is out of one's control. However, it cannot be domesticated somehow by maximizing abducibility. That is, one may try to get lucky by manipulating the environment so as to be exposed by profitable chances. We have maintained that such eco-cognitive manipulations are crucial also from an evolutionary point of view. Although there is not any sense of purposefulness in evolution, however, through the construction of more and more sophisticated cognitive niches, humans have tried to tame what we called the evolutionary luck. That is, they have tried to better control some evolutionary forces affecting their survival as a species.

References

1. Abe, A.: Cognitive chance discovery. In: Stephanidis, C. (ed.) UAHCI 2009. LNCS, vol. 5614, pp. 315–323. Springer, Heidelberg (2009)
2. Bardone, E.: Seeking Chances. From Biased Rationality to Distributed Cognition. Springer, Heidelberg (2011)
3. Godfrey-Smith, P.: Complexity and the Function of Mind in Nature. Cambridge University Press, Cambridge (1998)
4. Godrey-Smith, P.: Environmental complexity and the evolution of cognition. In: Sternberg, R., Kaufman, K. (eds.) he Evolution of Intelligence, pp. 233–249. Lawrence Erlbaum Associates, Mawhah (2002)
5. Jablonka, E., Lamb, M.J.: Evolution in Four Dimensions. Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life. The MIT Press, Cambridge (2005)
6. Kawasaki, G.: Enchantment: The Art of Changing Hearts, Minds, and Actions. Penguin&Portfolio, New York (2011)
7. Laland, K.N., Odling-Smee, F.J., Feldman, M.W.: Cultural niche construction and human evolution. Journal of Evolutionary Biology 14, 22–33 (2001)
8. Laland, K., Odling-Smee, J., M.W, F.: Niche construction, biological evolution and cultural change. Behavioral and Brain Sciences 23(1), 131–175 (2000)
9. Magnani, L.: Abduction, Reason, and Science. Processes of Discovery and Explanation. Kluwer Academic/Plenum Publishers, New York (2001)
10. Magnani, L.: Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning. Springer, Heidelberg (2009)
11. Magnani, L.: Understanding Violence. The Intertwining of Morality, Religion and Violence: A Philosophical Stance. Springer, Heidelberg (2011)

12. Magnani, L., Bardone, E.: Sharing representations and creating chances through cognitive niche construction. The role of affordances and abduction. In: Iwata, S., Ohsawa, Y., Tsumoto, S., Zhong, N., Shi, Y., Magnani, L. (eds.) Communications and Discoveries from Multidisciplinary Data, pp. 3–40. Springer, Berlin (2008)
13. Maynard-Smith, J., Szathmry, E.: The Origins of Life: From the Birth of Life to the Origin of Language. Oxford University Press, Oxford (2000)
14. McBurney, P., Parsons, S.: Chance discovery using dialectical argumentation. In: Terano, T., Nishida, T., Namatame, A., Tsumoto, S., Ohsawa, Y., Washio, T. (eds.) New Frontiers in Artificial Intelligence: Joint JSAI 2001 Workshop Post Proceedings, pp. 414–424. Springer, Berlin (2001)
15. Odling-Smee, F., Laland, K., Feldman, M.: Niche Construction. A Neglected Process in Evolution. Princeton University Press, New York (2003)
16. Ohsawa, Y., McBurney, P. (eds.): Chance Discovery. Springer, Berlin (2003)
17. Pritchard, D.: Epistemic luck. Journal of Philosophical Research 29, 193–222 (2004)
18. Turner, J.S.: Extended phenotypes and extended organisms. Biology and Philosophy 19, 327–352 (2004)
19. Turner, J.S.: The Tinkerer's Accomplice: How Design Emerges from Life Itself. Harvard University Press, Cambridge (2007)

Relation between Chance Discovery and Black Swan Awareness

Akinori Abe

NTT Communication Science Laboratories
3-1 Wakamiya, Morinosato, Atsugi-shi, Kanagawa 243-0198, Japan
abe.akinori@lab.ntt.co.jp, ave@ultimaVI.arc.net.my

Abstract. We have been researching for chance discovery more than 10 years. The aim of chance discovery is discover something, but chance discovery is rather different from data mining. In 2007 Taleb introduced a concept "Black Swan." Taleb uses this rare black swan metaphor to explain how usual person tend to ignore rare or novel events and the importance of being aware of such rare or novel events. The main concept of Black Swan seems coincide with that of chance discovery. Thus, in this paper, I compare a chance with Black Swan and present the future feather of chance discovery.

Keywords: chance discovery, Black Swan, rare or novel event.

1 Introduction

We have been researching for chance discovery more than 10 years. Of course, the aim of chance discovery is discover something, but chance discovery is rather different from data mining. In general, data mining tries to discover trends or general matters which will be useful for us. For instance, a trend in stock price is usually data mined. Data mining usually adopts stochastic models. Since abnormal values are usually ignored, stochastic models such as the Black-Scholes model could not explain the end of the bubble phenomenon [14]. This situation is the same as in data mining. In fact, recently data mining has changed its scope to discover exception or outlier. Actually, those who can predict this type of extraordinary situation can make money. Thus the change of data mining's research interest is very natural. On the other hand, according to the definition of chance discovery [11], chance discovery's main aim has been to discover rare, hidden, potential or novel event(s) / situation(s) that can be conceived either as a future opportunity or risk. In 2007 Taleb introduced a concept "Black Swan" [15]. Black swans had never seen in Europe, but are native to Australia. Therefore, Europeans believed all swans were white and the sighting of the first black swan might have been an interesting surprise for a few ornithologists. Taleb uses this rare black swan metaphor to explain how usual person tend to ignore rare or novel events and the importance of being aware of such rare or novel events. The main concept of Black Swan seems coincide with that of chance discovery.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 495–504, 2011.

In this paper, I compare a chance with Black Swan and present the future feather of chance discovery. For chance discovery I mainly use a framework prposed in [2], [3], and [4] etc.

2 Chance Discovery

In many papers I have described definition and concept of chance discovery. However, in this paper, for a comarison I brief illustrate chance discovery.

In fact, the following definition is rather differs from the original definition in [11] to reflect the recent research interests.

> A chance is rare, hidden, potential or novel event(s) / situation(s) that can be conceived either as a future opportunity or risk.

Then "chance discovery" research is a type of research to establish methods, strategies, theories, and even activities to discover a chance. In addition, it aims at discovering human factors for chance discoveries. Therefore not only researchers in computer science and engineering but also researchers with different expertise such as psychologists, philosophers, economists and sociologists take part in chance discovery research.

I formalized the procedure of chance disocvery by using abduction and analogical mapping ([2], [3], and [4] etc.). Actually, chance discovery was formalized by Abductive Analogical Reasoning (AAR)[1].

Based on the formalization, I defined two types of chance.

- **Type 1:** When some of the hypotheses are unknown
  When some of the necessary hypotheses are unknown (not found), the current observation cannot be explained. In the context of chance discovery, chance seems to be a set of unknown hypotheses, and this type of inference can be regarded as pure abduction.
  For instance, when we must predict the occurrence of a serious earthquake, we show that if such events (symptoms) as active faults can be found, a serious earthquake will occur, and then attempts to find the events must be carried out.
  Let a knowledge base (Σ) be as follows:

$$
\begin{array}{l}
earthquake \coloneq movement, distortion, oceanic\_plate. \\
faults. \\
distortion. \\
movement.
\end{array}
$$

  By the usual hypothetical reasoning, since *oceanic_plate* is not included in a hypothesis base as the possible hypothesis, the observation (O) — *earthquake*— cannot be explained. However, when we want to show a possibility of an earthquake, we must discover and show symptoms (= chance).

When we apply AAR[1] to solve this problem, AAR explains an observation *earthquake* by generating a hypothesis $(\neg S)$ *oceanic_plate*.

$$\Sigma \models \neg oceanic\_plate \vee earthquake. \tag{1}$$

Then, in the above example, chance is *oceanic_plate*, which was not found (in the knowledge base). It is logically discovered and suggested by abduction. AAR shows that "if the hypothesis *oceanic_plate* is justified, the observation will be explained." In this case, with a suggestion from abduction, the user will be aware of the occurrence of an earthquake, and can assume that *oceanic_plate* is a chance of earthquake; therefore a possible way to predict an earthquake will be to find *oceanic_plate*. Consequently, if we can find *oceanic_plate*, we can predict (explain) the occurrence of a serious earthquake.

– **Type 2:** When some of the rules are unknown
When some of the necessary rules are unknown, even if we are aware of all the symptoms, the future observation cannot be explained (predicted). In this case, reasons are usually shown afterward.
For instance, when a serious earthquake like the Great Hanshin Earthquake occurs, even if we know that there are a lot of active faults near the Kansai area, we cannot predict an earthquake because we did not know the relationship between active faults and the Hanshin Earthquake. Therefore, when we want to predict an earthquake, we must predict unknown rules from future observation (to be predicted) and the current symptoms.
Let a knowledge base (Σ) be the above knowledge base. In this case, as shown, *oceanic_plate* can be shown as a chance. However, sometimes we must be aware of existing things $(faults)$ that are similar to the known but non-existing things (*oceanic_plate*). In such cases, we predict or generate new rules as hypotheses. AAR can explain an observation by generating new hypotheses by referring to the known knowledge. If we use this type of inference, the following inference can be achieved. Since the knowledge base has $faults$ as a fact, AAR uses $faults$ as one of the symptoms (= hypotheses) to explain an observation. When the observation cannot be explained by $faults$, AAR, by referring to "$earthquake \coloneq movement, distortion, oceanic\_plate$.", generates a new rule " $earthquake \coloneq movement, distortion, faults$." to explain an observation *earthquake* when *oceanic_plate* is similar to $faults$.

$$\Sigma \models S'' \vee earthquake. \tag{2}$$

$$\neg S'' = \neg movement \vee \neg distortion \vee \neg faults \vee earthquake. \tag{3}$$

In this case, $faults$ is a chance, and by the similar experiences "$earthquake \coloneq movement, distortion, faults$.", forthcoming observation can be explained (predicted). AAR shows that "if the rule "$earthquake \coloneq movement$,

[1] For logical formalization, please see [1].

distortion, *faults*." is justified, the observation will be explained by *faults*[2]." This is a suggestion of the occurrence of an earthquake that makes the user aware of the the occurrence of the earthquake. If the user thinks the newly generated rule is plausible, he/she can decide *faults* as a chance and predict an earthquake by using the rule. Of course, if there is no similar knowledge, some rules can be generated to explain the observation only in an abductive way. However, in this case, generated rules may be shortcut rules or wrong rules. More discussions have been done in [3] and [4].

Thus hypothesis (type 1) and rule (type 2) are considered as chance to be discovered. In this framework, chances are abduced. That is, the first type is to show (suggest) unseen or unknown events as chance and the second type is to show (suggest) the known events as chance by generating new rules. Thus, by abduction chances are logically discover then potential factors (chance) that will cause serious affairs are suggested. In addition, AAR is a reasoning that combine abduction and analogical mapping. Only by abduction, as shown above, we sometimes cannot guarantee the generated hypotheses set. With analogical mapping, we can generate a more plausible hypotheses set that is supported by the existing clause set. Then abduction can perform discovery and suggestion of chance and analogical mapping can perform adjustment and confirmation of chance. By the combination of abduction and analogical mapping, more plausible chance can be suggested.

3 Black Swan

In 2007 Taleb published "Black Swan" [15]. In the book, Taleb introduced a concept "Black Swan[3]" as an event with the following three attributes.

1. It is outlier, as it lies outside of the realm of regular expectations, because nothing in the past can convincingly point to its possibility.
2. It carries an extreme impact.
3. In spite of its outlier status, human nature makes us concoct explanations for its occurrence after the fact, making it explainable and predictable.

Taleb pointed out that one single observation can invalidate a general statement derived from millennia of confirmatory sighting of millions of white swan. All you need is one single black bird. As examples, Taleb showed the rise of Hitler and the subsequent war, the rise of Islamic fundamentalism, the spread of the Internet, and the market crush of 1987 (and its more unexpected recovery). Those events will not be able to predict by the previous pile of knowledge. By the concept of "Black Swan," Taleb deals with humans' blindness with respect to

[2] In fact, a certain movement of faults is also caused by the movement of oceanic plates, and it would be rather difficult to find a similarity between faults and oceanic plates in a simple way. Anyway, this scenario is only used for explanation.

[3] Black swans are native to Australia, but had never been seen in Europe.

randomness, particularly the large deviations. That is, Taleb pointed out that we humans are arrogant. We think we know everything. We study the normal, and rule out the random and the unimaginable. Because Black Swans are rare and unpredictable, the statistical measures "experts" use to predict "risk" usually don't even consider the possibility of Black Swans.

Thus Taleb pointed out the limitation of general data mining or predictive inference based on stochastic models. In addition, Taleb points out the importance of not ignoring rare or novel events.

Then "Black Swan Theory" can be defined as "The event is a surprise (to the observer) and has a major impact. After the fact, the event is rationalized by hindsight." can be obtained. Unlike the earlier philosophical black swan problem, the "Black Swan Theory" refers only to unexpected events of large magnitude and consequence and their dominant role in history. Such events, considered extreme outliers, collectively play vastly larger roles than regular occurrences.

4 Big Earthquake in Japan, 2011

On 11 March 2011, east and north side of Japan was suffered from very big earthquake. Several article such as the Washington Post and Foreign Policy called it "Japan's Black Swan." That is, the 9.0 earthquake arose from a failure of human imagination. In fact the earthquake arose from a movement in the tectonic plates involved. However, the failure to prepare better for the affects of that earthquake and the subsequent tsunami resulted from a preceding failure to imagine that an a disaster of that size could occur. In addition, in fact, such big earthquake has not occurred more than 1000 years. Therefore, we ignore possibility of occurring such big earthquake even if some researchers showed their analyses of the big earthquake in 896 and alerted the possibility of huge miserable situations. Of course, some analysts alerted the place of an auxiliary power was not suitable for a huge tsunami which has not occurred for 1000 years. However, no problem has occurred since 1970 (the year when the service started). Thus accidents in nuclear facilities also arose from a failure of human imagination.

In Tabel's homepage [17], Taleb analyzes Japan's accidents in nuclear facilities caused by earthquake as follows:

The Japanese Nuclear Commission had the following goals set in 2003: "The mean value of acute fatality risk by radiation exposure resultant from an accident of a nuclear installation to individuals of the public, who live in the vicinity of the site boundary of the nuclear installation, should not exceed the probability of about 1×10^6 per year (that is , at least 1 per million years)." That policy was designed only 8 years ago. Their one in a million-year accident almost occurred about 8 year later (I am not even sure if it is at best a near miss). We are clearly in the Fourth Quadrant there.

Then Tabel lists issues:

1) Small probabilities tend to be incomputable; the smaller the probability, the less computable.

2) Model error causes the underestimation of small probabilities & their contribution (on balance, because of convexity effects). Any model error, just as any undertainty about flying time causes the expected arrival to be delayed (you rarely land 4 hours early, more often 4 hours late on a transatlantic flight, so "unforeseen" disturbances tend to delay you).
3) The problem is more acute in Extremistan, particularly the manmade part. The probabilities are underestimated but the consequences are much, much more underestimated.
4) [...] because of globalization, the costs of natural catastrophes are increasing in a nonlinear way.
5) Casanova problem (survivorship bias in probability): If you compute the frequency of a rare event and your survival depends on such event not taking place (such as nuclear events), then you underestimated that probability.
6) Semi-technical Examples: to illustrates the point (how models are Procrustean beds)[4]

Tabel focues on undestimation of the possibility of huge earthquake. For instance, Taleb points out that if we use binomial distribution model, there is no such thing as "measurable risk" in the tails, no matter what model we use. For underestimation of tail events, Tabel discusses in detail in [16].

Ohsawa who has been researching for chance discovery tries to discover possible symptoms of earthquake. After the big earthquake, several aftershocks occurred. Ohsawa collected data of aftershocks and adopted $Kamishibai - KeyGraph^{\textcircled{R}}$ [13] to discovered possible points of next coming earthquake (for instance, in the blog dated on 24 March, 2011 — `http://d.hatena.ne.jp/kokashita/20110324/1300977044` (in Japanese)). In the analysis, Ohsawa focuses on points which are rare or novel (earthquakes have occurred not so frequently or not) but will be influenced by previous earthquakes. In fact, every events should occur in causality. As well-known, earthquakes also occur by a pile of many factors. $Kamishibai - KeyGraph^{\textcircled{R}}$ visualizes such sequences (causality) and hidden sequences (requiring certain techniques). Users can be aware of hidden but important points by observing a graph generated by $Kamishibai - KeyGraph^{\textcircled{R}}$. In fact, Ohsawa points out that red points shown in Fig. 1 are warning points where earthquake do not regularly occur[5]. By observing the graph, users can be aware of rare or easy-to-ignored points as possibly-earthquake-occurring points.

In the previous section, I illustrated two types of chance discovery by using example of earthquake prediction. I adopt abduction which also considers causality of events to generate a set of hypothesis. That is, by observing various events, abduction abduces necessary but rare, hidden, or ignored events for earthquake. Actually, since I do not use probabilistic model for inferences, during inference

[4] Tabel showed two example of binomial distribution with B[N, p] probability of success, and a Gaussian with the probability of exceeding a certain number. Details are discussed in [17].

[5] Letters in the figure is written in Japanese. They are names of region. Point 88 will be a new point which $Kamishibai - KeyGraph^{\textcircled{R}}$ generated.

Fig. 1. Analysis of earthquake occurring by $Kamishibai - KeyGraph^{\textcircled{R}}$

rareness or novelty are not explicitly expressed. However during abduction, several hypotheses are newly generated. Users will evaluate rareness or novelty and importance of the hypotheses.

5 Chance and Black Swan

By Black Swan, Taleb focuses on underestimation of novel or unknown events. Actually, referring to [16], Taleb's analysis is based on stochastic models, but for certain problems Taleb focuses on awareness of such novel or unknown events. In fact, Black Swan actually exists, but since we do not know it is a black coloured swan, we tend to ignore it. In the worst case, we reject such event as an illusion or an exception. Then, we tend to ignore or not consider such significant events for predicting and preparing such accidents as in nuclear facilities in Fukushima. Of course, it is sometimes ridiculous and inefficient to prepare for all rare occurring events. However, an important point is that we should select events which might cause significant or serious situation in the future. This sense was proposed in the chance discovery's definition. Accordingly, we are not necessary to focus on all rare or novel events, but we can especially focus on special events which might cause significant or serious situation such as the end of economic bubble in Japan in the future. For that several approaches have been proposed. I cannot intorduce all the proposals. Please refer to [5], [11] and [12] etc. Important strategies in chance discovery are:

1. to determine missing or hidden events
2. to display such missing or hidden events

3. to provide a certain scheme for the user to consider such missing or hidden events

In my formalization, abduction determines missing or hidden events. Then if we use a certain visualization system such as $Kamishibai - KeyGraph^{®}$, it will be possible to display such missing or hidden events to users. For instance, in [7], I proposed an accident or incident determination system by abduction. I also proposed communication error determination system by adopting $Kamishibai - KeyGraph^{®}$ [8]. Main objective of the both system is to determine rare or novel event which might lead us to serious accident or incident. In fact, both systems suggest possible point where reasons of accident or incident occurred. Final decision is performed by users. Sometimes it is difficult to make a decision in such situations. Users will be at a loss which hypothesis should be take for the future. Therefore, it is necessary to provide a proper scheme to deal with such difficult or complex situation. For such situation, I introduced a concept "curation" to chance discovery [9]. For curation I defined the following key schemes:

- Curation is a task to offer users opportunities to discover chances.
- Curation should be conducted with considering implicit and potential possibilities.
- Chances should not be explicitly displayed to users.
- However, such chances should be rather easily discovered and arranged according to the user's interests and situations.
- There should be a certain freedom for user to arrange chances.

Actually, curation has just began and is an new and on-going research. The above definition can be modified and better scheme will be proposed for chance discovery. However, its fundamental philosophy (for instance, the fiest one) will not change.

For Black Swan, actually it is not a symptom but an event which just occurs. In this point, chance and Black Swan are different. For instance, in the case of the tragedy caused by the earthquake, Black Swan is an accident in nuclear facilities. On the other hand, chance will be the place of an auxiliary power.

However, the procedure focusing on events in long tail will be coincide with procedures in chance discovery. And it is important to focus on what is sighting but ignoring. Chance discovery, in addition, focuses on what is currently not sighting or difficult to be sighting. To be aware of Black Swan is important but it will be a rather serious problem to discover such Black Swan. According to Taleb's assumption, Black Swan does actually exist, but is easy to ignored. It will be necessary to provide a certain mechanism such as an affordance collection. In [6], based on abduction, I formalized a concept of affordance based support system for dementia persons. However, I did not consider a better affordance collection strategy. As I proposed in [10], mechanism such as an extended KeyGraph will be necessary.

Black Swan and chance discovery can be combined. When an event which seems Black Swan can be predicted or estimated, by a logical procedure, it will be possible to determine the event should cause benefit or serious risk. When it

will cause benefit or serious risk, by chance discovery procedure, it will be able to discover rare or novel events which will cause Black Swan. Then we will be able to prepare to benefit or serious risk.

6 Conclusions

In this paper, I compared chance with Black Swan. Chance discovery mainly deal with discovering rare or novel events that might cause benefit or risk in the future. Thus various tools and scheme such as *KeyGraph*® and abduction have been applied or developed. On the other hand, Taleb uses this rare black swan metaphor to explain how usual person tend to ignore rare or novel events and the importance of being aware of such rare or novel events which carry extreme impacts. In fact, Black Swan does exist, but is not recognized as a bird. Then sometimes unhappy or miserable situation occurs. Therefore, such Black Swan should be aware of beforehand. I explain this situation by using affordance collection. I think it will be difficult to discover Black Swan, if it does not stand in front of us. The problem shown by Taleb is how to recognize something standing in front of us as Black Swan. Of course, for chance discovey, this type of problem should also be solved. For affrodance collection in [10] I proposed an extended KeyGraph. In addition, recently I have introduced curation to chance discovery [9]. In order to discover more proper chance or Black Swan, the above strategy will be necessary.

References

1. Abe, A.: Abductive Analogical Reasoning. Systems and Computers in Japan 31(1), 11–19 (2000)
2. Abe, A.: The Role of Abduction in Chance Discovery. In: Proc. of SCI 2001, vol. VIII, pp. 400–405 (2001)
3. Abe, A.: The Role of Abduction in Chance Discovery. New Generation Computing 21(1), 61–71 (2003)
4. Abe, A.: Abduction and Analogy in Chance Discovery. In: [11], ch. 16, pp. 231–248. Springer, Heidelberg (2003)
5. Abe, A., Ohsawa, Y. (eds.): Readings in Chance Discovery. International Series on Natural and Artificial Intelligence, Advanced Knowledge Intelligence, vol. 3 (2005)
6. Abe, A.: Cognitive Chance Discovery. In: Stephanidis, C. (ed.) UAHCI 2009, Part I. LNCS, vol. 5614, pp. 315–323. Springer, Heidelberg (2009)
7. Abe, A., Ohsawa, Y., Kuwahara, N., Ozaku, I.H., Sagara, K., Kogure, K.: Scenario Violation as Gaps between Activity Patterns. New Mathematics and Natural Computation 6(2), 193–208 (2010)
8. Abe, A., Ohsawa, Y., Ozaku, I.H., Sagara, K., Kuwahara, N., Kogure, K.: Communication Error Determination System for Multi-layered or Chained Situations. Fundamenta Informaticae 98, 123–142 (2010)
9. Abe, A.: Curation in Chance Discovery. In: Proc. ICDM 2010, 5th International Workshop on Chance Discovery, pp. 793–799 (2010)

10. Abe A.: Curation and Communication in Chance Discovery. In: Proc. of 6th International Workshop on Chance Discovery (IWCD6) in IJCAI 2011 (to appear, 2011)
11. Osawa, Y., McBurney, P. (eds.): Chance Discovery. Springer, Heidelberg (2003)
12. Ohsawa, Y., Tsumoto, S. (eds.): Chance Discoveries in Real World Decision Making. Data-based Interaction of Human Intelligence and Artificial Intelligence Series: Studies in Computational Intelligence. Springer, Heidelberg (2006)
13. Ohsawa, Y., Ito, T., Kamata, M.: Kamishibai KeyGraph as Scenario Map Visualizer for Detecting Transient Causes from Sequential Data. In: Proc. of PAKDD 2008 Working Notes of Workshops on Data Mining for Decision Making and Risk Management, pp. 272–283 (2008)
14. Takeuchi, K.: Where probabilistic and stochastic society goes? revue de la pensée d'aujourd'hui 28(1), 84–99 (2000) (in Japanese)
15. Taleb, N.N.: The Black Swan, Allen Lane (2007)
16. Taleb, N.N.: Estimation and Forecasting Errors, Regress Arguments, and the Underestimation of Tail Events (2011), http://www.fooledbyrandomness.com/errors.pdf
17. Taleb, N.N.: Opacity: What We Do Not See, A Philosophical Notebook (2011), http://www.fooledbyrandomness.com/notebook.htm

On Kernel Information Propagation for Tag Clustering in Social Annotation Systems

Guandong Xu[1,3], Yu Zong[2,4,⋆], Rong Pan[3], Peter Dolog[3], and Ping Jin[2]

[1] School of Engineering & Science, Victoria University, Vic 8001, Australia
[2] Department of Information and Engineering, West Anhui University, Liuan, China
[3] Department of Computer Science, Aalborg University, Dk-9220, Denmark
[4] Department of Computer Science and Technology, University of Science and Technology of China, 230036, China

Abstract. In social annotation systems, users label digital resources by using tags which are freely chosen textual descriptors. Tags are used to index, annotate and retrieve resource as an additional metadata of resource. Poor retrieval performance remains a major challenge of most social annotation systems resulting from the severe problems of ambiguity, redundancy and less semantic nature of tags. Clustering method is a useful approach to handle these problems in the social annotation systems. In this paper, we propose a novel clustering algorithm named kernel information propagation for tag clustering. This approach makes use of the kernel density estimation of the KNN neighbor directed graph as a start to reveal the prestige rank of tags in tagging data. The random walk with restart algorithm is then employed to determine the center points of tag clusters. The main strength of the proposed approach is the capability of partitioning tags from the perspective of tag prestige rank rather than the intuitive similarity calculation itself. Experimental studies on three real world datasets demonstrate the effectiveness and superiority of the proposed method.

1 Introduction

In past years the emergence of Web 2.0 applications has created a new era for sharing and organizing documents in online social communities. The shared documents could range diversely from the social bookmarks *Del.icio.us*[1] to scientific publications on *CiteUlike*[2]. One of the common characteristic these kinds of document possessing is the phenomenon of *Folksonomy* - users choose their own free style terms (i.e. tags) to annotate various documents indicating their own perceptions or conceptual judgments on these resources for better indexing and annotation. In other words, *Tag*, as one kind of specific lexical information that is user-generated metadata with uncontrolled vocabulary, plays a crucial role in such social collaborative systems.

⋆ Corresponding Author.
[1] www.delicious.com
[2] www.citeulike.org

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 505–514, 2011.

Recently tagging has been widely used in recommender systems for many applications [3,7,14]. The common usage of tags in these systems is to add the tagging attribute as an additional feature to re-model users or resources over the tag vector space, and in turn, making tag-based recommendation or personalized recommendation. However, as the tags are of syntactic nature, in a free style and do not reflect sufficient semantics, the problems of redundancy, ambiguity and less semantics of tags are often incurred in all kinds of social tagging recommender systems. In order to deal with these difficulties, recently clustering method has been introduced into social tagging recommender systems to find meaningful topic information conveyed by tag aggregates. The aim of tag clustering is to reveal the coherence of tags from the perspective of how resources are annotated and how users annotate in collaborative annotations.

In the context of tag clustering, most of the researches on tagging clustering are directly using the traditional clustering algorithms such as K-means [11] or Hierarchical Agglomerative Clustering [12] on tag data, which possess the inherent drawbacks, such as the sensitivity of initial values and high computational cost etc. On the other hand, various tags used in the tagging data apparently have different importances in tag groups due to the semantic or domain topic tendency of tags. Bearing this intrinsic phenomena in mind, we propose to make use of the individual significance of each tag in tagging data for tag clustering. The basic idea behind our approach is that the propagated centrality (or prestige) degree of one tag derived from the tag neighborhood graph does reveal the importance that is contributed to the tag cluster forming. In particular, we devise a new Kernel Information Propagation Tag Clustering (KIPTC) algorithm to obtain the centrality (i.e. prestige) degree of each tag via a random walk over the graph approach. In order to evaluate the effectiveness of the proposed approach, we conduct experiments on real tagging datasets. The contributions of our paper are as follows:

- We address the tag clustering problem in social annotation systems via a graph-based optimization approach.
- We propose a new Kernel Information Propagation Tag Clustering algorithm, in which we define the Prestige Rank to capture the importance of tags in the KNN neighbor directed graph, and devise an iterative updating mechanism based on random walk to identify the global prestige rank.
- We conduct comparative experiments on three real world datasets to evaluate the effectiveness of the proposed algorithm.

The remainder of this paper is organized as follows. We review the related work in Section 2 and introduce the preliminaries in Section 3. The details of KIPTC algorithm are discussed in Section 4. Experimental evaluation results are reported in Section 5. Section 6 concludes this paper and outlines the future work.

2 Related Work

In past years, many studies have been carried out on tagging clustering. [12] demonstrates how tag clusters serving as coherent topics can aid in the social

recommendation of search and navigation. In [6] topic relevant partitions are created by clustering resources rather than tags. By clustering resources, it improves recommendations by distinguishing between alternative meanings of query. While in [1], clusters of resources are shown to improve recommendation by categorizing the resources into topic domains. A framework named Semantic Tag Clustering Search, which is able to cope with the syntactic and semantic tag variations is proposed in [2]. P. Lehwark et al. use Emergent-Self-Organizing-Maps (ESOM) and U-Map techniques to visualize and cluster tagged data and discover emergent structures in collections of music [8].

Kernel density estimate method is a widely used statistical approach for non-parameter density estimation in high dimensional space. For example, [9] has used it in capturing the local characteristic and density distribution in sparse high dimensional space. In this paper, our approach is originated from the concept of kernel information, and extends it to reveal the centrality of tag in tag aggregates, which is fundamentally different from the above clustering approaches.

3 Preliminaries

3.1 Social Tagging System Model

In this paper, our work is to deal with the tagging data. A typical social tagging system has three types of objects, users, tags and resources which are interrelated with one another. Social tagging data can be viewed as a set of triples [5,4], with each (u, r, t) representing a user u annotates tag t to resource r. Therefore a social tagging system can be described as a four-tuple, where there exists a set of users, U; a set of tags, T ; a set of resources, R; and a set of annotations, A^N. We denote the tagging data in the social tagging system as D i.e., $D =< U, R, T, A^N >$. The annotations are represented as a set of triples containing user u, tag t and resource r: $A^N \subseteq \langle u, r, t\rangle : u \in U, r \in R, t \in T$. Therefore a social tagging system can be viewed as a tripartite hyper-graph [10] with users, tags and resources represented as nodes and the annotations represented as hyper-edges connecting users, resources and tags.

3.2 A Working Example of Tag Clustering

In the above model, the tag is usually in a very high dimension due to the free style of tag texts, which results in the problem of redundancy and ambiguity; in turn, bringing in the difficulty in tag computing such as the similarity calculation of tag vectors. Therefore, clustering is often employed to capture the topical aggregates of tags, i.e. one kind of structural semantics of tags. In real applications, to fulfill this we usually decompose the tripartite graph of social tagging data to form a resource-tag matrix by accumulating the frequency of each tag in the resource vector along users. In this expression, each tag is described by a set of resources, to which this tag has been assigned, i.e., $t_i = (w_{i1}, \cdots, w_{im})$, where w_{ik} denotes the the occurrence frequency on resource r_k dimension of tag t_i. Thus, the similarity between any two tags is defined as follow:

Definition 1. *Given two tags* $t_i = (w_{i1}, \cdots, w_{im})$ *and* $t_j = (w_{j1}, \cdots, w_{jm})$*, the similarity is defined as the Cosine function of angle between two vectors of* t_i *and* t_j*:*

$$Sim\,(t_i, t_j) = \frac{t_i \cdot t_j}{\|t_i\| \cdot \|t_j\|} \tag{1}$$

Upon the mutual tag similarity is determined, various clustering algorithms could be applied to partition the tags. Fig. 1 gives a simple working example to show how five tags are assigned into two groups by using different clustering strategies (in red or black dashed circles). We will use this example to demonstrate our approach in the later section.

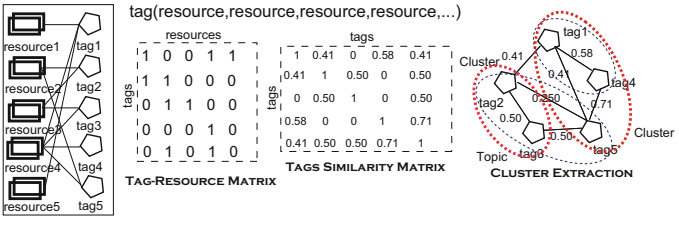


Fig. 1. A Working Example of Tag Clustering

4 Kernel Information Propagation for Tag Clustering Algorithm

So far most tag clustering approaches are mainly dependent on the co-occurrence matrix of tagging data, e.g. the occurrence matrix of resource-tag, to partition tags into various groups. Different from these approaches, our method is instead, to reveal the global prestige degree of each tag via a graph-based partition manner originated from the calculation of kernel density distribution. In the follow section, we discuss the details of our proposed tag clustering algorithm based on kernel information propagation.

4.1 KNN Directed Graph and Kernel Density

According to Definition 1, a similarity matrix S could be constructed from the tagging data, to indicate the affinity of tags. From S, we can find KNN neighbors of each tag and then create a KNN directed graph G, i.e. $G =< V, E >$, where V is the node set of tags and E is the directed edge set between each pair of tags, $< p, q > \in E$ denotes that tag q is a KNN-neighbor of tag p.

Fig.2 shows an example of a part of graph G with two-fold relationships. In one fold, the central black point P has five KNN neighboring nodes denoted by heavy

black circle with arches pointing from P to them, reflecting the neighborhood relationships of P. In the other fold, P is the KNN neighbor of each light circle nodes with arches directing from these nodes to P. In this manner, a KNN directed graph G is constructed and its adjacency matrix A of G is defined as:

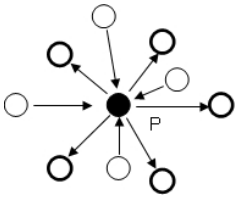


Fig. 2. An Example of KNN Neighbor Directed Graph

Definition 2. *Given a KNN directed graph G, its adjacency matrix is defined as A, where $A(p,q) = 1$, if the directed arch $< p,q >$ exists, and $A(p,q) = 0$, otherwise.*

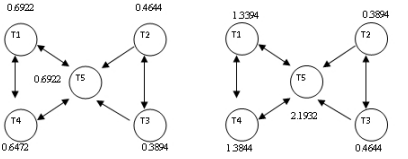
The kernel density estimate method [9] has mainly been used in capturing the local characteristics and density distribution in high dimensional space. In the context of KNN neighbor directed graph, particularly the kernel density function indicates the density distribution of similarity function. Moreover the estimated KNN kernel density information of each node represents the local centrality degree of the node in a possible cluster and the number of arches pointed to the node reflects the "respectful" degree of the node contributed by its neighbors. Bearing this idea in mind, in this paper, we intend to adopt the concept of kernel density of each node, and then define the measure of Local Prestige (LP) to reflect the local centrality information of each node in the KNN neighbor directed graph.

Definition 3. *Given a node $p \in G$ and a backlink node set $B(p)$, of which p is in the KNN neighbors, the Local Prestige (Centrality) of p is defined as the sum of the KNN kernel density of $B(p)$:*

$$LP(p) = \sum_{q \in B(p)} f(q) \tag{2}$$

where $f(q)$ denotes the estimated KNN kernel density of node q.

According to Definition 3, we can see that $LP(p)$ is actually the aggregated kernel density of supporting nodes which choose node p as their KNN neighbors. Intuitively, the higher value of $LP(p)$ the more centrality the node p possesses within the cluster and the more likely the node p becomes the cores of the cluster. Fig.3(a) and (b) illustrate the initial kernel density values and their local prestige degrees of five nodes in the working example.



(a) Kernel Density (b) Local Prestige

Fig. 3. The kernel Information and Local Prestige of Tags in KNN Neighbor Directed Graph

4.2 Kernel Information Propagation for Tag Clustering Algorithm

LP captures the local centrality information of each node in the KNN neighbor directed graph G. Apparently, the prestige score of one node is determined by the number of nodes pointing to this node and their local prestige scores. Meanwhile, known from the theory of directed graph, the prestige of a node is divided evenly and is propagated to other nodes that are pointed by it. Hence the node p gets a boost of its prestige from the nodes that point to p, i.e. the iterative prestige propagation along the directed arches. Inspired by this thought, we envision a new tag clustering algorithm based on Kernel Information Propagation (KIPTC). In particular, we first adopt the Random Walk with Restart (RWR) algorithm [13] to deal with the prestige propagation for global Prestige Rank, and then make use of the global Prestige Rank to form tag clusters. The whole process consists of two main stages. In the stage of global prestige propagation, via the random walk along the constructed KNN neighbor directed graph, the global Prestige Rank is calculated as follows:

Definition 4. *Given PR_i indicating the Prestige Rank scores of all nodes at ith iteration and M denoting the transition matrix of the KNN neighbor directed graph, the updated Prestige Rank at (i+1)th step, PR_{i+1}, is given by:*

$$PR_{i+1} = \alpha \cdot M \cdot PR_i + (1 - \alpha) \tag{3}$$

where α is the damping factor which controls the convergence of the algorithm, and the element of M is determined by $A(q,p)/N(q)$ and $N(q)$ is the total number of outgoing arches of the node q.

The execution of Prestige Rank propagation is repeated until it converges to a stable status. As for the above example shown in Fig.1, the Prestige Rank scores of five nodes are **0.454**, **0.114**, **0.108**, **0.450** and **0.604**, respectively, with $\alpha = 0.85$. From these scores, we can further infer that the node #5 has the highest prestige score amongst all nodes, indicating the highest appropriateness of being a core within one cluster.

After the global prestige scores are calculated, we then utilize them to determine the cores of clusters and include other tag members, which are closely next to the core tag. As the whole KNN neighbor directed graph constitutes a number of connected subgraphs, we then conduct the graph traversal operation

Algorithm 1. Kernel Information Propagation for Tag Clustering

Input: The tag set V and the neighborhood parameter K
Output: The cluster result C

Generate the tag similarity matrix S based on Eq.(1), and calculate the kernel density of each tag;
Construct a KNN directed graph based on Definition 2;
Calculate the Local Prestige LP of each tag using Eq.(2);
while *Not all nodes are calculated* **do**
    update Prestige Rank (PR) via RWR by using Eq.(3);
    Sort the PR scores in a descending order;
end
for *each unvisited tag* v, $v \in V$ **do**
    Select the tag v with the highest PR score;
    Form $C_v = DFS(G, v)$, where C_v denotes a cluster with v as the core;
    $C \leftarrow C \cup C_v$;
end
Return C.

to identify the cluster member nodes. The process to generating the clusters is divided as: (1) we select a tag v, $v \in V$, with the highest PR as the starting core of a cluster; (2) we use the Depth First Search (DFS) method to find the corresponding cluster members until all nodes within the subgraph containing v are completely searched. After that, the algorithm turns to locate a new starting center of another cluster. Steps (1) and (2) are iteratively executed until all the nodes in V are assigned to corresponding cluster. Looking back to the working example, by selecting node #5 as the first cluster core, we include node #1 and #4 as its cluster members via DFS. Then we turn to choose node #2 as the core of the second cluster and add node #1 as its member, eventually resulting in two clusters of $C_1 = 1, 4, 5$ and $C_2 = 2, 3$. Below Algorithm 1 gives the pseudo codes of KIPTC algorithm.

5 Experimental Evaluations

5.1 Datasets and Evaluation Metrics

To evaluate our approach, we conduct preliminary experiments on three real world datasets: MedWorm[3], MovieLens[4] and DMOZ[5]. We perform the experiments using an Intel Core 2 Duo CPU (2.4GHz) workstation with 4G memory, running windows XP. All the algorithms are implemented in Matlab 7.0.

The statistical results of these three datasets are listed in Table 1. These three datasets are pre-processed to filter out some noisy and extremely sparse data subjects to increase the data quality.

[3] http://www.medworm.com/
[4] http://www.movielens.org/
[5] http://www.michael-noll.com/dmoz100k06/

Table 1. Statistics of Experimental Datasets

Property	MedWorm	MovieLens	Dmoz
Number of users	949	4,009	5,016
Number of resources	261,501	7,601	13,771
Number of tags	13,507	16,529	25,311
Total entries	1,571,080	95,580	97,587
Average tags per user	132	11	123
Average tags per resource	5	9	11

In our study, we expect to assign the similar tags serving for the same topic into the same tag cluster. Hence we assume that tag clusters with good quality are composed by lots of similar tags, and the tags in different tag clusters are dissimilar to each other. In particular, we define *Similarity*, *Dissimilarity* and *Silhouette* metric to validate our method.

Definition 5. *Given a tag cluster set* $C = \{C_1, \cdots, C_{|C|}\}$, *the Similarity and Dissimilarity are defined as:* $Similarity(C) = \frac{1}{|C|} \sum_{k=1}^{|C|} \frac{2 \cdot Sim(t_i, t_j)}{|C_k| \cdot (|C_k| - 1)}$, $t_i, t_j \in C_k$; *and* $Dissimilarity(C) = \frac{1}{|C|} \sum_{k=1}^{|C|} \frac{Dissim(k)}{|C_k| \cdot (|T| - |C_k|)}$, *where* $Dissim(k) = \sum_{k=1}^{|C|} Sim(t_i, t_j)$, $t_i \in C_k, t_j \notin C_k$, *and* $|T|$ *is the total number of tags.*

Definition 6. *Given the Similarity and Dissimilarity of cluster* C, *its Silhouette measure is defined as:* $Silhouette\,(C) = 1 - \frac{Dissimilarity(C)}{Similarity(C)}$

Obviously, the higher the silhouette the better the clustering quality is.

5.2 Experiments and Discussions

To evaluate the impact of K (i.e., neighborhood size) selection on clustering, we conduct experiments to evaluate the quality of tag clusters with varying K. Table 2 gives the comparisons in terms of Silhouette on Medworm, MovieLens and Dmoz datasets, respectively. From Table 2, we can first find that the cluster results on three datasets are very close under different K settings. This observation validates that the selection of K does not impose a significant impact on clustering quality. Interestingly, the clustering results derived from Medworm look better than those of Movielens and Dmoz, which might be due to the fact that tags used in Medworm dataset is focused on a more specialized medical domain, while the domain topics of MovieLens and Delicious datasets span more diversely. In order to evaluate the effectiveness of the proposed method, we also implement the traditional clustering algorithm, i.e., K-means on these three real world datasets with the cluster number being set to be the same as that of KIPTC obtained in DFS. The experimental results are shown in Table 3. According to the comparison results of Table 3, we can find that the quality of clustering results obtained by KIPTC is consistently better than that of K-means in terms of Silhouette measure. This finding concludes that KIPTC algorithm has shown

Table 2. Clustering Comparisons of Silhouette with Varying K

	Medworm	Movielens	Dmoz
K=4	0.9752	0.4809	0.7497
K=8	0.9757	0.4908	0.7364
K=12	0.9749	0.4746	0.7476

Table 3. Comparisons of Various Clustering on Silhouette

	Medworm	Movielens	Dmoz
K-means	0.9261	0.1414	0.1912
KIPTC	0.975	0.5377	0.7522

the capability of finding better clustering results than K-means. The reason for this is probably due to that our proposed algorithm is able to, not only capture the local prestige information of tags in the KNN neighbor directed graph G, but also obtain the global centrality of tags via a prestige information propagation. By making use of the global prestige rank, the KIPTS algorithm can effectively locate the more appropriate cluster cores and form the meaningful tag clusters accordingly. In contrast, the K-means clustering algorithm mainly relies on the similarity matrix and demonstrates the poor clustering results, especially with the diverse datasets, e.g., MovieLens and Dmoz.

6 Conclusion and Future Work

Tag clustering is a useful method to find out tag clusters embedded in tagging data and has a potential in improving the performance of tag-based recommender systems. In this paper, we propose a novel tag clustering based on kernel information propagation via random walk on graph. We first use the KNN neighbor directed graph and Kernel density estimate method to find out the local prestige information of each tag, and then employ the random walk with restart algorithm to iteratively propagate the prestige rank until convergence. At last, we use the prestige scores of tags to locate the appropriate cluster core and conduct the graph traversal search to include cluster members. Experimental results conducted on three real world datasets have demonstrated the effectiveness and superiority of the proposed method in comparison to the traditional K-means clustering approach. The future work can be carried out along the direction of comparisons to more state-of-the-art clustering algorithms.

Acknowledgement. This work has been partially supported by EU FP7 ICT project M-Eco: Medical Ecosystem Personalized Event-Based Surveillance (No. 247829); grants from Natural Science Foundation of China (No. 60775037), the Key Program of National Natural Science Foundation of China (No. 60933013), and Research Fund for the Doctoral Program of Higher Education of China (20093402110017).

References

1. Chen, H., Dumais, S.: Bringing order to the web: Automatically categorizing search results. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 145–152. ACM, New York (2000)
2. van Dam, J., Vandic, D., Hogenboom, F., Frasincar, F.: Searching and browsing tag spaces using the semantic tag clustering search framework. In: IEEE Fourth International Conference on Semantic Computing (ICSC), pp. 436–439. IEEE, Los Alamitos (2010)
3. Durao, F., Dolog, P.: Extending a hybrid tag-based recommender system with personalization. In: SAC 2010: Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 1723–1727. ACM, New York (2010)
4. Guan, Z., Bu, J., Mei, Q., Chen, C., Wang, C.: Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 540–547. ACM, New York (2009)
5. Guan, Z., Wang, C., Bu, J., Chen, C., Yang, K., Cai, D., He, X.: Document recommendation in social tagging services. In: Proceedings of the 19th International Conference on World Wide Web, pp. 391–400. ACM, New York (2010)
6. Hayes, C., Avesani, P.: Using tags and clustering to identify topic-relevant blogs. In: International Conference on Weblogs and Social Media (March 2007)
7. Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag recommendations in folksonomies. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 506–514. Springer, Heidelberg (2007)
8. Lehwark, P., Risi, S., Ultsch, A.: Visualization and clustering of tagged music data. Data Analysis, Machine Learning and Applications, 673–680 (2008)
9. Liu, H., Lafferty, J., Wasserman, L.: Sparse nonparametric density estimation in high dimensions using the rodeo. In: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, San Juan, Puerto Rico (2007)
10. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 522–536. Springer, Heidelberg (2005)
11. Noll, M.G., Meinel, C.: Web search personalization via social bookmarking and tagging. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 367–380. Springer, Heidelberg (2007)
12. Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.: Personalized recommendation in social tagging systems using hierarchical clustering. In: Proceedings of the 2008 ACM Conference on Recommender systems, pp. 259–266. ACM, New York (2008)
13. Sun, J., Qu, H., Chakrabarti, D., Faloutsos, C.: Neighborhood formation and anomaly detection in bipartite graphs. In: ICDM, pp. 418–425 (2005)
14. Tso-Sutter, K.H.L., Marinho, L.B., Schmidt-Thieme, L.: Tag-aware recommender systems by fusion of collaborative filtering algorithms. In: SAC 2008: Proceedings of the 2008 ACM Symposium on Applied Computing. pp. 1995–1999. ACM, New York (2008)

An Efficient Itemset Mining Approach for Data Streams

Elena Baralis, Tania Cerquitelli, Silvia Chiusano, Alberto Grand, and Luigi Grimaudo

Politecnico di Torino - Dipartimento di Automatica e Informatica - Torino, Italy
{elena.baralis,tania.cerquitelli,silvia.chiusano,
alberto.grand,luigi.grimaudo}@polito.it

Abstract. This paper presents a new approach to efficiently discovering correlations among data items on a sequence of incoming data windows. The approach enables both on-line (e.g., mining only the most recent data) and off-line (e.g., analyzing aggregate data windows) queries, besides supporting user-defined item and support constraints.

Given a sequence of transactional data windows and a minimum support threshold, for each of the most recent data windows a projection is compactly stored in main-memory, including all items that have been frequently observed in the last windows. Users can easily perform constrained itemset extraction either from a single data window or from multiple ones. A summary of interesting itemsets mined from all available data is generated on a regular basis and compactly stored in a persistent data structure, to efficiently support further analysis (e.g., investigate only a selected past data window).

Experimental results obtained on both real and synthetic data streams show the effectiveness and the efficiency of the proposed approach in mining interesting itemsets by means of both on-line and off-line queries.

Keywords: Itemset extraction, knowledge discovery, data stream analysis.

1 Introduction

Pattern mining on data streams is a relevant research area which finds application in many real-life contexts (e.g., market basket analysis, network traffic analysis, context-aware applications, sensor network data). A data stream can be described as a sequence of incoming data, represented by either (i) single transactions [6,18,12,19,13] or (ii) transactional data windows (i.e., a set of transactions is added to the stream at the same time) [2,5,7,8,3] .

We focus on data streams characterized by a flow of new data windows arriving periodically. In this case, analysts may be interested in mining either the most recent or past data windows. For the first issue, the mining activity is performed on recent and relatively small portions of data. Thus, a real-time response and an accurate result are desirable. For the second issue, as past data windows might

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 515–523, 2011.

include a large amount of data, the analysis process may require a lot of time. Since the data stream is unbounded, efficient techniques should be devised to compactly represent either the relevant past data or the interesting extracted knowledge.

In this paper we propose a novel approach to efficiently performing itemset mining on a sequence of incoming data windows. Different constraints can drive the mining activity (e.g., a minimum support threshold, item constraints) to discover only the most relevant and interesting knowledge hidden in the analyzed data. Given a sequence of transactional data windows, constrained itemset mining can be performed (i) on a single data window, (ii) on multiple ones, or (iii) on past aggregate data windows. For the first two issues, we exploited an array-based data structure to compactly represent in main-memory the most relevant portion of each incoming data window. The extraction of frequent itemsets from streaming data presents additional challenges with respect to mining static data. Since only a subset of the transactional data is available at any given time, maintaining a succinct but complete representation of all meaningful data items becomes troublesome. The completeness of the result is thus often traded off against better scalability [18]. To enhance the accuracy of the mined knowledge, our approach also monitors infrequent data items in the current window that have been frequently observed in the past recent ones. The LCM v.2 algorithm [16] has been selected to efficiently perform the mining task on the most recent and interesting windows. The frequent itemsets, generated at regular intervals from the current portion of the stream available in main-memory, are compactly stored in a persistent data structure [4], which provides support for the mining activity on past data windows.

A preliminary experimental validation, performed on both real and synthetic datasets, shows the effectiveness and the efficiency of the proposed approach in mining relevant and interesting correlations on data streams.

The paper is organized as follows. Section 2 formalizes the problem addressed in this paper. Section 3 describes the proposed approach to efficiently discovering itemsets from a sequence of incoming data windows. Section 4 discusses the preliminary experiments that evaluate the effectiveness of the proposed approach. Finally, Section 5 draws conclusions and presents future developments of the proposed approach.

2 Problem Statement

Many real life data include streams of transactions, and thus can be described as a sequence of data windows arriving periodically. Two kinds of analysis can be performed over such data: (i) Mining the most recent data windows by means of on-line queries and (ii) discovering interesting and recurrent patterns on aggregate past data windows by means of off-line queries. In both cases, users can extract itemsets by enforcing item and/or support constraints.

On-Line Analysis. On-line itemset mining can be perfomed either from a single data window or from multiple ones. Let $\mathcal{I}=\{i_1, i_2, \ldots, i_n\}$ be a set of items.

A transactional data window w is a collection of transactions, where each transaction is a set of items in $\mathcal{I}$. When a new data window w_i arrives at time i (i.e., i uniquely identifies the window), its transactions are added to the data stream. Note that the distribution of items in the windows may be skewed. Let $MinSup$ be the minimum support threshold and w'_i the instance of w_i obtained by removing some items. Each item in w'_i is characterized by its frequency in w'_i, called *window local support*.

Given a minimum support threshold $MSup \geq MinSup$, frequent itemset mining from w'_i is the extraction of the complete set of itemsets with a window local support larger than or equal to $MSup$.

The itemset mining activity on a set of recent windows, not necessarily in sequence, and driven by item and/or support constraints, can be described as follows. Let $\mathcal{W}=\{w'_1, \ldots, w'_n\}$ be a set of recent data windows, not necessarily in sequence. Each item in $\mathcal{W}$ is characterized by its frequency in $\mathcal{W}$, called *window global support*. Given a user-specified set of items $C \subseteq \mathcal{I}$, also called item constraint, and a minimum support threshold $MSup$, constrained itemset mining from $\mathcal{W}$ is the extraction of all itemsets with a window global support larger than or equal to $MSup$ and including at least the items in C.

Off-Line Analysis. Users can be interested in discovering interesting patterns from past data windows, represented in aggregate form. Let $\Omega=\{w'_1, \ldots, w'_k\}$ be the set of recent and continuous data windows, called *aggregate data window*, whose infrequent items with respect to a minimum support threshold $MinSup_\Omega$ have been discarded. Given a minimum support threshold $MSup \geq MinSup_\Omega$, itemset mining from Ω is the extraction of all itemsets with support in Ω larger than or equal to $MSup$.

3 Data Stream Analysis

The itemset mining task can be straightforwardly performed on a collection of static data, but novel challenges arise when dealing with continuously incoming data. In streaming data, we cannot predict whether a given item will be of interest over a time period encompassing several data windows (i.e., frequent w.r.t. the *window global support*, see Section 2). Since maintaining the complete transactional data windows in main memory may be unfeasible, items in each data window may be filtered according to their *window local support* to discard infrequent ones. This approach does not remove any locally relevant item from each window, and thus ensures that the complete set of itemsets is extracted from each data window (without any loss of information). However, the completeness of the result set obtained by mining multiple windows cannot be guaranteed. Some items may be locally infrequent in some windows, but globally frequent on a window set (i.e., their global support satisfies the minimum support threshold). These items are thus relevant for the itemset extraction, but they are underrepresented in the window set. Hence, these items and all of their supersets will appear in the result set with a lower support, or they might not be included in the result set at all, if they no longer meet the minimum support threshold.

To improve the completeness of the result set and the accuracy of the itemset supports, our approach to frequent itemset mining takes into account correlation among data windows. In addition to maintaining locally frequent items, each data window also includes items that, although locally infrequent, have been frequently observed in the past windows. This way, potentially interesting data patterns that exhibit a variable frequency over time can be effectively identified. The proposed approach is outlined in the following.

On-Line Itemset Mining. Given a sequence of transactional data windows w_i and a minimum support threshold $MinSup$, each item in w_i is additionally characterized by a timestamp representing the identifier of the most recent window in which the item has occurred at least $MinSup$ times. For each data window w_i, its projection w_i' is compactly stored in main memory. Let n be a user-defined parameter indicating the number of past windows whose frequent items shall be monitored in the current one. The projected instance w_i' of w_i includes all items satisfying at least one of the following conditions: (i) their support is larger than or equal to $MinSup$, and (ii) their timestamp is no less than $i-n$. Furthermore, for each data window a header table, denoted as H_{w_i}, listing all items monitored in w_i with their corresponding frequencies and timestamps, is stored.

Figure 1 reports a simple sequence of three transactional data windows, w_1, w_2, and w_3 used as a running example, while the corresponding header tables H_{w_i} are shown in Figure 2. Parameter n has been set to 2 and $MinSup = 2$ (i.e., $MinSup = 40\%$).

w_1		w_2		w_3	
TID	**Items**	**TID**	**Items**	**TID**	**Items**
T1	a, b, e, f, g	T6	e	T11	b, c, e, i
T2	a, b	T7	b, e, i	T12	a, b, h, l
T3	e, f, g	T8	b, e	T13	b, d, g, m
T4	a, b, g	T9	d, f, h, i	T14	b, d, g, i, n
T5	a, b, g, h, i	T10	a, c, d, f, g, h	T15	b, d, g

Fig. 1. Running example

Being w_1 the first monitored transactional data window, only frequent items in w_1 are initially observed (i.e., H_{w_1}=$\{a, b, g, e, f\}$, see Figure 2). When analyzing w_2 and w_3, all items that were frequent in previous (1 or 2) windows are also monitored. Thus, H_{w_2} also includes items $\{a, g\}$, while H_{w_3} includes items $\{a, e, h\}$ (see Figure 2). Transactional data are represented in memory by means of an array-based data structure, as proposed in [16], on which the extraction takes place. Each transaction is represented in memory as one array, while transactions including the same frequent items are represented by a single array.

Constrained itemset mining based on a minimum support threshold and/or item constraints can be easily performed either on a single data window or on multiple ones. When the mining process is performed on a single data window with a minimum support threshold larger or equal to $MinSup$, the complete

Window	w_1					w_2								w_3						
Item	a	b	g	e	f	e	b	d	f	h	i	a	g	b	d	g	i	a	e	h
Local support	4	4	4	2	2	3	2	2	2	2	2	1	1	5	3	3	2	1	1	1
Timestamp	1	1	1	1	1	2	2	2	2	2	2	1	1	3	3	3	3	1	2	2

Fig. 2. Header tables for the 3 data windows

set of itemsets is extracted. An approximate result set may be returned when multiple windows are involved in the query (see previous discussion).

On-line frequent itemset extraction is performed in two steps. First, eligible items, i.e., items which satisfy the support constraint, are selected. Given $MSup$ and $\mathcal{W}$=$\{w_1, \ldots, w_n\}$, i.e., the set of recent windows which we would like to query, the corresponding header tables $\mathcal{H}$=$\{H_{w_1}, \ldots, H_{w_n}\}$ are read from main-memory. For each item in $\mathcal{H}$ the *window global support* is computed by summing the item local supports for all windows in $\mathcal{W}$. Items with a window global support lower than $MSup$ are pruned. Consider, for example, $\mathcal{W}$=$\{w_2, w_3\}$, and $MSup \geq 4$ (i.e., $MSup \geq 40\%$). The set of eligible items with their corresponding window global support is $\mathcal{H}$=$\{< b : 7 >, < d : 5 >, < e : 4 >, < g : 4 >, < i : 4 >\}$. Note that, thanks to the proposed approach, item e has not been discarded in w_3, although infrequent, on account of its high support in w_2. Thus, its support can be correctly computed in the joint query on windows $\{w_2, w_3\}$. The same holds for item g.

Then, the extraction process is performed for each eligible item by exploiting LCM v.2 [16], a very efficient state-of-the-art algorithm whose source code, available at [15], has been integrated into our framework. LCM v.2 currently outperforms any other state-of-the-art approach, thanks to an efficient support counting technique, the occurrence deliver, and to the exclusive use of array-based data structures instead of sophisticated data structures such as prefix trees [10,11].

Off-Line Itemset Mining. Since a data stream cannot be stored in main-memory due to its unbounded size, we exploited a persistent data structure, called Array-Tree [4], to compacly store frequent itemsets mined from k (a user-defined parameter) aggregate past data windows and allow efficient queries.

The Array-Tree compactly represents the set of frequent itemsets extracted from a sequence of k incoming data windows by enforcing a given minimum support threshold $MinSup_\omega$. The Array-Tree exploits (i) a prefix-tree like structure based on arrays to succinctly store its nodes, and (ii) both prefix-path sharing (i.e., each itemset is represented by a single tree path, but a prefix-path may represent the common prefix of multiple itemsets) and subtree sharing (i.e., portions of the tree with the same structure, in terms of nodes, entries and pointers are written on disk only once) to reduce data replication in the tree, thus increasing its compactness. Furthermore, it provides support to user queries either enforcing a minimum support constraint, or requesting itemsets including a specified set of items.

4 Preliminary Experimental Results

We validated our on-line approach by means of different experiments addressing the following issues: (i) Performance of itemset extraction by varying both the selected queried windows and the support threshold and (ii) accuracy of the mined results.

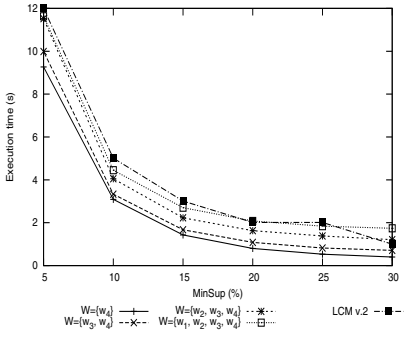
We report experiments on 4 representative datasets whose characteristics are in Table 1. Connect and Kosarak [15] are real and medium-size datasets. The last two datasets (i.e., TxPyIiCwDz) reported in Table 1 have been synthetically generated by means of the IBM generator [1] by setting different parameters (i.e., T average size of transactions, I number of different items, P average length of maximal patterns, C correlation grade between patterns, and D number of transactions). All experiments were performed on a 2.66GHz dual-processor quad-core Intel Core 2 Quad Q9400 PC with 8Gbyte main memory running Linux kernel v.2.6.32-28-server. The reported execution times are real times, including both system and user times, obtained from the Unix time command as in [15].

Table 1. Dataset characteristics

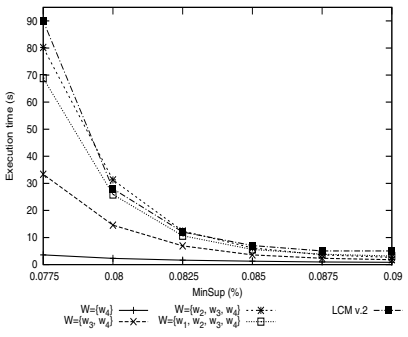
Dataset	Transactions	Items	AvgTrSz	Size (MB)
CONNECT	67,557	129	43	9.11
KOSARAK	990,002	41,270	8.1	30.55
T22P22I100kC1D3M	3,000,000	48,068	27.93	473.25
T24P22I50KC1D10M	10,000,000	32,442	29.20	1787.05

Itemset Extraction Performance. To simulate a sequence of incoming data windows we have split each original dataset in four data windows respectively with the same number of transactions. Figure 3 shows the run time for frequent itemset extraction by enforcing different support thresholds. Given the minimum support threshold $MinSup$, in each data window we monitored (i) frequent items in the observed window and (ii) all frequent items observed in the last 3 windows. As shown in Figure 3, the proposed approach has been exploited to perform different mining sessions by querying different recent data windows (i.e., windows in $\mathcal{W}$). By increasing the number of the considered windows the execution time increases because a large number of transactions needs to be mined. Since a large number of frequent itemsets are mined by decreasing the minimum support threshold ($MSup$), the execution time required to perform the mining process increases in all considered data windows.

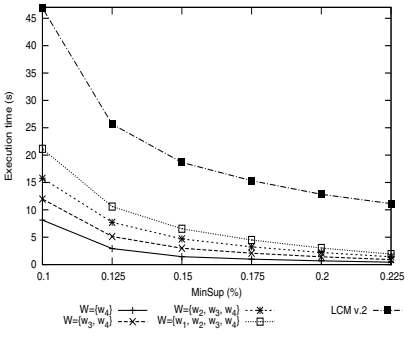
Figure 3 also reports the run time for the LCM v.2 algorithm [16] on the complete flat file dataset. For any considered supports, our approach on $\mathcal{W}$= $\{w_1, w_2, w_3, w_4\}$ often yields performance comparable to and sometimes better than LCM v.2. The slight difference in execution time is mainly due to the time devoted by LCM v.2 to reading the complete dataset and loading its frequent projection in main memory, while our approach performs the mining process on the projections already available in main memory. Thus, the access time to



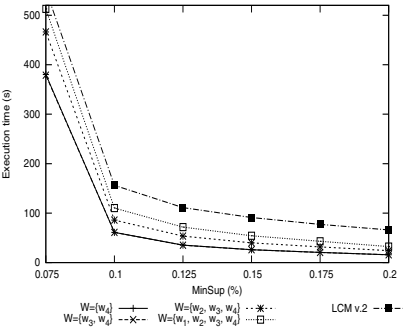
(a) Connect, $MinSup = 5\%$



(b) Kosarak, $MinSup = 0.0775\%$



(c) T22P22I100kC1D3M, $MinSup = 0.1\%$



(d) T24P22I50KC1D10M, $MinSup = 0.075\%$

Fig. 3. Frequent itemset extraction time

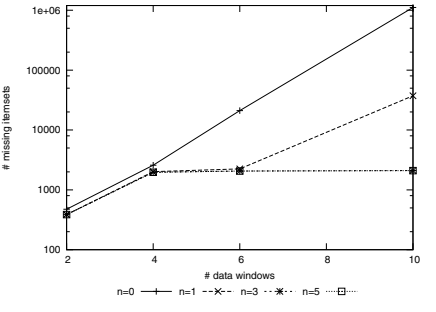


Fig. 4. Effect of stream partitioning on the T24P22I50KC1D10M dataset, $MinSup$=0.075%

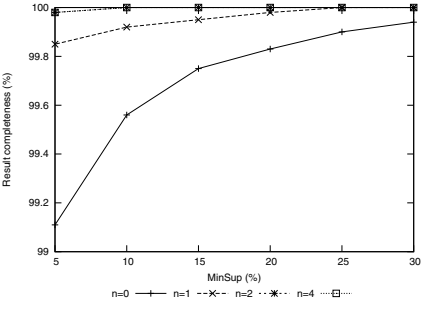


Fig. 5. Accuracy on the Connect dataset, 4 data windows

the main memory window projections is lower. Furthermore, the accuracy of the mined knowledge yielded by our approach is very good, since the difference between the number of frequent itemsets mined by our approach and LCM v.2 on the complete dataset is always negligible (see below for further discussion) for any considered support thresholds.

Accuracy of the Extracted Knowledge. To investigate the effectiveness of our approach in enhancing the accuracy of the extracted knowledge, we analyzed: (i) Effect of data stream partitioning and (ii) effect of parameter n (see Section 3), representing the number of windows whose frequent items are monitored in the current one.

Figure 4 analyzes the number of itemsets missing from the result set yielded by data stream partitioning. T24P22I50KC1D10M is discussed as a representative dataset. The analysis has been performed by splitting the original dataset in 2, 4, 6, and 10 data windows, and by considering different values of parameter n. As shown in Figure 4, the number of missing itemsets increases as the data stream is partitioned into more data windows, while it decreases when the value of parameter n is increased. Thus, by setting an appropriate value for this parameter, better accuracy can be achieved.

Furthermore, we analyzed the effect of parameter n by enforcing different support thresholds. Since the Connect dataset is characterized by the densest data distribution, it is discussed as a representative example. Figure 5 shows the percentage of frequent itemsets extracted in each configuration with respect to the complete result set. Experiments demonstrate that the result completeness decreases by lowering the minimum support threshold, because a larger number of potentially meaningful items are discarded during the analysis of the data stream. By setting higher values for parameter n, the result completeness improves, because in each data window a higher number of frequent items from the past windows is monitored.

5 Conclusion and Future Works

This paper describes an efficient approach to mining a sequence of incoming data windows. The experimental results showed the effectiveness and the efficiency of the proposed approach in mining interesting itemsets on data streams.

As future work, we plan to: (i) Design an efficient algorithm to query a set of selected materialized Array-Trees, (ii) provide support for different interestingness constraints (e.g., as proposed in [14]) to drive both on-line and off-line querying, and (iii) exploit the proposed approach to mine more compact itemset representations (e.g., closed itemset [17], maximal itemset [9]).

References

1. Agrawal, N., Imielinski, T., Swami, A.: Database mining: A performance perspective. IEEE Trans. on Knowledge and data Engineering 5(6) (1993)
2. Aumann, Y., Feldman, R., Lipshtat, O.: Borders: An efficient algorithm for association generation in dynamic databases. JIIS 12 (1999)

3. Baralis, E., Cerquitelli, T., Chiusano, S.: Constrained itemset mining on a sequence of incoming data blocks. Int. J. Intell. Syst. 25, 389–410 (2010)
4. Baralis, E., Cerquitelli, T., Chiusano, S., Grand, A.: Array-tree: A persistent data structure to compactly store frequent itemsets. In: IEEE Conf. of Intelligent Systems, pp. 108–113 (2010)
5. Cheung, D.W.-L., Han, J., Ng, V., Wong, C.Y.: Maintenance of discovered association rules in large databases: An incremental updating technique. In: ICDE, pp. 106–114. IEEE Computer Society, Los Alamitos (1996)
6. Cheung, W., Zaiane, O.R.: Incremental mining of frequent patterns without candidate generation or support constraint. In: IDEAS, pp. 111–116 (July 2003)
7. Ganti, V., Gehrke, J., Ramakrishnan, R.: DEMON: Mining and monitoring evolving data. IEEE Trans. Knowl. Data Eng. 13(1), 50–63 (2001)
8. Ganti, V., Gehrke, J.E., Ramakrishnan, R.: Mining data streams under block evolution. SIGKDD Explorations 3(2) (January 2002)
9. Gouda, K., Zaki, M.J.: Efficiently mining maximal frequent itemsets. In: ICDM 2001: Proceedings of the 2001 IEEE International Conference on Data Mining, pp. 163–170. IEEE Computer Society, Washington, DC, USA (2001)
10. Grahne, G., Zhu, J.: Efficiently using prefix-trees in mining frequent itemsets. In: FIMI 2003 Workshop on Frequent Itemset Mining Implementations (November 2003)
11. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: In SIGMOD 2000, Dallas, TX (May 2000)
12. Li, K., Wang, Y.-y., Ellahi, M., Wang, H.-a.: Mining recent frequent itemsets in data streams. In: Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2008, vol. 04, pp. 353–358. IEEE Computer Society, Washington, DC, USA (2008)
13. Liu, H., Lin, S., Qiao, J., Yu, G., Lu, K.: An efficient frequent pattern mining algorithm for data stream. In: Proceedings of the 2008 International Conference on Intelligent Computation Technology and Automation, vol. 01, pp. 757–761. IEEE Computer Society, Washington, DC, USA (2008)
14. Pei, J., Han, J., Lakshmanan, L.V.S.: Pushing convertible constraints in frequent itemset mining. Data Min. Knowl. Discov. 8(3), 227–252 (2004)
15. F. repository, http://fimi.cs.helsinki.fi/
16. Uno, T., Kiyomi, M., Arimura, H.: LCM ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. In: FIMI (2004)
17. Wang, J., Han, J., Pei, J.: Closet+: searching for the best strategies for mining frequent closed itemsets. In: KDD 2003: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 236–245. ACM, New York (2003)
18. Wang, Y., Li, K., Wang, H.: Maintaining only frequent itemsets to mine approximate frequent itemsets over online data streams. In: IEEE Symposium on Computational Intelligence and Data Mining, pp. 381–388 (2009)
19. Yang, C.-H., Yang, D.-L.: Imbt-a binary tree for efficient support counting of incremental data mining. In: CSE (1), pp. 324–329 (2009)

The Representation of Inconsistent Knowledge in Advanced Knowledge Based Systems

Mark Burgin[1] and Kees (C.N.J.) de Vey Mestdagh[2]

[1] Department of Computer Science, University of California,
Los Angeles, United States of America
[2] Centre for Law & ICT, University of Groningen,
Groningen, The Netherlands

Abstract. Contradiction handling is one of the central problems in AI. There are different approaches to dealing with contradictions and other types of inconsistency. We describe an approach based on logical varieties, which are complex structures constructed from logical calculi. Being locally isomorphic to a logical calculus, globally logical varieties allow representation of contradictory knowledge in a consistent way, providing much more flexibility and efficacy for AI than standard logical methods. Problems of logical variety immersion into a logical calculus are studied. Such immersions extend the local structure of a logical calculus to the global structure of a logical variety, demonstrating when it is possible to use standard logical tools, such as logical calculi, and when it is necessary to go beyond this traditional technique. Finally a particular logical variety, the Logic of Reasonable Inferences, applied to the design of legal knowledge based systems is described.

Keywords: Inconsistent knowledge, Logical Varieties, Logic of Reasonable Inferences, Legal Knowledge Based Systems.

1 Introduction

Minsky [20] was one of the first researchers in AI who attracted attention to the problem of inconsistent knowledge. He wrote that consistency is a delicate concept that assumes the absence of contradictions in systems of axioms. Minsky also suggested that in artificial intelligence (AI) systems this assumption was superfluous because there were no completely consistent AI systems. In his opinion, it is important to understand how people solve paradoxes, find a way out of a critical situation, learn from their own or others' mistakes or how they recognize and exclude different inconsistencies. Minsky [21] suggested that consistency and effectiveness may well be incompatible. He also writes [22]: "An entire generation of logical philosophers has thus wrongly tried to force their theories of mind to fit the rigid frames of formal logic. In doing that, they cut themselves off from the powerful new discoveries of computer science. Yes, it is true that we can describe the operation of a computer's hardware in terms of simple logical expressions. But no, we cannot use the same expressions to describe the meanings of that computer's output -- because that

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 524–537, 2011.

would require us to formalize those descriptions inside the same logical system. And this, I claim, is something we cannot do without violating that assumption of consistency." Then Minsky [22] continues, "In summary, there is no basis for assuming that humans are consistent - not is there any basic obstacle to making machines use inconsistent forms of reasoning". Moreover, it has been discovered that not only human knowledge but also representations/models of human knowledge (e.g., large knowledge bases) are inherently inconsistent (Delgrande, *et al*, [9]).

It is necessary to remark that inconsistencies bothered logicians from the time of Aristotle. However, the first logical systems treating contradictions appeared only in the 20th century. At first, they had the form of multivalued logics developed by Vasil'év and Łukasiewicz. Then the first relevant logics were built by Orlov. However, their work did not make any impact at the time and the first logician to have developed formal paraconsistent logic was a student of Łukasiewicz, Jaśkowski, [14]. Starting from this time, a diversity of different paraconsistent logics, including fuzzy logics, multivalued logics and relevant logics, has been elaborated (cf., for example, Routley, *et al*,[28]).

The perspective-bound character of information and information processing often results in natural inconsistency coming from different perspectives or from a faulty perception or from faulty information processing, such as processing on the basis of incomplete knowledge from a single perspective. As a result, now many understand that contradiction handling is one of the central problems in AI. Inconsistent knowledge/belief systems exist in many areas of AI, such as distributed knowledge base and databases, defeasible reasoning, dynamic expert systems, merging ontologies, ontology evolution, knowledge transition from one formalism to another, and belief revision.

There are three basic approaches to dealing with inconsistency. The first one is aimed at restoring consistency of an inconsistent knowledge system, e.g., a database (Rescher and Manor, [26]). Another approach is to tolerate inconsistency (Bertossi, *et al*, [5]) by including an inconsistent knowledge system into a paraconsistent or fuzzy logic (Priest, *et al*, [25]; Ross, [27]) and using this logic for inference in the given knowledge system. The third way is based on implicit or explicit utilization of logical varieties, quasi-varieties and prevarieties.

In comparison with non-monotonic logics, which form the base for the first approach, logical varieties, quasi-varieties and prevarieties provide tools for preserving all points of view, approaches and positions even when some of them taken together lead to contradiction. Due to their flexibility, logical varieties, quasi-varieties and prevarieties allow treating any form of logical contradictions in a rigorous and consistent way.

Paraconsistent logics, which form the base for the second approach, are inferentially *weaker* than classical logic; that is, they deem *fewer* inferences valid. Thus, in comparison with paraconsistent logics, logical varieties, quasi-varieties and prevarieties allow utilization of sufficiently powerful means of logical inference, for example, deductive rules of the classical predicate calculus. Besides, Weinzierl [31] explains why paraconsistent reasoning is not acceptable for many real-life scenarios and other approaches are necessary. In addition, paraconsistent logics attempts to deal with contradictions in a discriminating way, while logical varieties, quasi-varieties and prevarieties treat contradictions and other inconsistencies by a separation technique.

Although conventional logical systems based on logical calculi have been successfully used in mathematics and beyond, they have definite limitations that often restrict their applications. For instance, the principal condition for any logical calculus is its consistency. At the same time, knowledge about large object domains (in science or in practice) is essentially inconsistent (Burgin, [7]; de Vey Mestdagh, [19]; Nguen, [23]). From this perspective, Partridge and Wilks ([24]) write, "because of privacy and discretionary concerns, different knowledge bases will contain different perspectives and conflicting beliefs. Thus, all the knowledge bases of a distributed AI system taken together will be perpetually inconsistent." Consequently, when conventional logic is used for formalization, it is possible to represent only small fragments of the object domain. Otherwise, contradictions appear.

To eliminate these limitations in a logically correct way, logical prevarieties and varieties were introduced (Burgin, [7]). Logical varieties represent the natural development of logical calculi, being more advanced systems of logic, and thus, they show the direction in which mathematical logic will inevitably go. Including logical calculi as the simplest case, logical varieties and related systems offer several advantages over conventional logic:

1. Logical varieties, prevarieties and quasi-varieties give an exact and rigorous structure to deal with all kinds of inconsistencies.
2. Logical varieties allow modeling/realization of all other approaches to inconsistent knowledge. For instance, it is possible to use any kind of paraconsistent logics as components of logical varieties. In (Burgin, [7]), it is demonstrated how logical varieties realize non-monotonic inference.
3. Theoretical results on logical varieties provide means for more efficient application of logical methods to problems in different areas (e.g. the application of the Logic of Reasonable Inferences (a logical variety) to represent and process contradicting opinions in the legal domain as described below).
4. Logical varieties allow partitioning of an inconsistent knowledge system into consistent parts and to use powerful tools of classical logic for reasoning.
5. Logical varieties allow utilization of different kinds of logic (multifunctionality) in the same knowledge system. For instance, it is possible to use a combination of the classical predicate calculus and non monotonic calculus to represent two perspectives one of which is based on complete knowledge and the other on incomplete knowledge
6. Logical varieties allow separation of different parts in a knowledge system and working with them separately.
7. Logical varieties provide means to reflect change of beliefs, knowledge and opinions without loss of previously existed beliefs, knowledge and opinions even in the case when new beliefs, knowledge and opinions contradict to what was before. In (Burgin, [8]), they are applied to temporal databases.

These qualities of logic varieties are especially important for normative, in particular, legal, knowledge because this knowledge consists of a collection of formalized systems, a collection of adopted laws, a collection of existing traditions and precedents, and a collection of people's opinions. In addition, in the process of functioning, normative (legal) knowledge involves a variety of situational knowledge, beliefs and opinions. To analyze and use this diversity, it is necessary to have a

flexible system that allows one to make sense of all different approaches without discarding them in an attempt to build a unique consistent system. To formalize these characteristics of normative knowledge, a form of a logical variety, the Logic of Reasonable Inferences (LRI) was developed (de Vey Mestdagh [18]). The LRI was subsequently used as specification for the implementation of a knowledge based system shell, Argumentator (de Vey Mestdagh [19]). This shell has consequently been used to acquire and represent legal knowledge. The resulting legal knowledge based system has been successfully used to test the empirical validity of the theory about legal reasoning and decision making modeled by the LRI (de Vey Mestdagh [19]).

It is interesting that several other systems used for inconsistency resolution, e.g., Multi-Context Systems (Weinzierl [31]), are also logical varieties and prevarieties. For instance, bridge rules used in Multi-Context Systems for non-monotonic information exchange are functions that glue together components of a logical variety or prevariety (cf. Definition 2.1).

In section 2 we define the concepts of quasi varieties, prevarieties and varieties formally. In section 3 we describe the immersion of logical variety into a logical calculus. The mathematical results presented in this section are new. In section 4 we describe the Logic of reasonable Inferences as a form of logical variety and its use to represent inconsistent legal knowledge.

2 Logical Quasi-Varieties, Prevarieties, and Varieties

There are different types and kinds of logical varieties and prevarieties: Deductive or syntactic varieties and prevarieties, Functional or semantic varieties and prevarieties and Model or pragmatic varieties and prevarieties. Semantic logical varieties and prevarieties are formed by separating those parts that represent definite semantic units. In contrast to syntactic and semantic varieties, model varieties are essentially formal structures.

Syntactic varieties, quasi-varieties and prevarieties are built from logical calculi as buildings are built from blocks. That is why, we, at first remind the concept of a logical calculus.

Let us consider a logical language L and an inference language R.

Definition 2.1. A *syntactic* or *deductive logical calculus*, usually called *logical calculus*, is a triad (a named set) of the form $\boldsymbol{C} = (A, H, T)$ where $H \subseteq R$ and $A, T \subseteq L$, A is the set of axioms, H consists of inference rules (rules of deduction) by which from axioms the theorems of the calculus are deduced, and the set of theorems T is obtained by applying algorithms/procedures/rules from H to elements from A.

Let **K** be some class of syntactic logical calculi, R be a set of inference rules, and **F** be a class of partial mappings from L to L.

Definition 2.2. A triad $\mathbf{M} = (A, H, M)$, where A and M are sets of expressions that belong to L (A consists of axioms and M consists of theorems) and H is a set of inference rules, which belong to the set R, is called:

(1) a *projective syntactic* (**K,F**)-*quasi-prevariety* if there exists a set of logical calculi $\boldsymbol{C}_i = (A_i , H_i , T_i)$ from **K** and a system of mappings $f_i : A_i \to L$ and $g_i : M_i \to L$ $(i \in I)$ from **F** in which A_i consists of axioms and M_i consists of some (not necessarily all) theorems of the logical calculus $\boldsymbol{C}_i$, and for which the equalities $A = \bigcup_{i\in I} f_i(A_i)$, $H = \bigcup_{i\in I} H_i$ and $M = \bigcup_{i\in I} g_i(M_i)$ are valid (it is possible that $\boldsymbol{C}_i = \boldsymbol{C}_j$ for some $i \neq j$).

(2) a *syntactic* **K**-*quasi-prevariety* if it is a projective syntactic (**K,F**)- quasi-prevariety where all mappings f_i and g_i that define **M** are inclusions, i.e., $A = \bigcup_{i\in I} A_i$ and $M = \bigcup_{i\in I} M_i$;

(3) a *projective syntactic* (**K,F**)-*quasi-variety* with the depth k if it is a projective syntactic (**K,F**)-quasi-prevariety and for any $i_1 , i_2 , i_3 , \dots , i_k \in I$ either the intersections $\bigcap_{j=1}^{k} f_{ij}(A_{ij})$ and $\bigcap_{j=1}^{k} g_{ij}(T_{ij})$ are empty or there exists a calculus $\boldsymbol{C} = (A, H, T)$ from **K** and projections f: $A \to \bigcap_{j=1}^{k} f_{ij}(A_{ij})$ and g: $N \to \bigcap_{j=1}^{k} g_{ij}(M_{ij})$ from **F** where $N \subseteq T$;

(4) a *syntactic* **K**-*quasi-variety* with the depth k if it is a projective syntactic (**K,F**)-quasi-variety with depth k in which all mappings f_i and g_i that define **M** are bijections on the sets A_i and M_i, correspondingly.

(5) a (*full*) *projective syntactic* (**K,F**)-*quasi-variety* if for any $k > 0$, it is a projective syntactic (**K,F**)-*quasi-variety* with the depth k;

(6) a (*full*) *syntactic* **K**-*quasi-variety* if for any $k > 0$, it is a **K**-quasi-variety with the depth k;

(7) a *projective syntactic* (**K,F**)-prevariety if it is a projective syntactic (**K,F**)-quasi-prevariety in which $M_i = T_i$ for all $i \in I$;

(8) a *syntactic* **K**-*prevariety* if it is a syntactic (**K,F**)-quasi-prevariety in which $M_i = T_i$ for all $i \in I$;

(9) a *projective syntactic* (**K,F**)-*variety* with the depth k if it is a projective syntactic (**K,F**)-quasi-prevariety in which $M_i = T_i$ for all $i \in I$;

(10) a *syntactic* **K**-*variety* with the depth k if it is a projective syntactic (**K,F**)-quasi-variety with depth k in which $M_i = T_i$ for all $i \in I$;

(11) a (*full*) *projective syntactic* (**K,F**)-*variety* if for any $k > 0$, it is a projective syntactic (**K,F**)-variety with the depth k;

(12) a (*full*) *syntactic* **K**-*variety* if for any $k > 0$, it is a **K**-variety with the depth k.

We see that the collection of mappings f_i and g_i makes a unified system called a prevariety or quasi-prevariety out of separate logical calculi $\boldsymbol{C}_i$, while the collection of the intersections $\bigcap_{j=1}^{k} f_{ij}(A_{ij})$ and $\bigcap_{j=1}^{k} g_{ij}(T_{ij})$ makes a unified system called a variety out of separate logical calculi $\boldsymbol{C}_i$. For instance, mappings f_i and g_i allow one to establish a correspondence between norms/laws that were used in one country during different periods of time or between norms/laws used in different countries.

The main goal of syntactic logical varieties is in presenting sets of formulas as a structured logical system using logical calculi, which have means for inference and other logical operations. Semantically, it allows one to describe a domain of interest, e.g., a database, knowledge of an individual or the text of a novel, by a syntactic logical variety dividing the domain in parts that allow representation by calculi.

In comparison with varieties and prevarietis, logical quasi-varieties and quasi-prevarieties are not necessarily closed under logical inference. This trait allows better flexibility in knowledge representation.

Definition 2.4. The calculi $\boldsymbol{C}_i$ used in the formation of the prevariety (variety) **M** are called *components* of **M**.

An example of a logical variety is a distributed database or knowledge base, each component of which consists of consistent knowledge/data. Then components of this knowledge/database are naturally represented by components of a logical variety. Besides, in one knowledge base different object domains may be represented. In these domains some object may have properties that contradict properties of an object from another domain. As an example let us consider a knowledge base containing mathematical information. Suppose that this information concerns some large mathematical field like algebra or even its part - theory of groups. Mathematical logics are frequently considered to be the basis of mathematics and logical calculi are viewed as precise models and formalizations of real mathematical theories. But the theory of groups does not coincide with elementary (logical) theory of groups that is a deductive calculus. The field that is called in mathematics "the Theory of Groups" contains various subtheories (Hall, [13]).

In the theory of groups, such mathematical objects as finite and torsion-free groups are studied. In any finite group, the formula $\forall x\, \exists n\, (x^n = e)$ is valid where e is the identity element. At the same time, in torsion-free groups another formula $\forall x \forall n\, \neg (x^n = e)$ is true. Thus, if theory of groups with its subtheories, such as the theory of finite groups and theory of torsion-free groups, is represented as a single calculus, then both these formulae produce a contradiction. At the same time, a relevant logical variety in which subtheories are represented by its components provides means for consistent representation of the theory of groups.

Inference in a logical variety **M** is restricted to inference in its components because at each step of inference, it is permissible to use only rules from one set H_i applying these rules only to elements from the set T_i. This allows one to better model non-monotonicity of human thinking.

Indeed, the main difference between monotonic and non-monotonic reasoning arises from the different kinds of knowledge used in the process of inference. For instance, in the case of non-monotonic reasoning an inference rule of the following

type can be used: "*A* is true if *B* cannot be proved", i.e. to prove *A* the system relies on its ignorance of *B*. The statement *B* is not included in the system of initial axioms. That is why by the given above rule of inference, the statement A becomes true in the intellectual system. However, it is possible that *B* becomes proved at some stage of the inference. So in this situation, the intellectual system must invalidate *A* and even more - to revise each piece of knowledge depending on *A*. In this way the monotonic property of the consequence relation is violated. Usually, the statement *A* is excluded and the knowledge/belief revision takes place. Logical varieties allow not to eliminate knowledge/beliefs in the process of revision but to build a new component from which all knowledge/beliefs that contradict *B* are eliminated. In such a way, all previously obtained knowledge/beliefs are preserved.

Although any logical calculus is a logical variety, this particular case does not give anything new in logic because logical calculi already exist in logic. A non-trivial example of logical varieties is given by *many-sorted logics* (Meinke and Tucker, [17]; Abadi, *et al*, [1]). In these logics, the variables range over different domains. Consequently, logical variables are "typed" as variables in many computer programming languages. Many-sorted logics allow one not to work with the domain of discourse as a homogeneous collection of objects, but to partition this domain into several parts with various functions and relations connecting them. In this case, these parts being formalized form a model variety, while the system of logics that describe these parts forms a syntactic variety.

For instance, semantics of computer languages employ different types (domains) of data, such as the integers and the real numbers. Each domain has its own equality, relations, identities, and arithmetical operations. The logical language that describes the union of these domains will have two sorts of variables, real variables and integer variables. The meaning of a quantifier would be determined by the type of the variable it binds. The corresponding logic will be a logical variety built of two calculi. Intersection of these calculi will include such formulas as the commutative law

$$x + y = y + x$$

and the associative law

$$x + (y + z) = (x + y) + z$$

Any big mathematical theory, such as group theory, ring theory or topology (theory of topological spaces), forms a syntactic logical variety. For instance, group theory as the set of all consistent formulas in the formal language of group theory contains many subtheories, such as the theory of finite groups, the theory of toirsion-free groups, the theory of abelian groups, the theory of nilpotent groups, and the theory of prime groups. In the theory of finite groups, the formula that says that any element of a finite group, there is a number *n* such that taken to the power *n*, this element is equal to the unit of the group, is valid, while in the theory of toirsion-free groups, the negation of this formula is valid. This negation says that any element of a finite group, there is no number *n* such that taken to the power *n*, this element is equal to the unit of the group. Thus, group theory, as it is understood in mathematics and not the formal theory of groups in the logical sense, is not a calculus because it is incompatiple, but it is a logical variety, which has weight > 1.

We have a similar situation in the ring theory, which contains the theory of commutative rings and the theory of Lie rings. These two subtheories contain formalas such that one of them is the nagation of the other.

A similar situation also exists in other disciplines, for example, in archeology (VanPool and VanPool, [30]).

One more example of naturally formed logical varieties is the technique *Chunk and Permeate* built by Brown and Priest [6]. This technique suggests to begin reasoning from inconsistent premisses and proceeds by separating the assumptions into consistent theories (called by the authors *chunks*). These chunks are components of the logical variety shaped by them. After this, appropriate consequences arederived in one component (chunk). Then those consequences are transfered to a different component (chunk) for further consequences to be derived. This is exactly the way how logical varieties are used to realize and model nonmonotonic reasoning (Burgin, [7]). Brown and Priest suggest that Newton's original reasoning in taking derivates in the calculus, was of this form.

An interesting type of logical varieties was developed in artificial intelligence and large knowledge bases. As Amir and McIlraith write ([3], [15], [16]), there is growing interest in building large knowledge bases of everyday knowledge about the world, comprising tens or hundreds of thousands of assertions. However working with large knowledge bases, general-purpose reasoning engines tend to suffer from combinatorial explosion when they answer user's queries. A promising approach to grappling with this complexity is to structure the content into multiple domain- or task-specific partitions. These partitions generate a logical variety comprising the knowledge base content. For instance, a first-order predicate theory or a propositional theory is partitioned into tightly coupled subtheories according to the language of the axioms in the theory. This partitioning induces a graphical representation where a node represents a particular partition or subtheory and an arc represents the shared language between subtheories.

The technology of content partitioning allows reasoning engines to improve the efficiency of theorem proving in large knowledge bases by identifying and exploiting the implicit structure of the knowledge (Amir and McIlraith, [3]; McIlraith and Amir, [16]; MacCartney, *et al*, [15]),. The basic approach is to convert a graphical representation of the problem into a tree-structured representation, where each node in the tree represents a tightly-connected subproblem, and the arcs represent the loose coupling between subproblems. To maximize the effectiveness of partition-based reasoning, the coupling between partitions is minimized, information being passed between nodes is reduced, and local inference within each partition is also minimized.

Additional advantage of portioning is a possibility to reason effectively with multiple knowledge bases that have overlap in content (Amir and McIlraith, [3]).

The tools and methodology of content partitioning and thus, implicitly of logical varieties are applied for the design of logical theories describing the domain of robot motion and interaction (Amir, [2]).

Concepts of logical varieties and prevarieties provide further formalization for local logics of Barwise and Seligman ([4]), many-worlds model of quantum reality of Everett (Everett, [11]; DeWitt, [10]), and pluralistic quantum field theory of Smolin related to the many-worlds theory (Smolin, [29]).

3 Compatibility in Logical Varieties

An important problem of logic is to combine logics including all of them into one calculus. Gabbay ([12]) writes that "the problem of combining logics and systems is central for modern logic, both pure and applied. The need to combine logics starts both from applications and from within logic itself as a discipline. As logic is being used more and more to formalize field problems in philosophy, language, artificial intelligence, logic programming, and computer science, the kind of logics required becomes more and more complex." Logical varieties and prevarieties give a relevant context for solving this problem because it is possible to treat any system of logics as a logical variety or prevariety. Here we consider only deductive varieties.

Let us take a class **K** of logical calculi and a deductive **K**-variety $\mathbf{M} = \{C_i ; i \in I\}$.

Definition 3.1. A logical variety **M** is called:

1) *Discrete* if its components are disjoint;
2) *Classical* if all its components are classical deductive calculi;
3) *Connected* if any two of its components have a non-void intersection;
4) *Compatible* if it is a subset of a consistent calculus.
5) *Provably compatible* if it is possible to prove by classical methods that it is a subset of a consistent calculus.

Definition 3.2. A set of components $\{C_i ; i \in J\}$ of **M** is called (provably) compatible if the subvariety of **M** generated by these components is (provably) compatible

Lemma 1.

a) For any deductive variety M, there is a discrete deductive variety DM such that their upper levels are equal, i.e., T(M) = T(DM).
b) The discrete counterpart DM of a variety M preserves consistency.

Proposition 1. If **M** is compatible (**K**-compatible), then **DM** is compatible (**K**-compatible).

Theorem 1. For any number $n > 1$ there is a classical connected deductive logical variety M with n components such that any n - 1 components of **M** are provably compatible, **M** is compatible, but it is not provably compatible.

Remark 1. The condition that the variety is classical is essential. However, for limit ordinals similar results are not valid.

Theorem 2. For any classical deductive logical variety M, if any finite subset of components of **M** is compatible, then **M** is compatible.

Corollary 1. For any classical deductive logical variety **M** with a countable number of components, if any finite subset of components of **M** is compatible, then **M** is compatible.

The compatibility of a logical variety means that it is possible to immerse all components of this variety into one calculus from the class **K** of logical calculi. Thus, the obtained results show that the possibility of logic system immersion into one

calculus is undecidable for a finite number of logics (Theorem 1), while for an infinite number of logics, the decidability problem is reducible to the finite case (Theorem 2) and thus, is undecidable in general.

4 Knowledge Representation and Logical Inference in the Legal Domain

Although each domain of knowledge is more or less affected by the problem of inconsistent knowledge, this issue is particularly intense in the domain of legal knowledge, since it consists of the rules and procedures used to describe and solve legal conflicts, which presupposes contradictory and hence inconsistent perspectives. Human processors of legal knowledge follow formal and informal problem-solving methods in order to reduce the number of legal perspectives and eventually to decide, temporally and within a specific context, on a common perspective. The formal methods are based on universal properties of formally valid legal argument. The informal methods are based on legal heuristics consisting in tentative legal decision principles. The first category can be formalized by logic because it applies peremptorily to all legal perspectives. The second category cannot be fully formalized by logic because, although it is commonly applicable, it can always be refuted by a contradictory decision principle and even by the mere existence of an underlying contradictory argument.

Legal opinions range from informal to formal. On the informal side we find moral principles, social scripts, protocols, (technical) instructions, rules of thumb, rules of play etc. On the formal side we find legislation, legal principles, jurisprudence, policy rules etc. Legal opinions can be of a general (uninstantiated) and of a specific (instantiated) character. Legal procedures consist of (1) procedures to list all the normative opinions about a given situation that can be inferred from the given situation combined with the set of pre-existing normative opinions of the parties concerned and (2) procedures to reduce the number of normative opinions about the given situation to a (local and temporal) common opinion for (not necessarily *of*) the parties concerned. Both procedures involve *legal reasoning*. The second procedure also involves *legal decision-making*. Legal reasoning in the first class of procedures is concerned with the inference of normative opinions about the given situation (the object level). Legal reasoning in the second class of procedures is concerned with the inference of normative opinions about the reduction of normative opinions (the metalevel, e.g. “the judge is obliged to decide for a legally valid opinion”). However, the decision principles applied at this level also represent opinions, so there is no exhaustive or non-contradictory set of decisive opinions at this level either.

These properties of legal knowledge should be taken into consideration in order to be able to develop a tenable computational model of the application of legal knowledge. To achieve this aim logical varieties are used as a foundation for the Logic of Reasonable Inferences (LRI), which provides means for legal knowledge representation and models legal reasoning, using the language of classical first-order predicate calculus, as this language seems powerful enough to express legal rules and factual situations without losing any relevant information (de Vey Mestdagh, [18]). Being a logical variety the LRI allows for the representation of internally consistent, but mutually inconsistent, alternative opinions in one system, thus preserving all legal

perspectives. Each of these opinions (called a *position* below) from the logical variety of LRI is represented by a single classical calculus.

LRI uses the language $\mathcal{L}$ of the first order predicate calculus and is constructed as a logical variety **V** in $\mathcal{L}$. Expressions from $\mathcal{L}$ are called beliefs and components of **V** are called positions or convictions. The common part of all components (positions) is a set A of well formed formulas (wwfs) of $\mathcal{L}$ which are called axioms of **V**.

Definition 4.1. A *reasonable base* in $\mathcal{L}$ is a pair Δ defined as $\Delta = (A, H)$ where H is sets of *wffs* in $\mathcal{L}$, which are called (tentative) *assumptions* (*hypotheses* or *beliefs*).

The *assumptions* can model the rules of law that may or may not be applied in a given factual situation to derive a conclusion and contain all normative or subjective classifications of the factual situation. The *axioms* are intended to be valid in every justification and thus, restrict the number of possible justifications. These axioms represent the ascertained facts and previously ascertained conclusions (the permanent database in any implementation).

Definition 4.2. A *position* (or *conviction*) ϕ within a domain of rules $\Delta = (A, H)$ is a consistent set of wffs defined as $\phi = A \cup H'$ where $H' \subseteq H$.

Definition 4.3. A position (conviction) ϕ within a domain of rules $\Delta = (A, H)$ is called *logically closed* if it is a predicate calculus.

Thus, a position is a set of rules taken from the domain of rules and represents a conviction of an individual or a group of people. Note that all positions should at least contain all axioms of the domain of rules and each position is consistent by definition. This shows that all logically closed positions form a logical variety in which all intersection are equal to A.

Let Δ be a domain of rules. Define a new semantic derivability-relation $\models_r$ as $\Delta \models_r \varphi$ *iff* there exists a position ϕ within Δ which satisfies $\phi \models \varphi$ where $\models$ is the normal predicate calculus semantic derivability relation. If $\Delta \models_r \varphi$ holds, φ is said to be a *reasonable inference* from the domain of rules Δ. This the exact form of inference in logical varieties.

We can paraphrase this definition by stating that a *wff* can reasonably be inferred from an inconsistent set of *wff* iff it is derivable (in the normal predicate calculus sense) from a consistent subset of this set which contains at least the axioms. Note that if a domain of rules $\Delta = (A, H)$ is consistent (i.e. if $A \cup H = \Gamma$ is consistent), then $\Delta \models_r \varphi \iff \Gamma \models \varphi$ behaves exactly like $\models$ when applied to consistent theories.

Definition 4.4. A *justification* for a conclusion φ derived from a domain of rules Δ is a minimal position (with respect to set-inclusion) J within Δ such that $J \models \varphi$.

This definition is based on the intuitive concept of a set of rules and statements about the factual situation used to draw the conclusion. Note that a justification needs not be unique but it is always consistent, thus, satisfying our constraints.

Definition 4.5. A *context* in Δ is the union of n simultaneously derived conclusions ψ_i and their justifications J_i derived from Δ, i.e. a context is the set of tuples $\{ (\psi_i , J_i) \mid 1 \leq i \leq n \}$.

Definition 4.6. A context in Δ is called *consistent* if the justifications J_i derived from Δ satisfy the following condition:

$$\text{The union } \bigcup_{i=1}^{n} J_i \text{ is consistent}$$

This guarantees that simultaneously derived conclusions are not based on mutually inconsistent positions.

Definition 4.7. A *reasonable theory* Th Δ with a base $\Delta = (A, H)$ is the set of all wffs deducible in LRI from Δ, i.e., Th $\Delta = \{ \varphi \in \mathcal{L}$; there is a position $\phi \in \Delta$, such that $\phi \vDash \varphi \}$.

The construction of a reasonable theory shows that any reasonable theory Th Δ is a deductive logical variety V that has the form

$$V = \{ C_i \, ; i \in I \text{ and there is a position } \phi, \text{ such that } C_i = (\phi, d, T_\phi) \}$$

where d is the set of all deduction rules of classical the first-order predicate calculus and T_ϕ is the set of all formulas deducible from the position ϕ by rules from d. V is a first order predicate variety.

An earlier version of the LRI (de Vey Mestdagh, [18]) has been used to specify a knowledge based system shell, Argumentator (de Vey Mestdagh, [19]). This shell has consequently been used to acquire and represent legal knowledge. The resulting legal knowledge based systems have been successfully used to test the empirical validity of the theory about legal reasoning and decision making modeled by the LRI (de Vey Mestdagh, [19]).

5 Conclusion

Logical varieties, quasi-varieties and prevarieties eliminate certain limitations of conventional logical systems based on logical calculi. In comparison with paraconsistent logics, they allow utilization of sufficiently powerful means of logical inference, for example, deductive rules of the classical predicate calculus. In comparison with non-monotonic logics, logical varieties, quasi-varieties and prevarieties provide tools for preserving all points of view, approaches and positions even when some of them taken together lead to contradiction. Due to their flexibility, logical varieties, quasi-varieties and prevarieties allow treating any form of logical contradictions in a rigorous and consistent way.

The mathematical results of this paper explore the problem of applicability of the classical logic to knowledge systems. Indeed, it is possible to represent any logical system expressed in a logical language as a classical logical variety, prevariety or, at least, quasi-variety M. Then this system is embeddable in a classical logic if and only if all components of this variety, prevariety or quasi-variety M are compatible. However, our results show that this important problem is undecidable in a general case.

The LRI is a form of a logical variety that is aimed at building legal decision support systems and legal expert systems, which can be used by judges, jurors, lawyers, detectives, and attorneys. The legal system is characterized by different often contradictory opinions and the need to decide temporally within a certain legal context but to preserve the different points of view. The results within the legal domain can of course be generalized to any domain with similar characteristics. The LRI has been successfully implemented and applied in a legal knowledge based system to accommodate for these characteristics and has been empirically verified, by comparing the decisions made by the system with decisions made by human decision makers.

Generally, Legal Decision Support System based on logical varieties can help jurors and judges to find if witnesses are consistent in their depositions, if statements of different witnesses are compatible, if versions of persecution and defenders are consistent, and which of these versions is more grounded. It can help a judge to find what laws and/or what precedence cases are more compatible with the given case. Finally it can help detectives and attorneys to find which conjectures are compatible with evidence and with one another and which of these conjectures are more grounded.

References

1. Abadi, A., Rabinovich, A., Sagiv, M.: Decidable fragments of many-sorted logic. J. Symb. Comput. 45(2), 153–172 (2010)
2. Amir, E.: Dividing and Conquering Logic, Ph.D. Thesis, Stanford University, Computer Science Department (2002)
3. Amir, E., McIlraith, S.: Partition-Based Logical Reasoning for First-Order and Propositional Theories. Artificial Intelligence 162(1/2), 49–88 (2005)
4. Barwise, J., Seligman, J.: Information Flow: The Logic of Distributed Systems. Cambridge Tracts in Theoretical Computer Science, vol. 44. Cambridge University Press, Cambridge (1997)
5. Bertossi, L., Hunter, A., Schaub, T. (eds.): Inconsistency Tolerance. LNCS, vol. 3300. Springer, Heidelberg (2005)
6. Brown, B., Priest, G.: Chunk and Permeate: A Paraconsistent Inference Strategy, part I: The Infinitesimal Calculus. The Journal of Philosophical Logic 33, 379–388 (2004)
7. Burgin, M.: Logical Methods in Artificial Intelligent Systems. Vestnik of the Computer Society (2), 66–78 (1991) (in Russian)
8. Burgin, M.: Logical Tools for Program Integration and Interoperability. In: Proceedings of the IASTED International Conference on Software Engineering and Applications, pp. 743–748. MIT, Cambridge (2004)
9. Delgrande, J.P., Mylopoulos, J.: Knowledge Representation: Features of Knowledge. In: Bibel, W., Jorrand, P. (eds.) Fundamentals of Artificial Intelligence. LNCS, vol. 232, pp. 3–38. Springer, Heidelberg (1986)
10. DeWitt, B.S.: The Many-Universes Interpretation of Quantum Mechanics. In: Foundations of Quantum Mechanics, pp. 167–218. Academic Press, New York (1971)
11. Everett, H.: 'Relative State' Formulation of Quantum Mechanics. Reviews of Modern Physics 29, 454–462 (1957)
12. Gabbay, D.: Fibring Logics. Clarendon Press, Oxford (1999)

13. Hall Jr., M.: The theory of Groups. The Macmillan Company, New York (1959)
14. Jaśkowski, S.: Rachunek zdań dla systemów dedukcyjnych sprzecznych. Studia Societatis Scientiarun Torunesis (Sectio A) 1(5), 55–77 (1948)
15. MacCartney, B., McIlraith, S.A., Amir, A., Uribe, T.: Practical Partition-Based Theorem Proving for Large Knowledge Bases. In: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI 2003), pp. 89–96 (2003)
16. McIlraith, S., Amir, E.: Theorem proving with structured theories. In: Proceedings of the 17th Intl' Joint Conference on Artificial Intelligence (IJCAI 2001), pp. 624–631 (2001)
17. Meinke, K., Tucker, J.V. (eds.): Many-sorted logic and its applications. John Wiley & Sons, Inc., New York (1993)
18. Vey Mestdagh, C.N.J., de Verwaard, W., Hoepman, J.H.: The Logic of Reasonable Inferences. In: Breuker, J.A., Mulder, R.V., de Hage, J.C. (eds.) Proc. 4th Annual JURIX Conference on Legal Knowledge Based Systems, Model-Based Legal Reasoning, Vermande, Lelystad, pp. 60–76 (1991)
19. de Vey Mestdagh, C.N.J.: Legal Expert Systems. Experts or Expedients? The Representation of Legal Knowledge in an Expert System for Environmental Permit Law. In: Ciampi, C., Marinai, E. (eds.) The Law in the Information Society, Conference Proceedings on CD-Rom, Firenze, p. 8 (1998)
20. Minsky, M.: A Framework for Representing Knowledge. MIT, Cambridge (1974)
21. Minsky, M.: Society of Mind: A Response to Four Reviews. Artificial Intelligence 48, 371–396 (1991)
22. Minsky, M.: Conscious Machines. In: Machinery of Consciousness, 75th Anniversary Symposium on Science in Society, National Research Council of Canada (1991)
23. Nguen, N.T.: Inconsistency of knowledge and collective intelligence. Cybernetics and Systems 39(6), 542–562 (2008)
24. Partridge, D., Wilks, Y.: The Foundations of Artificial Intelligence. Cambridge University Press, Cambridge (1990)
25. Priest, G., Routley, R., Norman, J. (eds.): Paraconsistent Logic: Essays on the Inconsistent. Philosophia Verlag, München (1989)
26. Rescher, N., Manor, R.: On inference from inconsistent premisses. Theory and Decision 1(2), 179–217 (1970)
27. Ross, T.J.: Fuzzy Logic with Engineering Applications. McGraw-Hill P. C, New York (1994)
28. Routley, R., Plumwood, V., Meyer, R.K., Brady, R.T.: Relevant Logics and Their Rivals, Atascadero, Ridgeview, CA (1982)
29. Smolin, L.: The Bekenstein bound, topological quantum field theory and pluralistic quantum field theory, Penn State preprint CGPG-95/8-7; Los Alamos Archives preprint in physics, gr-qc/9508064 (1995), electronic edition `http://arXiv.org`
30. VanPool, T.L., VanPool, C.S. (eds.): Essential Tensions in Archaeological Method and Theory. University of Utah Press, Salt Lake City (2003)
31. Weinzierl, A.: Comparing Inconsistency Resolutions in Multi-Context Systems. In: Slavkovik, M. (ed.) Student Session of the European Summer School for Logic, Language, and Information, pp. 17–24 (2010)

Contextual Ontology Module Learning from Web Snippets and Past User Queries

Nesrine Ben Mustapha[1,2], Marie-Aude Aufaure[1],
Hajer Baazaoui Zghal[2], and Henda Ben Ghezala[2]

[1] Ecole Centrale Paris, MAS Laboratory, Business Intelligence Team, France
{nesrine.ben-mustapha,marie-aude.aufaure}@ecp.fr
[2] Laboratory RIADI, ENSI, La Manouba, Tunisia
hajer.baazaouizghal@riadi.rnu.tn

Abstract. In this paper, we focus on modularization aspects for query reformulation in ontology-based question answering on the Web. The main objective is to automatically learn ontology modules that cover search terms of the user. Indeed, the main problem is that current approaches of ontology modularization consider only the input existant ontologies, instead of underlying semantics found in texts. This work proposes an approach of contextual ontology module learning covering particular search terms by analyzing past user queries and snippets provided by search engines. The obtained contextual modules will be used for query reformulation. The proposal has been evaluated on the ground of semantic cotopy measure of discovered ontology modules, relevance of search results.

Keywords: Ontology, modular ontology, knowledge, ontology learning.

1 Introduction

With the increasing availability of ontologies on the Web, modularity principle has become an important issue to overcome scalability problems over ontology-based systems. Ontology module extraction (OME) approaches consist in reducing ontology to an ontology fragment covering a particular vocabulary. The extracted ontology modules are used for knowledge selection or reuse. Generally, the input of those approaches [1] consists in large ontologies and the output is a set of independent modules. Current approaches of modularization basically rely on static existant ontologies, which can be inconsistent to cover users need. Indeed, users search interest and also domain knowledge evolve with new discoveries and usages. As a consequence of this continuing evolution, automatic methods for ontology construction are required. For the best of our knowledge, machine learning strategies have not yet been explored for ontology modularization as mentioned in [1]. Unlike many previous approaches of modularization, the proposed method has been designed in an automatic and domain-independent way. Web information distributions were employed to assess the reliability of the extracted knowledge. We propose then a new approach of ontology module extraction from web snippets, and user's context (past user queries and selected documents).

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 538–547, 2011.

This paper is organized as following. In section 2, an overview of related works on ontology module extraction is presented. Section 3 describes the proposed approach of contextual *OM* extraction. In section 4, we describe an evaluation on the ground of two criteria which are the comparison of discovered ontology modules with an upper-level ontology MESH and the impact of *OM* extraction on the relevance and the ranking quality of search results. Finally, we conclude and discuss directions for future research.

2 Related Works

Ontology module extraction consists in reducing ontology to an ontology fragment that covers a particular vocabulary. In [5], the proposed approach called "ontology segmentation" takes one or several classes of the ontology as an input. It applies a generic algorithm to include all related classes that participate in the definition of the input classes, on the basis of class subsumption and OWL restrictions. Noy and Mussen [3] define a novel traversal view extraction technique for module extraction. Starting from one class of the considered ontology, relations of this class are recursively traversed to include related entities as in [5]. However, this technique is not automatic and takes into account the user involvement in selecting the relations to be traversed and associating to each of them a level of recursion, at which the algorithm should stop "traversing" relations. Besides, the proposed approach in in [4] is composed of: (1) the selection of relevant ontologies, (2) the modularization of the selected ontologies, (3) the merger of the relevant ontology modules in the case when the query terms are covered by several different ontologies. The input of the ontology modularization approach is made up with ontology and a set of terms that should be covered by the smallest part of the ontology. Unlike the algorithm in [5], all the superconcepts of a selected concept are not necessarily included (only the ones that are directly related to concepts of the module, i.e. the most specific common concepts).

In the approaches mentioned above, two main limits are noticed. First, the existing approaches of ontology modularization rely on static ontologies that can be inconsistent to cover basically user's need for information search on the Web. Second, modularization algorithms consider mainly the structure of the input ontology, instead of semantics or context. Consequently, we need semantics-based criteria so as to determine the border of ontology modules. Moreover, the contextuality of the module will considerably depend on the semantic covertness of the original input ontologies. However, ontologies on the web are not sufficiently consistent and contextualized to cover specific domain knowledge. Therefore, our proposed strategy should also differ from previously discussed strategies. Obtained modules have to be based on observing relevant interactions for knowledge selection, not on a human or human-driven dedicated specification, nor on the structural properties of the ontology (as in traversal view strategies). Ontology learning (OL) techniques could be a way to overcome these limits. In fact, for the best of our knowledge, unsupervised machine learning strategies have not yet been explored for ontology modularization as mentioned in[1].

Ontology learning (OL) aims at building ontologies from knowledge sources using a set of machine learning techniques and knowledge acquisition methods. OL from texts has been widely used in the knowledge engineering community. By applying a set of text mining techniques, a granular ontology is enriched with concepts and relationships. in this paper, we focus mainly on two categories of unsupervised techniques that don't need any background knowledge: Lexico-syntactic patterns (LSP)and distributional measures. In the last decade, with the enormous growth of Web information, the Web has become an important source of information for knowledge acquisition. Its main advantages are its huge size and its large degree of heterogeneity. OL from Web documents requires the same techniques as those used for ontology extraction from texts. A study of several types of available Web search engines and how they can be used to assist the learning process (searching Web resources and computing IR measures) were explored in [6].

The main challenge of the present work is to use OL for OM construction and its integration in the search process. This work is part of the generic approach. It aims to develop its modular semantic layer from the associations between queries and documents results in order to improve the contextualization of user's goal search and consequently, the answers' relevance of semantic search[7]. In the next section, our proposed approach will be detailed.

3 Contextual Ontology Module Learning Approach

In this section, we describe an approach of OM building in an automatic and domain-independent way, using past users queries and resulted snippets (returned by web search engines). Note that the term "snippet" is used here to denote a fragment of a Web page returned by remote search engines (such as GOOGLE or YAHOO) and summarizing the context of searched keywords. Our underlying hypothesis is that an OM is an ontology fragment that represents a question on specific domain knowledge. This OM can be used to annotate documents related to the specific knowledge component.

The main steps of the proposal are the following: question analysis, candidate answers extraction, context map extraction and contextual module representation using attributed graph. The input of the proposed approach is made up with questions and results pairs ($URls$) related to a specific topic. First, each question is analyzed by identifying the answers' patterns to be used in the next step. Second, these patterns are employed to reformulate queries in order to collect relevant snippets provided by a web search engine. Next, a concept network called context map is extracted from the obtained textual snippets by applying ontology learning techniques (LSP and web co-occurrence scores (WCS)). A top-level ontology (such as Mesh, Sensus) describing very general concepts that are the same across all knowledge domains can be used to identify question concepts and import related concepts and relations. The obtained context map acts as a skeleton on which the OM is built. it is represented using attributed graph.

3.1 Question Analysis

This step aims to obtain a reformulated question (RQ), taking into account answers patterns and users context (selected results from users). In order to conceptualize each question, a repository of predefined question patterns (RQP) is designed. Each query type (what, where, who) is associated to a set of answers patterns. According to the question pattern (QP) of the submitted query, answers pattern (AP) are selected and instantiated with question terms to search answers passage from web snippets. For instance, the query (*"what is a BMI?*) is a typical question of the following $QP1$ ("$what < be >< name >$ "). "$< name >< be >< Answer >$" is an AP assigned to this question pattern. A new RQ is represented by the following phrase: *"BMI is "*. However, query terms can be polysemic. Based on the same premise adopted in corpus-based approaches, we consider that the context can defined by a set of terms which co-occur frequently with query terms in the selected results". Those ones whose frequency is superior to a threshold are selected as belonging to the semantic signature (called also topic signature) of term "t".

In most of corpus-based approaches, the context of a word is usually defined as the word around them within certain of window of which size is usually set as two. Therefore, the analyzed query is extended with two terms that have the highest co-occurrence score (WCS)from the topic signature (TS). This score is described in section 3.2. The contextual query reformulation aims to eliminate the risk of collecting irrelevant snippets to the right sense of terms. For instance, the two high-ranked terms of the TS extracted from user's results related to the question "*What is BMI*" are "*height*" and "*weight*". The new RQ is the following query *"BMI is" AND height AND weight"*.

3.2 Candidate Answers Extraction

To extract candidate Answers (CA) from snippets, the RQ is submitted to the web search engine in order to collect the first β snippets. Using AP, words matching the tag $< answer >$ are selected as CA element. For instance, the following sentence *"BMI is a measure of body fat"* is tagged according to the pattern AP and the result is "BMI $< NAME >$, IS-A $< be >$ the measure of body fat $< answer >$". Then, the following term "measure of body fat" refers to a candidate answer. The corresponding CA values are selected based on WCS 1(superior to a threshold Tsα) (which can include the terms "*measure*", "*calculation*" and "*formula*").

$$score(\text{problem}, \text{choice}) = \frac{hits(problemAndchoice)}{hits(choice)} \tag{1}$$

These CA should be ranked and selected according to statistical assessment based on Web-based semantic similarity. Indeed, we used the scores below based on the measure proposed by Turney [2] to evaluate the co-occurrence score (WCS) between initial word problem (terms included in the Reformulated query) and related term candidate choice (candidate answer) by the following formula:

For the rest of this paper, we use the notation *hits(a)* to denote the number of search results that contain the query "a" made to a search engine. The concept candidate whose score is superior to an associated threshold Tsα is selected to be used in the context map construction. The threshold Tsα is based on the average of similarity values between terms of *TS* (that denotes the topic signature of query terms and obtained in the question analysis step) and the reformulated query *RQ*. For instance, the corresponding candidate values are selected based on web co-occurrence measures (superior to an threshold Tsα = 0.03.25) as follows: Score $(BMI, measure) = 0.28$; $Score(BMI, calculation) = 0.0373$; $Score(BMI, formula) = 0.37$.

The extracted answers candidates are used to construct new queries in order to collect new collection of snippets. New queries are automatically submitted to the search engine by extending the previous query as the following: BMI is a $< candidate_value >$, in order to collect the first β snippets. Then, a context map can be constructed from this collection and converted to an ontology fragment, as described in the following subsections.

3.3 Context Map Construction

The aim of this task is to construct context map that represents semantically the possible answers to user's information need, regarding its context. Note that the term "context map" refers to a network of domain terms and relationships extracted from textual passages. This task is mainly based on four operators: Concept identification (CIP), relation operators (RIP), Relation label and Concept learning (RLCP) operator and Concept and relation selection (CRS). The construction of context Map as shown by the figure 1, works as follows.

Concept Identification (CIP) and Relation Operators (RIP). Domain concepts are identified by CIP by using particular typed dependencies which are detected by a syntactic parser. For each type of dependency, a set of transformation is defined, in order to identify domain concepts. Those rules are based on lexico-syntactic patterns. A subset of the following rules is detailed in [11]. The parser provides grammatically typed dependency networks. Then, these networks are mined by the RIP in order to transform automatically the grammatical representation into semantic ones. The semantic representations are then used to create the context map.

Relation Label and Concept Learning (RLCP). The (RLCP) operator use the constructed context map after the identification of basic domain concept and domain relations in order to discover others possible label of relations and concepts using snippets and lexico-syntatctic patterns. For example, to discover new label of the relation "IS-A" that relates the concept "*BMI*" and "*measure*", the following query: "*BMI* * *measure*" is made to a search engine in order to import snippets that contain sentences regrouping this two terms in order to extract possible verbs relating them. According to the provided snippets, possible relation labels include the following verbs "*provide*", "*revert to*", "*give*", "*offer*" (figure 1).

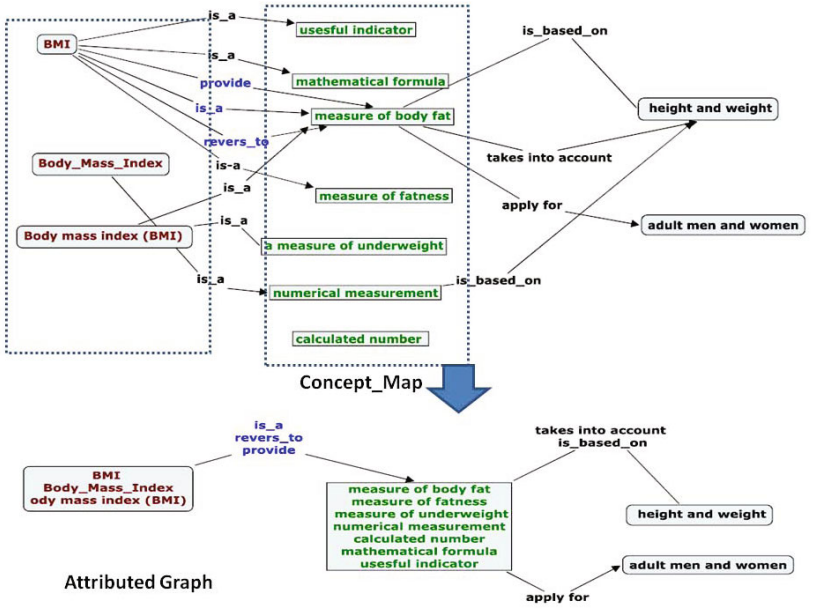


Fig. 1. Example of a context Map and extracted attributed graph related to BMI topic

On the other hand, new discovered labels are considered as new patterns for candidate concepts that can be related to domain concepts by means of these labels. Therefore, new queries are made to the search engine (such as "*BMI provides* *"), which provide relevant sentences containing these patterns. Then, new domain candidates are discovered.

Concept and Relation Selection (CRS). Applying the mentioned techniques does not mean that the extracted knowledge is enough strong to grant the definition of the module. This operators need to select the extractions that are sufficiently reliable. To perform this selection, we introduce a Web-based statistical analysis relying on co-occurrence measures computed directly from search engine. Co-occurrence measures are based on distributional hypothesis claiming that words that occur in the same context tend to have similar meanings. Several scores have been proposed in the past to compute Web-scale statistics, adapting the notion of co-occurrence and mutual information (computed as the probability of joint appearance of concepts in a corpus). Discovered candidate concept is represented by a node and it is weighted using the score presented by the following formula, taking into account the appropriate pattern:

$$\text{Score}_{\text{pattern}},\ (\text{choice}) = \underset{i=1}{\overset{Patterns}{Max}} \left(\frac{hits(\text{"}pattern_i(\text{"}concept\text{"},\text{"}candidate\text{"})}{hits(\text{"}candidate\text{"})}\right)$$

This formula computes the maximum probability of finding any no taxonomic relation involving the candidate concept and the domain concept in the scope of

web document containing that candidate. If this score is remarkably low than a threshold, the discovered concept or relation is rejected.

3.4 Ontology Module Representation

This step has as input the context map and as output an attributed graph. Since the OMs are supposed to be extracted from unstructured text, the discovered concepts and relations are not validated at one step. For this reason, we choose to rely on attributed graph because it is a powerfully enough to represent *OM* written in RDF, OWL or DAML+OIL. Besides, attributed graph is the model implemented in the ACG library, for graph transformation. Details about attributed graph are described in [9]. An attributed graph representation of the module $\mathcal{AG}_{\mathcal{M}}$ is a pair $(\mathcal{N}_G, \mathcal{E}_G)$, where $\mathcal{N}_G$ is a set of attributed nodes and $\mathcal{E}_G$ is a set of attributed edges.

An **attributed node** $\mathcal{N}_G = (T_N, AV_N)$ has a type T_N and a set of attribute values AV_N where Tn is the set of terms referring (eg. BMI, Body Mass Index) to a concept C and AV_N is the set of score's (WOC)values assigned to each of the terms belonging to CN.

An **attributed edge** $\mathcal{EG}_M = (T_E, \mathcal{RN}, AV_E, O_E, D_E)$ has a type T_E, a set of attribute values AV_E, an origin node O_E and a destination node D_E, where T_E denotes the type of a relation (hyponymy, meronymy, possession, verb label, etc.) and $\mathcal{RN}$ is a set of terms referring to the relation ($\mathcal{R}$).

An **attribute value** AV_E is a pair $(\mathcal{RN}, score)$ associating score's value to a term of ($\mathcal{RN}$). The figure 1 presents a reused module related to the disease subtopic.

4 Experimentation: Modular Ontology Learning for Semantic Search

In order to evaluate the feasibility of the proposed approach, we have tried to compare our approach with different related works on the modularization approaches. However, it has been difficult to compare the proposal with OM extraction approaches since the features and the usage of each input of those approaches are different. Therefore, it seems more logical to evaluate the present work according to OL creteria. We add that the evaluation of the proposed approach is based on two main criteria which are: (1) the comparison of OM learning process (figure 2) and (2) the impact of *OM* learning on the relevance of search results (figure 3).

In figure 2, two ontologies are compared. One one hand, we use a Taxonomic Precision (TP) which is a similarity measure based on the notion of semantic cotopy *sc*. It is recently presented and analysed in [10]. The reason to choose this measure was to take advantage of its ability to compare ontologies as whole structures. The values of TP are from the range [0, 1]. We use the Mesh Ontology (MSO) as an ontology reference. 80 queries on the topic of animal diseases were collected manually by using 80 concepts of MSO. 80 ontology modules which

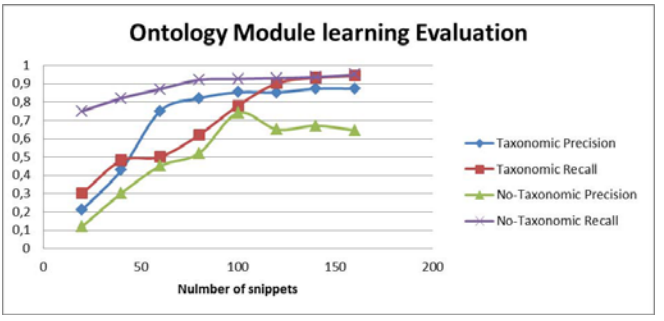


Fig. 2. ontology module learning evaluation

make up a large ontology (RO) were constructed according to the proposed approach to be compared with (MSO).Semantic Cotopy sc(c,O) of a concept c from ontology O is a set containing c and all super and sub-concepts of c in O, excluding the concept root of (O). Then, TP(c, RO, MSO) of concept c and two ontologies RO and MSO where $c \in RO$ and $c \in MSO$ is defined as follows:

$$TP(\mathrm{C}, \mathrm{RO}, \mathrm{MSO}) = \frac{sc(\mathrm{C}, \mathrm{MSO}) \bigcap sc(\mathrm{C}, \mathrm{RO})}{sc(\mathrm{C}, \mathrm{RO})} \tag{2}$$

A Taxonomic Recall (TR) can be assessed as follows:

$$TP(\mathrm{C}, \mathrm{RO}, \mathrm{MSO}) = \frac{sc(\mathrm{C}, \mathrm{MSO}) \bigcap sc(\mathrm{C}, \mathrm{RO})}{sc(\mathrm{C}, \mathrm{MSO})} \tag{3}$$

Therefore, the global TP and TR are computed respectively by the following formulas:

$$GTP(\mathrm{RO}, \mathrm{MSO}) = \frac{1}{|\mathrm{RO}|} \sum_{\mathrm{c} \in \mathrm{RO}} TP(\mathrm{c}, \mathrm{RO}, \mathrm{MSO}) \tag{4}$$

$$GTR(\mathrm{RO}, \mathrm{MSO}) = \frac{1}{|\mathrm{MSO}|} \times \sum_{\mathrm{c} \in \mathrm{RO}} TP(\mathrm{c}, \mathrm{RO}, \mathrm{MSO}) \tag{5}$$

No taxonomic precision and recall are calculated according to the same formula by substituting the sc(c, O) by the set containing concept c and all concepts related to c by a no taxonomic relationship. The figure 2 shows the evolution of the precision of taxonomic and non taxonomic structure according to the number of snippets used in the ontology module leaning. On the other hand, in order to evaluate the approach presented in this paper, the impact of the use of OMs during query reformulation is also experimented. First, we have computed the precision of results retrieved by means of query reformulation using discovered modules. Evaluation results contained in Figure 3 represent the obtained precision according to the number of retrieved documents (from 4 to 100). The first scenario represents the initial search, which is a keyword search on Yahoo. The second scenario represents the situation where there are similar cases in the

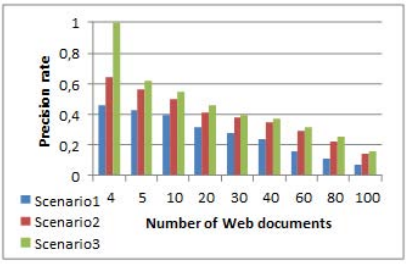


Fig. 3. Evaluation of result precision

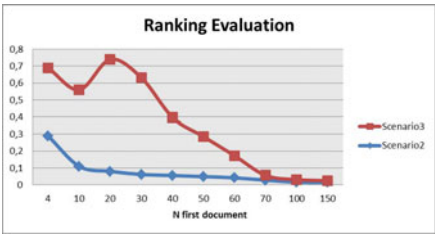


Fig. 4. Evaluation of results ranking

database. The search is based on using WorldNet to add synonyms. The third scenario represents the search for information based on the learned OMs by using 100 snippets for ontology module learning. The query reformulation is based on answers pattern extracted from constructed OMs. We have observe a significant improvement of the relevance of the retrieved information according to the amount of knowledge considered during query reformulation and *OM* creation. We have also noticed that this improvement is maintained as the number of documents increases, even though the quality of the retrieved document set decreases due to the higher amount of noisy and non-related documents retrieved. The results have revealed that: (1) Their accuracy was significantly improved by using modular ontologies; (2) Strongly, discovered ontology module are important to better contextualize users searches and (3) The relevance of documents are not based on the terms frequency but on the semantic relatedness between terms.

Second, in order to evaluate the ranking quality of results according to the formulated query, we used the well-known Normalized Discounted Cumulative Gain measure. While evaluating a ranking list, NDCG is computed according to the original paper [8], as follows:$NDGC(n) = Z_n \sum_{j=1}^{n} \frac{2^{r(j)}-1}{log(j+1)}$, Where $r(j)$ is the rating of the j-th document in the list, and the normalization constant Z_n is chosen so that the perfect list gets a NDCG score of 1. Figure 4 shows the evaluation results measured by the NDCG for the two scenarios previously described. The X-axis refers to the Web page rank. Again, it is shown that reformulated queries using pattern answers (extracted from obtained ontology modules) have contributed to improve significantly the document raking.

5 Conclusion and Future Work

In this paper, we have proposed a new approach of Ontology Module extraction from web snippets, and user's feedback (past user queries and selected documents). Unlike many previous modularization approaches, the originality of this work is that it has been designed in an automatic and domain-independent way, exploiting unsupervised techniques and the web as a large scale learning source.

The contribution resides in the following techniques: Web-based co-occurrence measures for the assessment of extracted knowledge (concepts and relationships) and Unsupervised method for context map construction and Attributed graph representation for a multi-label representation of ontology module. The evaluation of the proposal is based on two criteria which are the comparison of OM extraction process and the impact of module-based query reformulation on the relevance of search results. The evaluation of question answering system has revealed that the accuracy of the results was significantly improved by using modular ontologies. Our ongoing work aims at exploring ontology module construction for social search systems.

References

1. Stuckenschmidt, H., Parent, C., Spaccapietra, S.: Modular Ontologies: Concepts, Theories and Techniques for Knowledge Modularization. Springer, Heidelberg (2009)
2. Turney, P.: Mining the web for synonyms: PMI-I versus LSA on TOA. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 491–502. Springer, Heidelberg (2001)
3. Noy, N., Musen, M.: Specifying Ontology Views by Traversal. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 713–725. Springer, Heidelberg (2004)
4. D'Aquin, M., Sabou, M., Motta, E.: Modularization, a key for the Dynamic Selection of Relevant Knowledge Components. In: Proc. of the ISWC 2006 Workshop on Modular Ontologies (2006)
5. Seidenberg, J., Rector, A.: Web Ontology Segmentation: Analysis, Classification and Use. In: Proc. of the World Wide Web Conference, WWW (2006)
6. Sanchez, D., Moreno, A.: Learning non-taxonomic relationships from web documents for domain ontology construction. DKE 64(3), 600–623 (2008)
7. Elloumi, M., Ben-Mustapha, N., Baazaoui, H., Moreno, A., Sanchez, D.: Evolutive Content-Based Search System. In: KDIR (2010)
8. Jarvelin, K., Kekalainen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems 20(4), 422–446 (2002)
9. Ehrig, H., Ehrig, K., Prange, U., Taentzer, G.: Fundamental theory for typed attributed graphs and graph transformation based on adhesive hlr categories. Fundam. Inf. 31, 31–61 (2006)
10. Maedche, A., Staab, S.: Measuring similarity between ontologies. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 251–263. Springer, Heidelberg (2002)
11. Schutz, A., Buitelaar, P.: Relext: A tool for relation extraction from text in ontology extension. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 593–606. Springer, Heidelberg (2005)

Inconsistency-Tolerant Integrity Checking Based on Inconsistency Metrics

Hendrik Decker*

Instituto Tecnológico de Informática, Valencia, Spain

Abstract. A large diversity of different approaches to integrity checking has been proposed in the literature. We present a generic approach to integrity checking that is based on inconsistency metrics. It subsumes many known approaches to integrity checking by a uniform abstraction. As opposed to traditional approaches to integrity checking, it permits the tolerance of extant (i.e., surviving) inconsistency. Inconsistency metrics have been widely studied in the literature, but, so far, their applicability to integrity checking has largely remained unaddressed.

1 Introduction: Previous Work

The quality of the knowledge provided by database can be maintained by enforcing semantic integrity constraints that are imposed on the stored data. Integrity checking is an essential building block of integrity enforcement. For each transaction, it examines whether the transaction's updates would violate any constraint. In that case, at least a warning is issued, and typically, the transaction is either aborted or repaired, so that integrity violation is alleviated.

A large variety of different approaches to integrity checking can be found in the literature. Prominent ones are simplification-based [15,4], others may use, e.g., theorem proving [19] [2], compilation or partial evaluation [20] [13], query containment [12], incremental view materialization [17], special purpose languages [3] [18], etc. In this paper, we propose a generic manner of describing integrity checks in database systems, based on inconsistency metrics. It uniformly subsumes most of the methods for integrity checking mentioned so far, and also inconsistency-tolerant approaches that have been gaining attention recently [8].

Already the description of integrity checking in [8] is generic. It abstracts away from the diversity of technical details by which the many methods can be distinguished. In [8], every method is characterized merely by its i/o behaviour. The input always is a database state D, an integrity theory IC and an update U. The latter requests a transformation of D to a successor state D^U, obtained by deleting and/or inserting database clauses as specified in U. The output then is binary, e.g., either *ok*, by which the update request is accepted, so that it can be executed and committed, or *ko*, meaning that the update should not be accepted because it would harm the integrity of the database as stated by IC.

* Partially supported by FEDER and the Spanish grants TIN2009-14460-C03 and TIN2010-17139.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 548–558, 2011.

Moreover, the description in [8] significantly generalizes the traditional notion of integrity checking. Traditionally, integrity preservation always has been supposed to be total, i.e., no constraint violation is permitted in any committed state. Rather than requiring total integrity, only those instances of constraints in IC that are satisfied by D are required to remain satisfied by D^U in [8]; extant integrity violations in D, which may continue to survive in D^U, are tolerated. Such tolerance is necessary in large knowledge bases: experience teaches that inconsistencies are practically unavoidable and likely to accumulate over time in systems that undergo frequent changes over an extended period of time.

Although the description of integrity checking as an i/o mapping in [8] is generic, its characterization of soundness and completeness is not. Rather, it is 'case-based', i.e., it only captures those methods that preserve the satisfaction of certain instances called 'cases' of constraints across updates. Case-based inconsistency tolerance does not tolerate, e.g., the elimination of a constraint violation that is considered 'hard', while that elimination may entail a comparatively lightweight violation of some other constraint that is considered 'soft'.

Example 1. Let $\leftarrow \mathit{high\text{-}risk}(x)$ and $\leftarrow \mathit{low\text{-}risk}(x)$ be a hard and, resp., a soft constraint in a database D, defined by $\mathit{high\text{-}risk}(x) \leftarrow p(x, y, z)$, $y{+}z > t$, $y{>}z$ and $\mathit{low\text{-}risk}(x) \leftarrow p(x, y, z)$, $y{+}z > t$, $z{\geq}y$. They characterize entities with identifier x in the relation p as being of high or, resp., low risk, depending on the values of their attributes y and z, where t be some constant threshold value. Now, suppose that $t = 10$, and $p(a, 8, 3)$ be the only row of p in D that violates integrity. Further, let U be an update that modifies the entity a by assigning 3 to y and 8 to z. Unfortunately, inconsistency-tolerant case-based integrity checking does not accept U since U introduces the soft integrity violation $\mathit{low\text{-}risk}(a)$, although U eliminates the hard violation $\mathit{high\text{-}risk}(a)$. Hence, it is desirable to have a metric by which the degree or amount of inconsistency before and after the update can be compared. Thus, updates which do not augment the level of inconsistency can be accepted, while those that increase it should be rejected.

Inconsistency-tolerant integrity checking based on inconsistency metrics was first addressed in [10]. However, 'measure-based' integrity checking is only seen as one of several classes of integrity checking in [10], not in terms of a generic concept, as in this paper. Moreover, only inconsistency-tolerant methods have been of interest in [10], while, in this paper, we characterize each method for integrity checking as metric-based, no matter if it is inconsistency-tolerant or not. In fact, even the generalized notion of repair checking [1] in [5] can be subsumed under our notion of metric-based integrity checking.

A number of inconsistency metrics for assessing the quality of databases, and for maintaining it by measure-based integrity checking, has been discussed in [9]. Most of them were based on 'cases', as mentioned above. In this paper, we introduce a new one, based on 'causes', which have been introduced in [6].

Section 2 outlines the formal background and framework of the paper. Section 3 axiomatizes inconsistency metrics and discusses some examples of such metrics. Section 4 presents and discusses the concept of metric-based integrity checking. Section 5 concludes the paper.

2 Background and Framework

The formal background of this paper, including notations and terminology, is datalog [16]. Some of it is recalled in 2.1 and 2.2. In 2.3, we recapitulate "causes" [6]. Informally, causes are minimal explanations of constraint violations. Similarly, causes are used in [6] to explain answers to queries. In this paper, causes are used for two purposes: firstly, in 3.2, for measuring the inconsistency, i.e., lack of quality, in databases; secondly, in Section 4, as an example of metric-based inconsistency-tolerant integrity checking.

2.1 Databases, Updates, Constraints

An *atom* is of the form $p(t_1, ..., t_n)$, where p is a predicate of arity n ($n \geq 0$); the argument terms t_i are either constants or variables. A *literal* is either an atom A or a negated atom $\sim A$. A *fact* is an atom where all arguments are constants. A *database clause* is either a *fact* or is of the form $A \leftarrow B$, where the *head* A is an atom and the *body* B is a conjunction of literals; all variables in $A \leftarrow B$ are implicitly quantified universally in front of the formula. A *database* is a finite set of database clauses. Let $\mathcal{L}$ denote the underlying language and $\mathcal{L}^c$ the set of constants in $\mathcal{L}$, represented, w.l.o.g, by natural numbers. We use ';' to delimit elements in sets, while ',' symbolizes conjunction in the body of clauses.

An *update* is a finite set of database clauses to be inserted or deleted. For an update U of a database state D, let D^U denote the updated database, where all inserts in U are added to D and all deletes in U are removed from D.

An *integrity constraint* (in short, *constraint*) is a first-order predicate logic sentence, represented as a *denial* of the form $\leftarrow B$, where the body B states what must not hold. Implicitly, each variable in B is universally quantified at the front of $\leftarrow B$. An *integrity theory* is a finite set of constraints.

The DBMS is supposed to ensure that the database satisfies its integrity theory at all times, i.e., that all constraints are logical consequences of each state. To achieve this, database theory requires that, for each update U, the 'old' database D, i.e., the state to be updated by U, must satisfy all constraints, such that integrity checking can focus on those constraints that are possibly affected by the update. If those constraints remain satisfied, then the 'new' state D^U reached by committing U also satisfies all constraints.

Despite the theoretical insistence on total consistency at all times, the quality of a database is likely to suffer and deteriorate in the course of its evolution. Hence, it is necessary to have mechanisms that are able to tolerate certain amounts of integrity violations. As we are going to see, the cause-based approach developed in this paper is inconsistency-tolerant.

From now on, let D, IC, I, U and adornments thereof always stand for a database, an integrity theory, a constraint and, resp., an update. For convenience, we write $D(I) = true$ (resp., $D(I) = false$) if I is satisfied (resp., violated) in D. Similarly, $D(IC) = true$ (resp., $D(IC) = false$) means that all constraints in IC are satisfied in D (resp., at least one constraint in IC is violated in D). Moreover, we assume some familiarity with the completion $comp(D)$, as defined in [14].

2.2 Integrity Checking

In Definition 1, we are going to recall the definition of integrity checking in [8]. It describes each integrity checking method $\mathcal{M}$ as an i/o mapping that takes as input a triple (D, IC, U) and outputs either *ok* or *ko*, as already sketched in Section 1. If $\mathcal{M}$ is sound, $\mathcal{M}(D, IC, U) = ok$ indicates that U preserves integrity, i.e., $D^U(IC) = true$. If $\mathcal{M}$ is also complete, then $\mathcal{M}(D, IC, U) = ko$ indicates that U violates integrity i.e., $D^U(IC) = false$. Also the output *ko* of an incomplete method may mean that the update would violate integrity. But it may also mean that further checking is needed for determining the integrity status of D^U; if there are not enough resources to do so, then U should be cautiously rejected.

Definition 1. (*Sound and complete integrity checking*)
Let $\mathcal{M}$ be a method for integrity checking. $\mathcal{M}$ is called *sound* or, resp., *complete* if, for each (D,IC,U) such that $D(IC)$=*true,* (1) or, resp., (2) holds.

$$\text{If } \mathcal{M}(D, IC, U) = ok \text{ then } D^U(IC) = true. \tag{1}$$

$$\text{If } D^U(IC) = true \text{ then } \mathcal{M}(D, IC, U) = ok. \tag{2}$$

Although the efficiency of intergity checking methods usually is a prime criterion of their usefulness, Definition 1 clearly keeps aloof from such considerations, as well as from any detail of the implementation of $\mathcal{M}$ except its i/o behaviour. That abstraction has been useful for generalizing integrity checking in terms of inconsistency tolerance in [8], and is going to be useful also for conceiving the even more general metric-based concept of integrity checking in Section 4.

2.3 Causes

For a database D and a constraint $I = \leftarrow B$, we are going to define a 'cause' of the violation of I in D as a minimal 'explanation' of why I is violated in D, i.e., why $\exists B$ is *true* in D. An explanation E is going to be defined as an 'excerpt' of $comp(D)$ such that E entails the explained sentence. Roughly, such excerpts are sets of instances of predicate completions in $comp(D)$, or more precisely, of if- and only-if halves of completions. So, let, for each predicate p, the *iff-completion* of p be defined as in [14], and denoted by $\underline{p}$.

Example 2. Let $D = \{p(x,1) \leftarrow r(x);\ p(1,y) \leftarrow s(y,z);\ s(1,2);\ s(2,3)\}$. Then, $comp(D)$ contains the following iff-completions, each of which is a universally closed sentence with existentially quantified 'local' variables that do not occur in the head of any clause. We omit the universal quantifier prenex for all non-local variables in the completions below, and also the equality theory associated to $comp(D)$ that interprets $=$ as identity.

$p(x,y) \leftrightarrow (y = 1 \wedge r(x) \vee x = 1 \wedge \exists z(s(y,z)))$
$r(x) \leftrightarrow false$
$s(x,y) \leftrightarrow (x = 1 \wedge y = 2 \vee x = 2 \wedge y = 3)$

Definition 2. Let D be a database and p a predicate in $\mathcal{L}$.

a) Each formula obtained by substituting each $\forall$-quantified variable in $\underline{p}$ by some ground term is called a *basic instance* of $\underline{p}$.

b) The sentence obtained by replacing $\leftrightarrow$ in $\underline{p}$ or in any basic instance $\underline{p}^*$ of $\underline{p}$ with $\rightarrow$ is called the *only-if half* of $\underline{p}$, or, resp., of $\underline{p}^*$. The atom to the left of $\rightarrow$ of an only-if half H is called the *head* of H and is denoted by $h(H)$.

c) Let D^+ denote the set of ground instances of clauses in D, and D^- the set of basic instances of only-if halves of predicates in $\mathcal{L}$. Assume that D^- is factorized modulo renamings of variables.

For clarity, we represent elements of D^- in a simplified form, if possible. It is obtained by replacing equations with their truth value and applying common equivalence-preserving laws for the composition of subformulas with *true* or *false*. Elements of D^- the simpification of which is *true* are omitted.

Example 3. For p as in Example 2, $p(x,y) \rightarrow (y=1 \wedge r(x) \vee x=1 \wedge \exists z(s(y,z)))$ is the only-if half of $\underline{p}$. Its instance $p(1,1) \rightarrow (1=1 \wedge r(1) \vee 1=1 \wedge \exists z(s(1,z)))$, obtained by $\{x=1, y=1\}$, is clearly equivalent to $p(1,1) \rightarrow (r(1) \vee \exists z(s(1,z)))$. Similarly, the instance $p(2,3) \rightarrow (3=1 \wedge r(2) \vee 2=1 \wedge \exists z(s(3,z)))$ of $\underline{p}$ simplifies to $p(2,3) \rightarrow \mathit{false}$, which is equivalent to $\sim p(2,3)$. Similarly, the instance $s(2,3) \rightarrow (2=1 \wedge 3=2 \vee 2=2 \wedge 3=3)$ of the only-if half of $\underline{s}$ simplifies to $s(2,3) \rightarrow \mathit{true}$, which is equivalent to *true* and thus can be omitted.

Definition 3. Let D be a database and $I = \leftarrow B$ an integrity constraint.

a) Each subset of D^+ is called a *positive excerpt* of D, and each subset of D^- a *negative excerpt* of D. The usual equality theory of $comp(D)$ be associated by default to each negative excerpt.

b) An *excerpt* (P,N) of D consists of a positive excerpt P and a negative excerpt N of D. For each excerpt E, we denote its positive part by E^+, its negative part by E^-, and the union of E^+ and E^- by $\hat{E}$. We say that two excerpts E, E' *overlap* if $\hat{E} \cap \hat{E}' \neq \emptyset$.

c) An excerpt E is called an *explanation base of the violation of* I *in* D if $\hat{E} \models \exists B$.

d) An explanation base E of the violation of I in D is called a *cause of the violation of* I *in* D if there is no explanation base E' of the violation of I in D such that $\hat{E}' \subsetneq \hat{E}$.

Example 4

a) Let $D = \{p \leftarrow q;\ p \leftarrow \sim q\}$ and $I = \leftarrow p$. The two explanations of the violation of I in D are $(\{p \leftarrow \sim q\}, \{\sim q\})$ and $(D, \emptyset)$. Conversely, $(\{p \leftarrow q;\ q\}, \emptyset)$ is not an explanation of the violation of I in D since $q \notin D$.

b) Let $D = \{p(x) \leftarrow r(x);\ r(1)\}$ and $I = \exists x(r(x) \wedge \sim p(x))$. Clearly, $D(I) = \mathit{false}$. A denial form of I is $\leftarrow \mathit{viol}$, where viol is defined by $\{\mathit{viol} \leftarrow \sim q;\ q \leftarrow r(x), \sim p(x)\}$ (q is a fresh 0-ary predicate). Thus, each explanation of the violation of $\leftarrow \mathit{viol}$ in $D' = D \cup \{\mathit{viol} \leftarrow \sim q;\ q \leftarrow r(x), \sim p(x)\}$ explains the violation of I in D. Thus, $(\{\mathit{viol} \leftarrow \sim q\} \cup \{p(i) \leftarrow r(i) \mid i \in \mathcal{K}\},\ \{q \rightarrow \exists x(r(x) \wedge \sim p(x))\} \cup \{\sim r(i) \mid i \notin \mathcal{K}\})$ is a cause of viol in D', for each $\mathcal{K} \subseteq \mathcal{L}^c$ such that $1 \in \mathcal{K}$.

Note that, in examples 4a, b, the constraint, together with the database rules, is always always violated, independent of the presence or absence of facts in D. Hence, one could expect that causes also would be independent thereof. However, the absence of facts in D is used by some causes for explaining the violation of the constraint in 4a and, resp., b. That also illustrates that our concept of causes is applicable even in databases with unsatisfiable integrity theories.

3 Inconsistency Metrics

In 3.1, we present a general axiomatization of inconsistency metrics. It enhances the axiomatization in [9]. In 3.2, we introduce two metrics based on causes.

3.1 Axioms for Inconsistency Metrics

Let $\preccurlyeq$ symbolize an ordering that is antisymmetric, reflexive and transitive. For expressions E, E', let $E \prec E'$ denote that $E \preccurlyeq E'$ and $E \neq E'$. Further, for two elements A, B in a lattice, let $A \oplus B$ denote their least upper bound.

Definition 4. We say that $(\mu, \preccurlyeq)$ is an *inconsistency metric* (in short, a *metric*) if μ maps pairs (D, IC) to some lattice that is partially ordered by $\preccurlyeq$, and, for each pair (D, IC) and each pair (D', IC'), the following axioms (3) – (6) hold.

$$\text{If } \; D(IC) = \mathit{true} \text{ and } D'(IC') = \mathit{false} \; \text{ then } \; \mu(D, IC) \prec \mu(D', IC') \tag{3}$$

$$\text{If } \; D(IC) = \mathit{true} \; \text{ then } \; \mu(D, IC) \preccurlyeq \mu(D', IC') \tag{4}$$

$$\mu(D, IC \cup IC') \preccurlyeq \mu(D, IC) \oplus \mu(D, IC') \tag{5}$$

$$\mu(D, IC) \preccurlyeq \mu(D, IC \cup IC') \tag{6}$$

Axiom (3), called *violation is bad* in [9], ensures that the measured amount of inconsistency in any pair (D, IC) for which integrity is satisfied is always smaller than what is measured for any pair (D', IC') for which integrity is violated. Axiom (4), called *satisfaction is best*, ensures that inconsistency is lowest in any database that totally satisfies its integrity theory. Axiom (5) is a triangle inequality which states that the inconsistency of a composed element (i.e., the union of (D, IC) and (D, IC')) is never greater than the least upper bound of the inconsistency of the components. Axiom (6) requires that the values of μ grow monotonically with growing integrity theories. It is an open issue if (3) – (6) are orthogonal or not.

Occasionally, we may identify a metric $(\mu, \preccurlyeq)$ with μ, if $\preccurlyeq$ is understood.

Example 5. A simple example of a coarse, binary inconsistency metric β is provided by the equation $\beta(D, IC) = D(IC)$, with the natural ordering $\mathit{true} \prec \mathit{false}$ of the range of β, i.e., integrity satisfaction ($D(IC) = \mathit{true}$) means lower inconsistency than integrity violation ($D(IC) = \mathit{false}$).

More inconsistency metrics are defined and discussed in [9]. Axiom (6) of Definition 4 has not been contemplated in [9], but all examples of inconsistency metrics in [9] actually satisfy (6) as well. For instance, the function that maps pairs (D, IC) to the cardinality of the set of cases (instances) of violated constraints is a convenient quality metrics. Inconsistency can also be measured by taking such sets themselves, as elements of the lattice that is constituted by the powerset of all cases of IC, together with the subset ordering. Other metrics can be based on causes of violations, as outlined in the following subsection.

3.2 Cause-Based Inconsistency Metrics

Let $\mathsf{CauVio}(D, I)$ denote the set of causes of the violation of I in D. Further, let $\sigma(D, IC) = \{C \mid C \in \mathsf{CauVio}(D,I),\ I \in IC\}$ be the set of all causes of the violation of some constraint in IC. Then, $(\sigma, \subseteq)$ is an inconsistency metric, and so is $(\zeta, \leq)$, where ζ is defined by $\zeta(D, IC) = |\sigma(D,\ IC)|$ and $|\,.\,|$ denotes set cardinality. In words, σ collects and ζ counts causes of integrity violation.

Example 6. Let $I = \{\leftarrow \mathit{married}(x, y),\ \mathit{married}(x, z),\ y \neq z\}$. I states that no person x may be married to two different partners y and z, i.e., I forbids bigamy. Further, suppose that D contains $\mathit{married}(\mathit{sheik}, \mathit{wife}_1), \ldots, \mathit{married}(\mathit{sheik}, \mathit{wife}_n)$ $(n \geq 2)$ and that there is no other person married twice in D. Then, for each i, j such that $1 \leq i, j \leq n$ and $i \neq j$, $\{\mathit{married}(\mathit{sheik}, \mathit{wife}_i),\ \mathit{married}(\mathit{sheik}, \mathit{wife}_j)\}$ is a cause of the violation of I in D. Hence, the inconsistency in $(D, \{I\})$ as measured by ζ is $\zeta(D, \{I\}) = 1 + 2 + \ldots + n–1$. Thus, for $n > 3$, the inconsistency as measured by ζ that is caused by a man who is married to n different women is higher than the inconsistency of n men being married to just 2 women.

As already mentioned in 3.1, two metrics that are analogous to σ and ζ are featured in [9]. They compare or, resp., count sets of cases, i.e., instances of violated constraints. Those metrics correspond well to the class of integrity checking methods studied in [8], whose output *ok* entails that the given update does not increase the set (resp., the number) of extant violated cases. Still, quantifying the inconsistency of a database by constraint violations does not accurately reflect the amount of stored data that cause inconsistency. Hence, measuring quality by quantifying causes is preferable to quantifying cases.

Example 7. Suppose the predicate p in the constraint $I = \leftarrow p(x, x)$ (which requires the relation corresponding to p to be anti-reflexive) is defined by the two clauses $p(x, y) \leftarrow q(x, y), q(y, x)$ and $p(x, y) \leftarrow r(x, z), s(y, z)$. Further, suppose that the case $I' = \leftarrow p(c, c)$ of I is violated. With that information alone, as provided by measuring violated constraints, it is not clear whether the violation of I is due to the existence of the tuple $q(c, c)$ in the database or to the existence of one or several pairs of tuples of the form $r(c, z)$ and $s(c, z)$ in the join of r and s on their respective last column. In fact, an arbitrary number of causes for the violation of I and even of I' may exists, but the case-based inconsistency metrics in [9] do not give any account of that. As opposed to that, the cause-based metrics σ and ζ obviously do.

Another advantage of causes over cases is that the latter do not provide any means for computing reliable answers to queries in inconsistent databases, while the former do, as shown in [6].

4 Metric-Based Integrity Checking

In this section, we are going to define integrity checking generically by inconsistency metrics: updates are accepted only if they do not increase a measured amount of inconsistency. Also, we are going to show how to obtain inconsistency-tolerant integrity checking methods from metrics such as those in 3.2 and [9].

Definition 5. (*metric-based integrity checking*)
Let $\mathcal{M}$ be a mapping from triples (D, IC, U) to $\{ok, ko\}$, so that U is either accepted or, resp. rejected, and $(\mu, \preccurlyeq)$ an inconsistency metric. $\mathcal{M}$ is called a *sound*, resp., *complete method for integrity checking* if, for each triple (D, IC, U), (7) or, resp., (8) holds.

$$\text{If } \mathcal{M}(D, IC, U) = ok \text{ then } \mu(D^U, IC) \preccurlyeq \mu(D, IC). \tag{7}$$

$$\text{If } \mu(D^U, IC) \preccurlyeq \mu(D, IC) \text{ then } \mathcal{M}(D, IC, U) = ok. \tag{8}$$

If (7) holds, then $\mathcal{M}$ is also called *metric-based*, and, in particular, *μ-based*.

Definitions 1 and 5 are structurally quite similar. However, there are four essential differences. Firstly, the premise $D(IC) = true$ in Definition 1 is missing in Definition 5. This premise requires that integrity be totally satisfied before the update U. By contrast, inconsistency-tolerant integrity checking, as characterized by Definition 5, does not necessarily expect the total satisfaction of all integrity constraints. Rather, it ignores any extant violations (since the total integrity premise is absent), but prevents that inconsistency increases across updates, as guaranteed by the consequence of (7). So, the second difference to be mentioned is that the consequence of (7) clearly weakens the consequence of (1), and, symmetrically, the premise of (8) weakens the premise of (2).

Thirdly, Definition 1 is easily modified in order to characterize repair checking [1], viz. by replacing the premise $D(IC) = true$ with $D(IC) = false$. Then, $\mathcal{M}$ obviously checks whether U restores consistency in D^U. In fact, repair checking does not tolerate any remaining inconsistency in D^U, as opposed to metric-based integrity checking. A relaxation of repair checking which does tolerate inconsistency is presented in [5] and can be abstractly formalized similar to Definition 5.

The fourth difference between the traditional notion of integrity checking in Definition 1 and the metric-based notion in Definition 5 is that the former is significantly generalized by the latter, as expressed in Theorem 1.

Theorem 1. *Each sound or complete integrity checking method (Definition 1) is a sound or, resp., complete metric-based method (Definition 5).*

Proof. Obviously, (7) (resp., (8)) coincides with (1) (resp., (2)) for $\mu = \beta$ (cf. Example 5). If, additionally, $\mathcal{M}$ is a traditional integrity checking method that insists on the total integrity premise, as in Definition 1, then both definitions coincide. Hence, in general, Definition 1 is subsumed by Definition 5. □

In fact, Definitions 1 and 5 do not indicate how $\mathcal{M}$ would compute its output. However, for each metric $(\mu, \preccurlyeq)$, property (9), below, defines a sound and complete μ-based method, as already proved in [9].

$$\mathcal{M}^{\mu}(D, IC, U) = ok \text{ iff } \mu(D^U, IC) \preccurlyeq \mu(D, IC). \tag{9}$$

Hence, for $\mu = \sigma$ or $\mu = \zeta$, we obtain two sound and complete cause-based inconsistency-tolerant integrity checking methods for monitoring and controling quality. That is illustrated by the following example. It also illustrates that different modes or degrees of inconsistency tolerance can be obtained by suitable choices of metrics.

Example 8. Let D and IC be as in Example 6. Further, suppose that *sheik* divorces from $wife_1$, and is about to wed with $wife_{n+1}$, as expressed by the update request $U = \{delete\ married(sheik, wife_1),\ insert\ married(sheik, wife_{n+1})\}$. Thus, $married(sheik, wife_{n+1}) \in D^U$ and $married(sheik, wife_{n+1}) \notin D$, hence $\sigma(D^U, IC) \not\subseteq \sigma(D, IC)$, hence $\mathcal{M}^{\sigma}(D, IC, U) = ko$. On the other hand, we clearly have $\zeta(D^U, IC) = \zeta(D, IC)$, hence $\mathcal{M}^{\zeta}(D, IC, U) = ok$.

At last, the question arises how it can be that, according to Theorem 1, each and every integrity checking method is inconsistency-tolerant, as stated before. After all, several known methods have been shown to be *not* inconsistency-tolerant in [10,8]. Well, the proof of Theorem 1 and property (9) above provide the answer: each method $\mathcal{M}$ that insists on the total integrity premise $D(IC) = true$ in Definition 1 coincides with $\mathcal{M}^{\beta}$, which is a border case of inconsistency-tolerant integrity checking (cf. Example 5). In other words, only metrics that are less coarse than β yield real inconsistency tolerance.

5 Conclusion: Related and Future Work

Many methods and approaches to integrity checking in databases have been discussed in the literature. A generic formal definition of integrity checking that properly includes inconsistency-tolerant methods has been missing so far. Also the idea of measuring the inconsistency of databases by collecting or counting the causes of integrity violations is original of this paper.

In [7], a simple variant of causes is defined as a basis for integrity checking in databases where the bodies of clauses that define the constraints may not contain negation. Causes are not conceived as metrics in [7].

Other related work has been mentioned already in Section 1, except a well-known paper on inconsistency metrics in databases [11] and the literature on inconsistency tolerance in general, as referenced and discussed in [9,8].

In future work, we intend to assign application-specific weights to causes that violate constraints, for tuning cause-based metrics to given applications. Also, we are working on efficient ways to compute causes and cause-based inconsistency metrics. Moreover, we intend to use causes for repairing databases.

References

1. Afrati, F., Kolaitis, P.: Repair checking in inconsistent databases: algorithms and complexity. In: Proc. 12th ICDT, pp. 31–41. ACM Press, New York (2009)
2. Baltopoulos, I., Borgstroem, J., Gordon, A.: Maintaining Database Integrity with Refinement Types. University of Cambridge Whitepaper (2010)
3. Benedikt, M., Bruns, G.: On Guard: Producing Run-Time Checks from Integrity Constraints. In: Rattray, C., Maharaj, S., Shankland, C. (eds.) AMAST 2004. LNCS, vol. 3116, pp. 27–41. Springer, Heidelberg (2004)
4. Christiansen, H., Martinenghi, D.: On simplification of database integrity constraints. Fundam. Inform. 71(4), 371–417 (2006)
5. Decker, H.: Data Quality Maintenance by Integrity-preserving Repairs that Tolerate Inconsistency. To appear in Proc. 11th QSIC. IEEE CSP, Los Alamitos (July 2011)
6. Decker, H.: Answers that Have Integrity. Presented at the ICALP Workshop SDKB 2010, to appear in Post-Workshop, Proc. of SDKB. LNCS. Springer, Heidelberg (2011)
7. Decker, H.: Toward a Uniform Cause-based Approach to Inconsistency-tolerant Database Semantics. In: Meersman, R., Dillon, T., Herrero, P. (eds.) OTM 2010. LNCS, vol. 6427, pp. 983–998. Springer, Heidelberg (2010)
8. Decker, H., Martinenghi, D.: Inconsistency-tolerant Integrity Checking. IEEE TKDE 23(2), 218–234 (2011)
9. Decker, H., Martinenghi, D.: Modeling, Measuring and Monitoring the Quality of Information. In: Heuser, C.A., Pernul, G. (eds.) ER 2009. LNCS, vol. 5833, pp. 212–221. Springer, Heidelberg (2009)
10. Decker, H., Martinenghi, D.: Classifying integrity checking methods with regard to inconsistency tolerance. In: Proc. 10th PPDP, pp. 195–204. ACM Press, New York (2008)
11. Grant, J., Hunter, A.: Measuring inconsistency in knowledgebases. J. Intelligent Information Systems 27(2), 159–184 (2006)
12. Gupta, A., Sagiv, Y., Ullman, J., Widom, J.: Constraint checking with partial information. In: Proc. 13th PODS, pp. 45–55. ACM Press, New York (1994)
13. Leuschel, M., de Schreye, D.: Creating specialised integrity checks through partial evaluation of meta-interpreters. Journal of Logic Programming 36(2), 149–193 (1998)
14. Lloyd, J.: Foundations of Logic Programming, 2nd edn. Springer, Heidelberg (1987)
15. Nicolas, J.M.: Logic for improving integrity checking in relational data bases. Acta Informatica 18, 227–253 (1982)
16. Ramakrishnan, R., Gehrke, J.: Database Management Systems. McGraw-Hill, New York (2003)
17. Ross, K., Srivastava, D., Sudarshan, S.: Materialized View Maintenance and Integrity Constraint Checking: Trading Space for Time. In: Proc. SIGMOD 1996, pp. 447–458. ACM Press, New York (1996)

18. Rotaru, O., Petrescu, M.: A Database Integrity Pattern Language. Leonardo Journal of Sciences 5, 46–62 (2004)
19. Sadri, F., Kowalski, R.: A theorem-proving approach to database integrity. In: Minker, J. (ed.) Foundations of Deductive Databases and Logic Programming, pp. 313–362. Morgan Kaufmann, San Francisco (1988)
20. Zhu, L.: Enforcement of Integrity Constraints in Deductive Databases. PhD Thesis at Simon Fraser University (1992)

OLAP over Continuous Domains via Density-Based Hierarchical Clustering

Michelangelo Ceci[1], Alfredo Cuzzocrea[2], and Donato Malerba[1]

[1] Dipartimento di Informatica, Università degli Studi di Bari "Aldo Modo"
via Orabona, 4 - I-70126 Bari - Italy
{ceci,malerba}@di.uniba.it
[2] ICAR-CNR and University of Calabria
Via P. Bucci, 41C, I-87036 Rende, Cosenza, Italy
cuzzocrea@si.deis.unical.it

Abstract. In traditional OLAP systems, roll-up and drill-down operations over data cubes exploit fixed hierarchies defined on discrete attributes that play the roles of dimensions, and operate along them. However, in recent years, a new tendency of considering even continuous attributes as dimensions, hence hierarchical members become continuous accordingly, has emerged mostly due to novel and emerging application scenarios like sensor and data stream management tools. A clear advantage of this emerging approach is that of avoiding the beforehand definition of an ad-hoc discretization hierarchy along each OLAP dimension. Following this latest trend, in this paper we propose a novel method for effectively and efficiently supporting roll-up and drill-down operations over OLAP data cubes with continuous dimensions via a density-based hierarchical clustering algorithm. This algorithm allows us to hierarchically cluster together dimension instances by also taking fact-table measures into account in order to enhance the clustering effect with respect to the possible analysis. Experiments on two well-known multidimensional datasets clearly show the advantages of the proposed solution.

1 Introduction

Traditional OLAP data cubes are defined on top of discrete dimensions that expose fixed hierarchies [9]. To this end, attribute domains of these dimensions are first discretized, and then processed simultaneously in order to obtain the final cube, given a certain measure [9]. Despite this well-consolidated methodology, a recent trend focuses the attention on the problem of effectively and efficiently computing OLAP data cubes defined on top of continuous dimensions (e.g., [11,22,15]), as the latter are more suitable to capture real-life dynamics rather than the discrete case. Nevertheless, computing such kind of data cubes poses severe challenges, and several alternatives, such as approximation paradigms (e.g., [22]), have been studied to face-off drawbacks deriving from this challenge. On the other hand, OLAP has also been recognized not only as a "last-stage" technology, but also as an important enabling technology that allows us to enhance the expressive power and the quality of retrieved results of a number of Data Warehousing and Mining techniques (e.g., [5]).

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 559–570, 2011.

At the convergence of these two relevant challenges of Data Warehousing and Mining research, in this paper we propose and experimentally assess a novel framework, called OLAPBIRCH, whose main goal consists in integrating the clustering algorithm BIRCH [25] and OLAP. This integration permits to boost the benefits of both methodologies into a complex knowledge discovery framework for next-generation applications ranging from analytics to sensor-and-stream data analysis and social network analysis. OLAPBIRCH relies on top of a complex methodology according to which the capability of the clustering algorithm BIRCH are combined with OLAP in order to build a complex hierarchical data structure, called CF-Tree, whose nodes contain clusters retrieved by BIRCH from the target dataset and are organized in an OLAP-like fashion. This permits to exploit all the deriving benefits such as multidimensional and multi-resolution data exploration, roll-up and drill-down operations, interactive exploration, and so forth [9]. Particularly, supporting roll-up and drill-down operations play a significant role, due to the fact that CF-Tree materializes the retrieved clusters at each node, hence this allows us to significantly speed-up the response time due to executing these critical OLAP operations with respect to the baseline case represented by computing new clusters from pre-existent ones at each roll-up (or drill-down) operation. As an important contribution to actual research, OLAPBIRCH considers continuous dimensions instead than classical discrete ones, which is relevant in OLAP (e.g., [11,22,15]).

In more details, the proposed approach integrates a revised version of BIRCH in order to obtain a hierarchical organization of dimension instances according to similarity computed both on dimension values (from dimension tables) and measure values (from the fact table). We consider the BIRCH algorithm because, in its original formulation, it shows tree important peculiarities: *i)* Efficiency: The algorithm time complexity is linear in the number of instances to cluster and has a constant space complexity. *ii)* Hierarchical: The algorithm allows us to obtain a hierarchical clustering of instances. *iii)* Incremental: As new instances are given to the algorithm, the hierarchical clustering is revised and adapted by keeping into account memory constraints. All these properties well fit to work with data warehouses, where problems coming from the huge amount of data and require for efficient and incremental solutions. For the specific goal we tackle in this paper, it is also necessary to resort to a hierarchical clustering solution that would permit to give to the OLAP users the opportunity to perform roll-up and drill-down operations over continuous attributes. At this aim, we can exploit the peculiarity of BIRCH that provides high balanced hierarchies that become necessary in OLAP frameworks.

The paper is organized as follows. In Section 2, we discuss proposals related to our research. In Section 3, we present the proposed framework OLAPBIRCH. In Section 4, we present an empirical evaluation of the proposed framework an finally, in Section 5 we draw some conclusions.

2 Related Work

Two main areas are pertinent to our research, namely *clustering techniques over large databases*, and *integration of OLAP and Data Mining*. In this Section, we focus the attention on these research areas.

Clustering techniques over large databases is an active area of research that has attracted the attention from a large community of Data Mining researchers since several

decades [13]. A wide family of *clustering algorithms* for large databases has been developed, with alternate fortune. Among those, *CLARANS* [18] is a pioneeristic clustering algorithm that employs *randomized search over a partitioned representation of the target data domain* to discover clusters in spatial databases, which delineate a natural application scenario for clustering techniques [13]. *DBSCAN* [7] in another clustering algorithm that outperforms CLARANS [18] via introducing the notion of cluster density in order to discovery clusters of arbitrary shape. A major benefit of DBSCAN over CLARANS is represented by higher efficiency and scalability over large databases. This is mainly due to the fact that CLARANS assumes that all objects to be clustered can be housed in main memory, which is obviously unrealistic in real-life database application scenarios. Simultaneously to DBSCAN, [25] proposes BIRCH, which is described in detail in Section 3.1. *CURE* [10] is another clustering algorithm whose main idea consists in clustering the input dataset by means of representatives built from the target multidimensional data points via an original approach that combines random sampling and partitioning strategies, in order to *clustering multidimensional datasets in the presence of outliers*. *WaveCluster* [23] is a clustering algorithm based on the well-known *wavelet transforms*. The most distinctive characteristic of WaveCluster is represented by its *low dependence* on the fact input data may be sorted or unsorted, and its *low sensibility* on the presence of outliers in data. In addition to this, well-understood multi-resolution tools made available by wavelet transforms make WaveCluster able to discover clusters of arbitrary shape at different level of accuracy (of clustering). *CLIQUE* [1] instead focuses the attention on a more-refined clustering problem called *automatic subspace clustering of high-dimensional data*, i.e. clustering the target data with respect to a partition of the original features of the reference data source, even without the support of feature selection algorithms. CLIQUE makes use of a *bottom-up approach* that starts from the clustering at the lower dimensionality (of the target dimensional space) and progressively derives the clustering at the higher dimensionalities via the well-known *monotonicity property* inspired by the popular association rule mining algorithm *Apriori* [2]. Basically, this property argues that, if a collection of data points c is a cluster in a k-dimensional space S, then c is also part of a cluster in any $(k-1)$-dimensional projections of S. As we will highlight in the following, the subspace clustering problem in general, and the CLIQUE approach in particular, are very related to our research.

More recently, Data Mining researchers have concentrated their efforts towards the problem of *clustering methodologies in complex database environments*, which has produced a wide and rich literature on this specific topic. Among these, *CrossClus* [24] is a significantly-representative instance. CrossClus considers an applicative setting where data are stored in *semantically-linked database relations* and the clustering phase must be conducted across multiple relations rather than only one like most proposals assume. To this end, CrossClus devises an innovative methodology according to which the clustering phase is "propagated" across database relations by following associations among them, just starting from a small set of (clustering) features defined by users. Finally, a specialized clustering context directly related to our research deals with the problem of *clustering high-dimensional datasets* [16], since high-dimensional data are reminiscent of OLAP data. Here, *clustering scalability* and quality of clustering play the roles of

major research challenges, as it is well-understood that traditional clustering approaches are not effective on high-dimensional data [16].

Integration of OLAP and Data Mining is clearly relevant for our research too. Han initially introduces *OLAM methodologies* [12] as a first step towards achieving this so-important result by mainly focusing on the issue of extracting knowledge from OLAP data cubes. [3] proposes extending traditional OLAP functionalities over distributed database environments with the goal of generating *specialized data cubes storing association rules* rather than conventional SQL-based aggregates. A similar experience is done in [8], but with the difference of using data cubes as primary input data structures for *association rule mining*. Other kinds of OLAP/Data-Mining integration proposals are: (*i*) [19], which argues that applying Data Mining over data cubes definitely improves the effectiveness and the reliability of *decision-support systems*; (*ii*) [20,21], which propose integrating statistical tools within OLAP servers with the goal of supporting *discovery-driven exploration* of data cubes; (*iii*) [14], which makes use of data cubes as conceptual layer of the association rule mining phase; (*iv*) [6], which introduces *gradient analysis over OLAP data cubes*, a sophisticated data cube mining technique for detecting *significant* changes among fixed collections of cube cells; (*v*) [17], which proposes a framework for mining *inter-dimensional association rules* from data cubes on the basis of SUM-based aggregate measures as a crucial improvement over traditional COUNT-based frequency analysis that underlies conventional association rule mining methodologies.

3 OLAPBIRCH: Combining BIRCH and OLAP

In this section, for the sake of completeness, we first describe the BIRCH algorithm and then we describe proposed modifications that permit to integrate BIRCH in an OLAP framework.

3.1 BIRCH

The BIRCH algorithm [25] works on a hierarchical data structure that the authors call CF tree (Clustering Feature Tree). This data structure permits to partition the incoming data points in an incremental and dynamic way. Each node in the CF tree is called Clustering Feature: Given n data points in a cluster, each of which represented according to a d-size feature vector, CF vector of the cluster is defined as a triple $CF = (n, LS, SS)$, where LS is the linear sum and SS is the square sum of data points. The CF vectors are sufficient to compute information about subclusters like centroid, radius and diameter. They satisfy an important additivity condition, i.e. if $CF_1 = (n_1, LS_1, SS_1)$ and $CF_2 = (n_2, LS_2, SS_2)$ are the clustering features for sets of points S_1 and S_2 respectively, then $CF_1 + CF_2 = (n_1 + n_2, LS_1 + LS_2, SS_1 + SS_2)$ is the clustering feature for the set $S_1 \cup S_2$.

A CF tree is a balanced tree whose structure is similar to that of a B+tree and depends on two parameters: the branching factor B and a user defined threshold T that represents the maximum cluster diameter. Each non-leaf represents a cluster consisting of all the

subclusters represented by its entries. In particular, a non-leaf node N_j contains at most B entries of the form $[CF_i, c_i]_{i=1,..,B}$, where c_i is a pointer to the $i-th$ child node of N_j and CF_i is the clustering feature of the cluster identified by c_i. A leaf node contains at most L (typically $L = B$) entries each of the form $[CF_i]$. In the leaves, each node has two pointers $prev$ and $next$ which are used to chain all leaf nodes together. The tree size depends on the T value: the larger the T, the smaller the tree.

The algorithm BIRCH builds a CF tree in four steps. In the first step, BIRCH iteratively receives single data points and builds an initial CF-tree. A point is inserted by inserting the corresponding CF value into the closest leaf. If an entry in the leaf can absorb the new point without violating the threshold T condition, its CF is updated. Otherwise, a new entry is created in the leaf node, and, if the leaf node then contains more than L entries, it and maybe its ancestors are split. In this phase, in order to satisfy RAM constraints, BIRCH frequently rebuilds the whole CF tree by increasing the threshold T and tries to merge as many CF nodes as possible. The rebuild happens sufficiently fast since all needed data is already in RAM. At the same time outliers are removed from the tree and are stored to disk. The algorithm starts with maximum precision at $T = 0$ and as the CF tree grows larger than the available memory, it iteratively tries to find suitable cluster sizes by increasing T to be larger than the smallest distance between two entries in the tree. In the second step, the algorithm condenses the CF tree to a desirable size depending on the clustering algorithm employed in step three. This can involve removing outliers and further merging of clusters. In the third step, the algorithm employs a *global* clustering algorithm using the CF tree's leaves as input. This step, as claimed in [25], permits to avoid the undesirable effect of the skewed input order, and splitting triggered by space constraints. In this phase, the CF vectors allow for effective distance metrics computation. In the last step, a labeling procedure is performed. This means that, if desired, the actual data points can be associated with the generated clusters.

In the implementation we provide, the used distance measure is the variance increase distance [25] defined as follows:

Definition 1 (Variance Increase Distance)
Let C_1 and C_2 be two clusters, where $C_1 = \{x_i\}_{i=1..n_1}$ and $C_2 = \{x_i\}_{i=n_1+1..n_2}$. The variance increase distance between C_1 and C_2 is defined as:

$$D = \sum_{k=1}^{n_1+n_2} \left(x_k - \frac{\sum_{l=1}^{n_1+n_2} x_l}{n_1+n_2} \right)^2 - \sum_{i=1}^{n_1} \left(x_i - \frac{\sum_{l=1}^{n_1} x_l}{n_1} \right)^2 - \sum_{i=n_1+1}^{n_2} \left(x_i - \frac{\sum_{l=n_1+1}^{n_2} x_l}{n_2} \right)^2$$

In our implementation, the clustering algorithm for the third step is the well-known DBSCAN [7] algorithm that performs a density based clustering. Density based clustering is performed on cluster centroids (that can be easily computed from the CF vectors and represent aggregated data) and allows us to further aggregate data that show similar peculiarities.

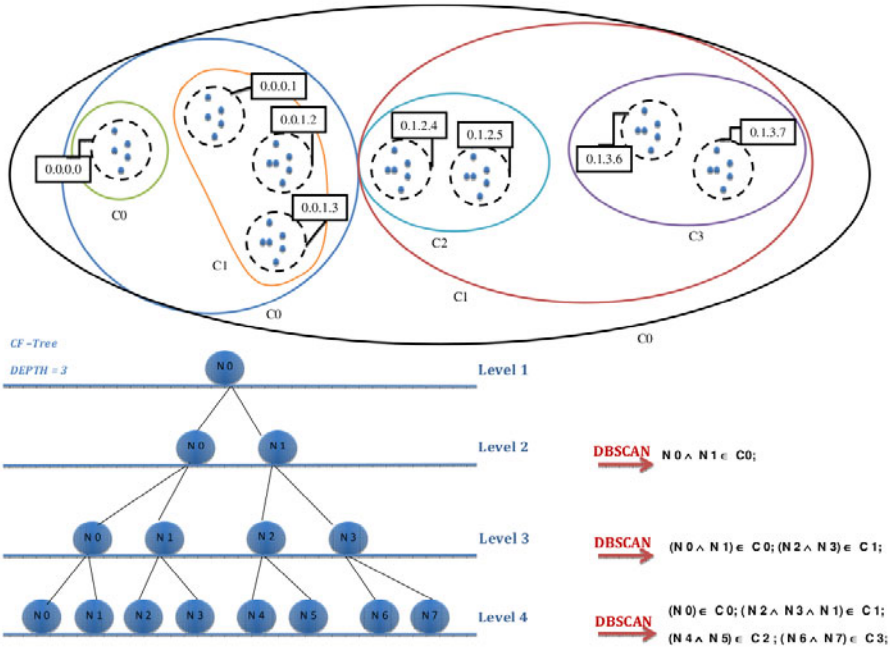


Fig. 1. OLAPBIRCH: An example of CF tree

3.2 OLAPBIRCH

The integration of the implemented BIRCH algorithm in the OLAP solution we present, is not a trivial task since different issues have to be considered: First, OLAP queries can consider all the levels of the hierarchy and not only the last level. This means that it is necessary to have refined clusters not only in the last level of the hierarchy, but also in intermediate levels. Second, in OLAP frameworks, the user is typically able to control size of hierarchies, but this is not possible in the original BIRCH algorithm. Third, although the last step of the BIRCH algorithm is not mandatory, this step is necessary in our framework in order to simplify the computation of OLAP queries. Fourth, in order to avoid the combinatorial explosion that is typical in multidimensional clustering, it is necessary to focus only on interesting continuous dimension attributes.

In order to face with the first issue, we revised the clustering algorithm in order to allow the system to run the *global* clustering algorithm also in intermediate nodes of the tree. At this purpose, we extended the CF tree structure by providing pointers $prev$ and $next$ to each internal node. This allows us to linearly scan a each single level of the tree. In Figure 1, we report a graphical representation of the CF tree structure used in the proposed framework.

As for the second issue, in addition to the memory space constraints, we consider also an additional constraint that forces tree rebuilding when a maximum number of levels (MAX_LEV) is exceeded. This is coherent with the goal of having a limited number of levels, as in classical OLAP systems.

As for the third issue, given the maximum number of levels MAX_LEV and the branching factor B, it is possible to use a numerical representation of the complete path of clusters for each dimension instance so that the classical B+tree index structure can be used in order to allow efficient computation of range queries[11]. The representation is in the form $< d_1 d_2 \ldots d_{MAX\_LEV} >$, where each d_i is a sequence of $\lceil log_2 B \rceil$ bits that permits to identify each subcluster. The number obtained in this way is then used to perform roll-up and drill-down operations.

Finally, as for the fourth issue, in order to integrate the algorithm in an OLAP framework, we defined a language that permits to specify the attributes to be considered in the clustering phase. At this purpose, we have exploited the Mondrian[1] project that permits to represent a multidimensional schema of a data warehouse by means of an XML file. This file permits to define a mapping between the multidimensional schema and tables and attributes stored in the database. Main elements in this XML file are: the data source, cubes, measures, the fact table, dimensions and hierarchies.

For our purposes, we have modified the DTD in order to allow different types of hierarchies. The modified portion of the DTD is:

```
<!ELEMENT Hierarchy ((\%Relation;)?,(Level)*,
          (MemberReaderParameter)*,(Attribute)+, (Depth))>
    <!ATTLIST Hierarchy
     hasAll (true|false) #REQUIRED
     allMemberName CDATA #IMPLIED
     allMemberCaption CDATA #IMPLIED
     primaryKey CDATA #IMPLIED
     primaryKeyTable CDATA #IMPLIED
     defaultMember CDATA #IMPLIED
     memberReaderClass CDATA #IMPLIED>
<!ELEMENT Attribute EMPTY>
    <!ATTLIST Attribute
     name CDATA #IMPLIED
     table CDATA #REQUIRED
     column CDATA #REQUIRED
     nameColumn CDATA #REQUIRED
     type (Numeric) Numeric #REQUIRED>
<!ELEMENT Depth EMPTY>
    <!ATTLIST Depth value (Numeric) Numeric #REQUIRED>
```

The DTD so modified permits to add two new elements ($< Attribute >$ and $< Depth >$) to the elements defined in $< Hierarchy >$. The $< Attribute >$ element permits to define one or more attributes to be used in the clustering procedure. Properties that can be defined in the $< Attribute >$ tag are: name - attribute name; table - table that contains the attribute; column: database column name; nameColumn: database column name (alias); type: SQL attribute type. The $< Depth >$ element permits to specify the maximum depth of the CF-tree.

[1] http://sourceforge.net/projects/mondrian/files/mondrian/

The CF-tree is updated when a new dimension tuple is saved in the data warehouse while DBSCAN is run only when OLAP queries are executed and clusters are not updated. This permits to focus our attention only to levels that are actually used in the queries. It is noteworthy that, differently from [22], the global clustering is run on compact representations of data and does not pose efficiency problems.

Example 1 Let us consider the database schema reported in Figure 2 where $lineitem$ is the fact table and $orders$ is a dimensional table. By selecting, in the XML file, the attributes $orders.o\_totalprice$ and $orders.o\_orderpriority$:

```
< Attribute name="totalprice" table="orders" column="
      o_totalprice" nameColumn="o_totalprice"
      type="Integer"/>
< Attribute name="orderpriority" table="orders" column="
      o_orderpriority" nameColumn="o_orderpriority"
      type="Integer" />
< Depth value="20"/>
```

we have that the OLAP engine performs clustering on the following database view:

```
SELECT l_quantity, l_extendedprice, l_discount, l_tax,
    o_totalprice, o_orderpriority
FROM lineitem, orders
WHERE l_orderkey = o_orderkey
```

4 Experimental Evaluation and Analysis

In order to evaluate the effectiveness of the proposed solution, we performed experiments on two real world datasets. The first dataset is the SPAETH Cluster Analysis Datasets[2], a small dataset that allows us to visually evaluate the quality of extracted clusters. The second dataset is the well-know TPC-H benchmark (version 2.1.0)[3]. In Figure 2 we report the relational schema of TPC-H implemented on PostgreSQL, which we used as supporting DBMS. The TPC-H database holds data about the ordering and selling activities of a large-scale business company. For experiments we used the 1GB version of TPC-H [4] containing more that 1×10^6 tuples in the fact table. On this last dataset, we performed experiments on the scalability on the algorithm and we collected results in terms of running times and cluster quality. Clustering is performed on the fact table measures as well as on the the attributes orders.o_totalprice and orders.o_orderpriority as specified in Example 1. In order to force the OLAPBIRCH framework to work in the worst case scenario, running times are obtained by forcing the system to work when the examples are given one by one. The cluster quality is measured according to the *weighted average cluster diameter square measure*:

$$Q = \sum_{i=1..K} n_i(n_i - 1)D_i^2 / \sum_{i=1..K} n_i(n_i - 1) \quad (1)$$

[2] http://people.sc.fsu.edu/∼jburkardt/datasets/spaeth/spaeth.html

[3] Transactions Processing Council Benchmarks. Available from: http://www.tpc.org

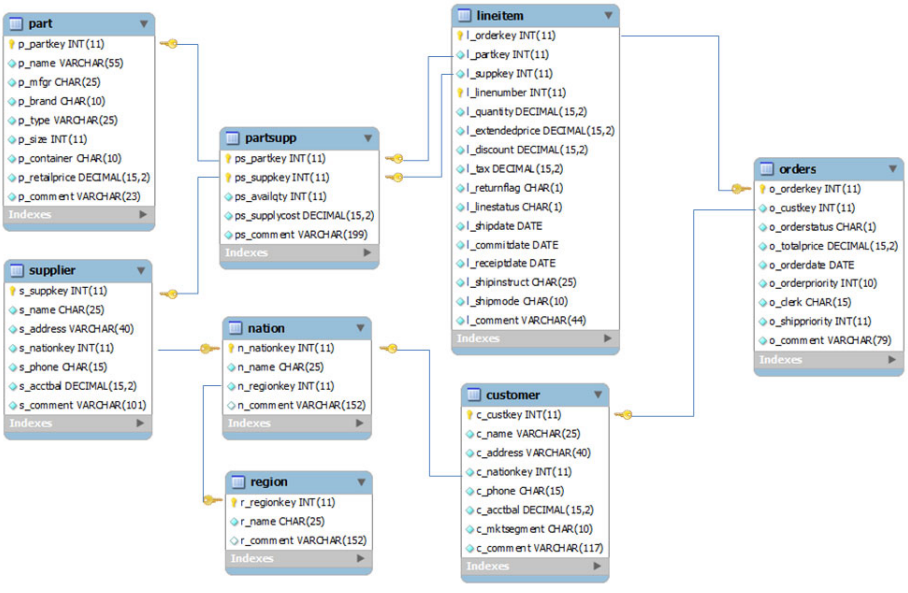


Fig. 2. TPC-H database schema

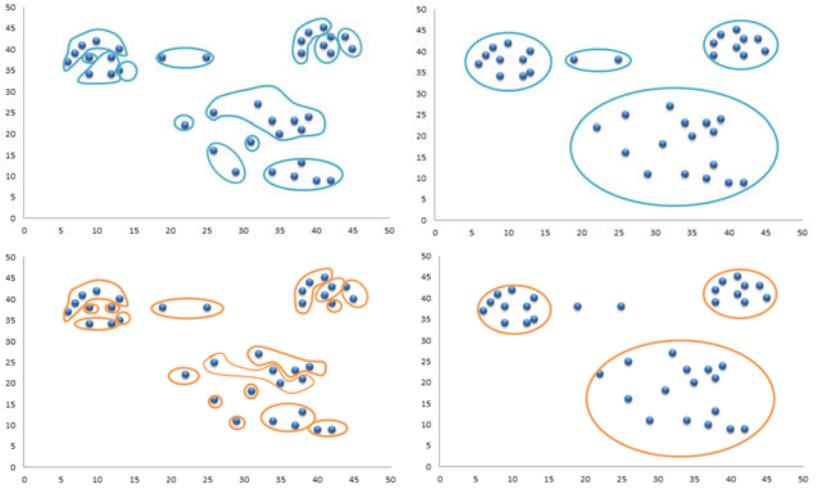


Fig. 3. Clustering effect on Spaeth dataset. CF-tree is obtained with B=L=2. Left: OLAP-BIRCH without DBSCAN, Right: OLAPBIRCH with DBSCAN; Top: $LEVEL = 6$, Bottom: $LEVEL = 7$. Points outside clusters are considered outliers.

where K is the number of obtained clusters, n_i is the cardinality of the i-th cluster and D_i is the diameter of the i-th cluster. The smaller the Q value, the higher the cluster quality.

Table 1. TPC-H: scalability results. MAX_LEV= 20, $B = L = 2$.

No of points	Running time (s)	Q	No of rebuilds
60×10^3	776	0.08	5
100×10^3	1819	0.07	5
500×10^3	41,646	0.018	5
1.1×10^6	172,800	0.039	9

In Figure 3, we report a graphical representation of obtained clusters. As we can see, the global clustering (DBSCAN) is necessary in order to have good quality clusters. Moreover, by increasing the depth of the tree, it is possible to have more detailed clusters.

Results obtained on the TPC-H database are reported in Table 1. From these results, it is possible to see that the quality of the clusters does not deteriorate when the number of examples increases. We can also see that the number of times that the CF-tree is rebuilt is very small, even for huge datasets.

Figure 4 shows a different perspective of the obtained results. In particular, it shows that there is strong correlation between the region dimension (that is not considered during the clustering phase) and the obtained clusters. This means that numerical properties of the orders change in distribution between the regions where the order is performed.

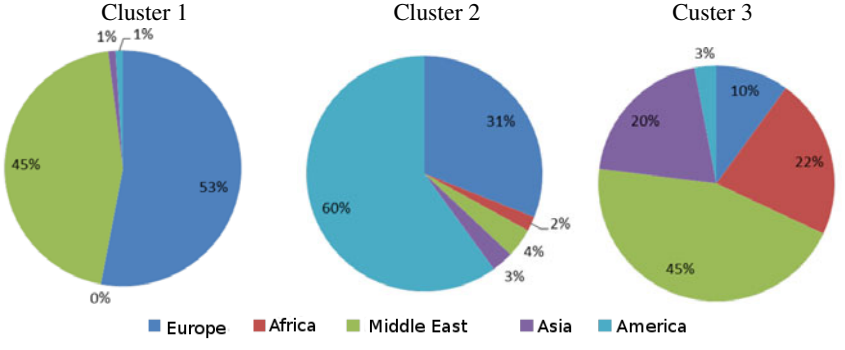


Fig. 4. TPC-H: Data distribution over the Region dimension

5 Conclusions and Future Work

In this paper we have presented the framework OLAPBIRCH. This framework integrates a clustering algorithm in an OLAP engine in order to support roll-up and rill-down operations on numerical dimensions. OLAPBIRCH integrates a revised version of the BIRCH clustering algorithm that permits to incrementally revise hierarchical clustering for all the levels of the hierarchy. Preliminary results show the effectiveness of the proposed solution on large real world datasets. For future work we intend to

compare our framework with competitive frameworks and prove its applicability in answering to range queries and we intend to run experiments by varying input parameters in order to give better insights on their definition.

Acknowledgment. This work is partial fulfillment of the research objective of ATENEO-2009 project "Estrazione, Rappresentazione e Analisi di Dati Complessi".

References

1. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data. Data Min. Knowl. Discov. 11(1), 5–33 (2005)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Bocca, J.B., Jarke, M., Zaniolo, C. (eds.) VLDB 1994, Proceedings of 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile, September 12-15, pp. 487–499. Morgan Kaufmann, San Francisco (1994)
3. Chen, Q., Dayal, U., Hsu, M.: An OLAP-based scalable web access analysis engine. In: Kambayashi, Y., Mohania, M., Tjoa, A.M. (eds.) DaWaK 2000. LNCS, vol. 1874, pp. 210–223. Springer, Heidelberg (2000)
4. Cuzzocrea, A.: Improving range-sum query evaluation on data cubes via polynomial approximation. Data Knowl. Eng. 56(2), 85–121 (2006)
5. Cuzzocrea, A., Serafino, P.: Clustcube: An olap-based framework for clustering and mining complex database objects. In: SAC (2011)
6. Dong, G., Han, J., Lam, J.M.W., Pei, J., Wang, K.: Mining multi-dimensional constrained gradients in data cubes. In: VLDB, pp. 321–330 (2001)
7. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, pp. 226–231 (1996)
8. Goil, S., Choudhary, A.N.: Parsimony: An infrastructure for parallel multidimensional analysis and data mining. J. Parallel Distrib. Comput. 61(3), 285–321 (2001)
9. Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., Pirahesh, H.: Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub totals. Data Min. Knowl. Discov. 1(1), 29–53 (1997)
10. Guha, S., Rastogi, R., Shim, K.: Cure: An efficient clustering algorithm for large databases. Inf. Syst. 26(1), 35–58 (2001)
11. Gunopulos, D., Kollios, G., Tsotras, V.J., Domeniconi, C.: Selectivity estimators for multidimensional range queries over real attributes. VLDB J. 14(2), 137–154 (2005)
12. Han, J.: Towards on-line analytical mining in large databases. SIGMOD Record 27(1), 97–107 (1998)
13. Hinneburg, A., Keim, D.A.: Clustering methods for large databases: From the past to the future. In: SIGMOD Conference, p. 509 (1999)
14. Imielinski, T., Khachiyan, L., Abdulghani, A.: Cubegrades: Generalizing association rules. Data Min. Knowl. Discov. 6(3), 219–257 (2002)
15. Karayannidis, N., Sellis, T.K.: Hierarchical clustering for olap: the cube file approach. VLDB J. 17(4), 621–655 (2008)
16. Kriegel, H.-P., Kröger, P., Zimek, A.: Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. TKDD 3(1) (2009)
17. Messaoud, R.B., Rabaséda, S.L., Boussaid, O., Missaoui, R.: Enhanced mining of association rules from data cubes. In: DOLAP, pp. 11–18 (2006)
18. Ng, R.T., Han, J.: Clarans: A method for clustering objects for spatial data mining. IEEE Trans. Knowl. Data Eng. 14(5), 1003–1016 (2002)

19. Parsaye, K.: Olap and data mining: Bridging the gap. Database Programming and Design 10, 30–37 (1997)
20. Sarawagi, S.: idiff: Informative summarization of differences in multidimensional aggregates. Data Min. Knowl. Discov. 5(4), 255–276 (2001)
21. Sarawagi, S., Agrawal, R., Megiddo, N.: Discovery-driven exploration of olap data cubes. In: Schek, H.-J., Saltor, F., Ramos, I., Alonso, G. (eds.) EDBT 1998. LNCS, vol. 1377, pp. 168–182. Springer, Heidelberg (1998)
22. Shanmugasundaram, J., Fayyad, U.M., Bradley, P.S.: Compressed data cubes for olap aggregate query approximation on continuous dimensions. In: KDD, pp. 223–232 (1999)
23. Sheikholeslami, G., Chatterjee, S., Zhang, A.: Wavecluster: A wavelet based clustering approach for spatial data in very large databases. VLDB J. 8(3-4), 289–304 (2000)
24. Yin, X., Han, J., Yu, P.S.: Crossclus: user-guided multi-relational clustering. Data Min. Knowl. Discov. 15(3), 321–348 (2007)
25. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: An efficient data clustering method for very large databases. In: Jagadish, H.V., Mumick, I.S. (eds.) SIGMOD Conference, pp. 103–114. ACM Press, New York (1996)

Non-separable Transforms for Clustering Trajectories

Alfredo Cuzzocrea and Elio Masciari

ICAR-CNR – Institute of Italian National Research Council
{cuzzocrea,masciari}@icar.cnr.it

Abstract. Trajectory data refer to time and position of moving objects generated by different sources using a wide variety of technologies (e.g., RFID tags, GPS, GSM networks). Mining such amounts of data is challenging, since the possibility to extract useful information from these peculiar kind of data is crucial in many application scenarios such as vehicle traffic management, hand-off in cellular networks and supply chain management. In this paper, we address the problem of trajectory data streams clustering, that revealed really challenging as we deal with data (trajectories) for which the order of elements is relevant. We propose a complete framework starting from data preparation task that allows us to make the mining step quite effective. Since the validation of data mining approaches has to be experimental we performed several tests on real world datasets that confirmed the efficiency and effectiveness of the proposed techniques.

1 Introduction

Data Clustering is one of the challenging mining techniques exploited in the knowledge discovery process[5]. Clustering huge amounts of data is a difficult task since the goal is to find a suitable partition in a unsupervised way (i.e. without any prior knowledge) trying to maximize the similarity of objects belonging to the same cluster and minimizing the similarity among objects in different clusters. Many different clustering techniques have been defined in order to solve the problem from different perspectives, i.e. partition based clustering (e.g. *K-means*[10]), density based clustering (e.g. *DBScan*[3]), hierarchical methods (e.g. *BIRCH*[15]) and grid-based methods (e.g. *STING* [14]). Moreover, clustering methods have been exploited in a wide variety of application scenarios ranging from transactional data, text documents, XML data, etc. A main problem for clustering data is that there is a great degree of uncertainty both in the data selection phase and in the definition of clusters, moreover due to complexity issues some algorithms do not scale-up very well when the size of the dataset becomes really huge.

Trajectory data can be generated in a wide variety of applications, such as GPS systems [4], supply chain management [9] , vessel classification by satellite images [8]. As an example of such data consider moving vehicles, where both cars and trucks leave a digital trace by the personal or vehicular mobile devices

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 571–580, 2011.

that can be collected via a wireless network infrastructure. Furthermore, mobile phones continuously signalling their locations (cell), are at each moment connected to their GSM network. Consider the case of a phone setting up a call and moving through the network, there may occur the so called hand-off problem, i.e. the cell where the user is moving through does not have enough bandwidth to place the call. In the same way, GPS-equipped portable devices can record their latitude-longitude positions at each moment that they are exposed to a GPS satellite, and transmit their trajectories to a collecting server. Due to this large amount of moving objects data generated every day, there is a great need for analyzing it efficiently in order to extract useful information.

Trajectory data carry information about actual position and timestamp of moving objects at a detail level often unnecessary. Indeed, many proposals split the search space in regions having the suitable granularity and represent them as areas tagged by an annotated symbol. The sequence of regions(symbols) define the trajectory traveled by a given object. Based on the above representation of trajectory data (i.e. region based instead of a sequence of multidimensional points) mining steps are performed based on proper techniques. Note that regioning is a common assumption in trajectory data mining [8,4]. On the contrary, we allow to exploit crucial information about common trajectories "shape", i.e. not only the length, but the direction of movement and the eventual turn made by the trajectory. We propose a framework that suitably pre-elaborate trajectories for the subsequent mining step performed using non-separable transforms that allow to work directly on the multidimensional representation of the trajectories thus making the mining process lossy.

Approach outline. In this paper a solution for data pre-elaboration based on a proper filtering of the multidimensional data based on *Lifting Schemes* [13] is proposed. The aim of lifting is to represent a spatial signal (i.e. the whole trajectory) using a shorter sequence by a proper filtering step that allow prediction and update of proper coefficients. Lifting schemes are successfully exploited in image processing since they offer really good performances without any loss of information. They are introduced because in many context it is mandatory to work on the overall trajectory considering all the involved dimensions. After the pre-processing step, we exploit non-separable Fourier transforms since they can be computed efficiently even if the data set size is huge. To this end we define a clustering strategy based on Fourier Analysis in order to catch "structural" dissimilarity between trajectories. The basic intuition exploited here is that a trajectory has a "natural" interpretation as a time series (namely, a discrete-time signal), in which numeric values summarize some relevant features of the elements belonging to the trajectory. In a sense, the analysis of the way the signal shapes differ can be interpreted as the detection of different regions crossed by the trajectory. Moreover, the analysis of the frequencies of common signal shapes can be seen as repeated crossing of the same region. In this context, the proposed approach can be seen as an efficient technique, which can satisfactorily evaluate how much two trajectories are similar w.r.t. the structural features previously discussed. We point out that in this work we disregard speed and time

information since they are out of the scope of this paper so we assume that points in the trajectories appear at fixed time interval. This is not at all a limitation since we are interested in discovering common structures in (static) trajectory datasets. The choice of comparing the frequency spectra of trajectories is due to both effectiveness and efficiency issues. Indeed, the exploitation of Non-Separable Fourier Transforms (in particular we use Discrete Fourier Transform - DFT [11] since it enable us to analyze and manipulate signals in a very powerful way) leads to abstract from structural details which, in most application contexts, should not affect the similarity estimation such as, e.g., different numbers of occurrences of the same location (i.e. simple traffic problems) or small shifts in the actual positions where a point appears in the trajectory. This eventually makes the comparison less sensitive to minor mismatches. Moreover, a frequency based approach allows to estimate the similarity through simple measures (e.g., vector distances) which are computationally less expensive than techniques based on the direct comparison of the original trajectory structures. To summarize, we propose to represent a trajectory as a time series of fixed frequency, in which each trajectory point corresponds to an impulse. By analyzing the frequency spectra of the signals so far obtained, we can hence state the degree of similarity between trajectories. It is worth noticing that the overall cost of the proposed approach is only $O(N \log N)$, where N is the maximum number of regions in the trajectories to be compared. Moreover, it exhibits a satisfactory effectiveness even when compared with other existing approaches. As mentioned above our approach is quite intriguing since there are some application scenarios for which the monodimensional representation (i.e. trajectories partition) is not well suited so we need to work on the two dimensional trajectory. Since manipulating the original trajectory could be quite difficult (especially for distance definition matter) we need to transform it properly. In this case we use a powerful mathematical tool that exploit Fourier Non-Separable Transform by choosing a proper basis. The transform so far obtained allows us to compare a compact representation of the trajectory that reflect all the features of the original one. As a matter of fact, we need to properly define a distance measure that reflect the desired features of our clustering scheme.

As for every data mining task the evaluation has to be experimental in nature, we performed several tests in order to assess the validity of our proposal and the results obtained are quite convincing.

2 Problem Statement

In this paper we tackle the problem of data clustering from large corpus of trajectory data streams. While for transactional data a tuple is a collection of features, a trajectory is an ordered set (i.e., a sequence) of timestamped points. Trajectory data are usually recorded in a variety of different formats, and they can be drawn from a continuous domain. We assume a standard format for input trajectories, as defined next.

Definition 1 (Trajectory). *Let P and T denote the set of all possible (spatial) positions and all timestamps, respectively. A trajectory is defined as a finite sequence $s_1, \cdots, s_N$, where $N \geq 1$ and each s_i is a pair (p_i, t_i) where $p_i \in P$ and $t_i \in T$.*

We assume that P and T are discrete domains. For continuous locations a viable approach is to partition the space into regions to map the initial locations into discrete regions labeled with a timestamped symbol. Notice that the way chosen for the assignment of symbols to locations is totally irrelevant for the clustering goal since we are interested in clustering the trajectories as a whole. The granularity level of region partitioning can be set according to the application requirements as will be discussed below. The problem of finding a suitable partitioning for both the search space and the actual trajectory is a core problem when dealing with spatial data. Every technique proposed so far, somehow deals with regioning and several approaches have been proposed such as partitioning of the search space in several regions of interest (RoI)[4] and trajectory partitioning [7,6] by using polylines. However, in this paper we present an approach that does not introduce any approximation since it consider the whole trajectory points as will be exploited in next sections.

2.1 Multiresolution Analysis for Trajectories

Trajectory data are usually two dimensional thus when a dramatic accuracy is required we cannot disregard any point. Therefore, there is a need for a multiresolution analysis to take into account both the spatial dimensions in the preprocessing step.

Definition 2. *Let $\{S_m\}$ be a set of subspace of $L^2(\mathcal{R})$ and $m \in \mathcal{Z}$ such that the following hold:*

- $S_m \in S_{m+1}$;
- $\bigcup_{j \in \mathcal{Z}} = L^2(\mathcal{R})$;
- $\bigcap_{j \in \mathcal{Z}} = \emptyset$;
- $x(t) \in S_m \Leftrightarrow x(\frac{t}{2} \in S_{m-1})$.

$\{S_m\}$ is a multi-resolution system.

The choice of the $\{S_m\}$ defines the analysis being performed in particular if we choose orthogonal subspace we have *orthogonal multi-resolution* analysis. We exploit here L space since it has a proper norm.

For trajectory data multi-resolution analysis aims at representing each trajectory being analyzed (and then elaborated using a mathematical transform) using a reliable set of coefficients. In order to perform this step allowing a perfect reconstruction of trajectories we adopt the so called *Lifting Scheme* approach.

2.2 Lifting Schemes

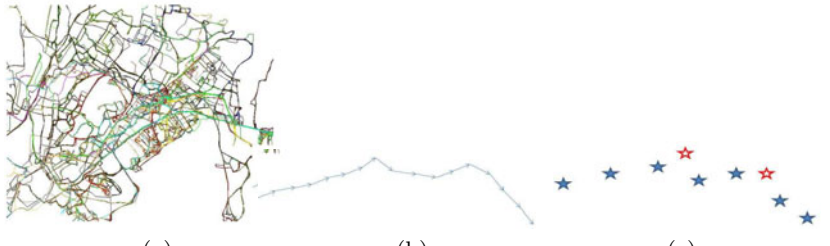
An effective approach for multi-resolution analysis is the so called *lifting scheme*. It was originally introduced for filtering signal and due to its intuitive features and more important because of its ability to exactly reconstruct the original input sequence it has been widely used also as a support for image compression in wavelet based systems. Moreover, it is well suited as a preprocessing step for non-separable transforms. In order to perform the proper lifting we need to define a filtering function that could be for example Least Mean Square, Regression or Kalman filtering.

Let $\mathcal{F}$ be a filtering function and Tr_x the input trajectory. We split Tr_x in two subsequences Tr_{x_o} and Tr_{x_e} that are respectively the sequence of odd and even indices. The lifting is performed by iteratively updating the subsequences with their predicted version in order to obtain two shorter sequences that are representative of the original sequence (say it Tr'_x) and the trend (say it Tr_h) respectively. Obviously, when performing this step some errors could arise (say it Tr_e). More formally, let $\mathcal{P}$ and $\mathcal{U}$ two filtering function exploited for prediction and update, a generic lifting step works as follows:

- $Tr_e(k) = Tr_{x_o}(k) - \mathcal{P}(Tr_{x'}(k))$;
- $Tr_h(k) = Tr_{x_e}(k) - \mathcal{U}(Tr_{x_e}(k))$.

The proposed lifting scheme will be used for implementing the non separable transform exploited for clustering in next sections, when the trajectory size becomes unpractical we can reduce it by a lifting step. Since in our case we deal with trajectories the proposed scheme may seem rather obscure, to better clarify the proposed pre-elaboration step consider the following example.

Consider the trajectories depicted in Figure 1(a) and the sample trajectory zoomed for clarity in Figure 1(b).



(a) (b) (c)

Fig. 1. A set of trajectories (a) a sample trajectory (b) and its lifted version (c)

Applying the lifting scheme defined above will produce the sequence in Figure 1(c) where the solid blue stars represent the trend sequence, while the red stars represent the error sequence. As it is easy to see the number of points taken into account after lifting is much smaller and this feature is particularly useful when considering (real) longer trajectories having more complex shapes.

2.3 Exploiting Fourier Transform for Spatial Quincunx Lattices Based Clustering

In this section we exploit non-separable transforms in order to effectively manage two dimensional trajectories. Non-separable transforms allow us to consider the whole trajectory taking into account both dimensions in the computation thus avoiding any approximation due to mono-dimensional transform composition. We will exploit them in a suitable way in order to catch similarity among trajectories. The first step to be performed for any mathematical transform is to define the basis function and the features of the search space where data reside. After a deep investigation of the trajectories being analyzed we assumed that trajectories resides on a *Quincunx Lattice*[1]. This will allow us to treat with a proper detail even trajectories laying close to the edges of the search space.

It is well known from algebra that there is a connection between polynomial algebra and signal transforms, in particular we briefly recall that the most important algebraic polynomials are *Chebyshev* polynomials that are defined using a recurrence equation:

$$C_n(x) = 2 \cdot x \cdot C_{n-1}(x) - C_{n-2}(x), n \geq 2$$

whose solution depends on the initial condition (i.e. C_0 and C_1).

Before defining the proper transform we briefly recall the notion of polynomial algebra that naturally induces the notion of transform. An algebra $\mathcal{A}$ is a vector space closed under multiplication and for which the distributivity law holds. As an example consider the set of polynomials $\mathcal{A} = \mathcal{C}[x]$ in one variable. Let $p(x)$ a given polynomial s.t. $p(x) \in \mathcal{C}[x]$, the set of polynomials of degree less than $deg(p)$ allowing addition and multiplication modulo p induce the following algebra:

$$\mathcal{A} = \mathcal{C}[x]/p(x) = \{q(x) s.t. deg(q) < deg(p)\}$$

$\mathcal{A}$ is called a polynomial algebra. As a vector space, $\mathcal{A}$ has dimension $dim(\mathcal{A}) = deg(p)$. Further, $\mathcal{A}$ is obviously generated by x, since every element of A is a polynomial in x. Similarly to the one variable case, we can define polynomial algebras in more variables. For our case (two variables) as example and define a polynomial algebra as[2]:

$$\mathcal{A} = \mathcal{C}[x]/(p(x,y), q(x,y))$$

Once defined the polynomial algebra the polynomial transform is defined by an isomorphic decomposition of $\mathcal{A}$ w.r.t. a given basis. In particular in our case

[1] It is always possible to find a basis that allow this representation for the search space.

[2] Note that here we need to compute modulo two polynomials to ensure that the dimension of A is finite.

given the degree of p (say it n) and q (say it m) we can construct a matrix $\mathcal{P}_{b,a} = [p_l(a_k; b_k)]; 0 < l, k < m \cdot n$. The matrix so far defined is the Fourier Transform for the given algebra(more details can be found in [2]). If we consider the basis B defined using the *Chebyshev* polynomials (i.e. $B = (T_i(x), T_j(y)); 0 < i, j < n$) we obtain the quincunx lattice shown in Fig. **??**. After properly constructed the algebra (details are not relevant for this paper see [12]) the spatial signal spectrum is:

- $u_k = \cos(\frac{k+1/2}{n/2} \cdot \pi), 0 < k < n/2$
- $v_l = \cos(\frac{l+1/2}{n/2} \cdot \pi), 0 < l < n/2$
- $w_{k,l,\pm} = \pm\frac{1}{2}\sqrt{(1+u_k)\cdot(1+v_l)}$

Once obtained the spatial signal spectrum we can easily define the distance between two trajectories by considering their spectrum. In particular, given two trajectories Tr_1 and Tr_2 and their spectrum $Q(Tr_1)$,$Q(Tr_2)$, for u_k and v_l coefficient we compute their distance as $d_{\widehat{\alpha\beta}} = \arccos(u1_k) - \arccos(u2_k)$ and $d_{\widehat{\gamma\delta}} = \arccos(v1_l) - \arccos(v2_l)$. For each pair of $w_{k,l,\pm}$ we compute the distance as $d_w = \sqrt{w1^2_{k,l,\pm} - w2^2_{k,l,\pm}}$. Finally, we define the *Quincunx* based distance as:

$$dist_Q(Tr_1, Tr_2) = \sqrt{\sum_{k=1}^{n^2} d_w(k)^2} \cdot \cos(\max\{\mu(d_{\widehat{\alpha\beta}}), \mu(d_{\widehat{\gamma\delta}})\})$$

where $\mu(d_{\widehat{\alpha\beta}})$ is the average angle distance between each pair of u_k and $\mu(d_{\widehat{\gamma\delta}})$ is the mean between each pair of v_l. The distance so far defined is able to catch dissimilarity between trajectories since it consider the max angular distance between the two direction (the cos argument) while taking into account the overall extension of the spatial signal. We point out that the choice to exploit a *Quincunx* lattice allows us to better cover the whole search space (since it is a directed lattice) and exploiting a non-separable transform (i.e. defined on both dimensions not as a simple composition of two mono-dimensional transforms) we do not introduce any approximation error. We compute the clustering of the lifted trajectories by running k-means++, a stable k-means improvement that has been proposed in [1], using a distance-based probabilistic algorithm $O(log(k))$ that makes it competitive with optimal clustering and avoid the initial cluster assignment problem, we do not report here the code for this step due to space limitations. Of course the distance effectiveness rely on the ability to catch the main features of the trajectories. As will be shown in the experimental section we performed several experiments to asses the validity of the approach and the results obtained are quite convincing.

3 Experimental Results

In this section, we present some of the several experiments we performed to assess the effectiveness of the proposed approach in clustering trajectories. To this

purpose, a collection of tests is performed, and in each test some relevant groups of homogeneous trajectories (*trajectory classes*) are considered. The direct result of each test is a similarity matrix representing the degree of similarity for each pair of trajectories in the data set. The evaluation of the results relies on some *a priori* knowledge about the trajectories being used that was obtained by domain expert. We performed several experiments on a wide variety of real datasets. We present here only the results obtained applying the transform to the lifted trajectories due to space limitations. We compare our approach with the one presented in [6]. More in details we analyzed the following data:

1) *School Bus*: it is a dataset consisting of 145 trajectories of 2 school buses collecting (and delivering) students around Athens metropolitan area in Greece for 108 distinct days[3]; 2) *Animals*: it is a dataset containing the major habitat variables derived for radio-telemetry studies of elk, mule deer, and cattle at the Starkey Experimental Forest and Range in northeastern Oregon[4].

In order to perform a simple quantitative analysis we produce for each test a similarity matrix, aimed at evaluating the resulting intra-cluster similarities (i.e., the average of the values computed for trajectories belonging to the same cluster), and to compare them with the inter-cluster similarities (i.e., the similarity computed by considering only trajectories belonging to different classes). To this purpose, values inside the matrix can be aggregated according to the cluster of membership of the related elements: given a set of trajectories belonging to n prior classes, a similarity matrix S about these trajectories can be summarized by a $n \times n$ matrix CS, where the generic element $CS(i,j)$ represents the average similarity between cluster i and cluster j.

$$CS(i,j) = \begin{cases} \frac{\sum_{x,y \in C_i, x \neq y} DIST(x,y)}{|C_i| \times (|C_i|-1)} & \text{iff } i = j \\ \frac{\sum_{x \in C_i, y \in C_j} DIST(x,y)}{|C_i| \times |C_j|} & \text{otherwise} \end{cases}$$

where $DIST(x,y)$ is the chosen distance metric ($TRACLUS$ or our Fourier based).

The higher are the values on the diagonal of the corresponding CS matrix w.r.t. those outside the diagonal, the higher is the ability of the similarity measure to separate different classes. In the following we report a similarity matrix for each dataset being considered, as it will be clear it has proven that our techniques are quite effective for clustering the datasets being considered. For comparison purposes we summarize also the results obtained by using the technique described in [6] and reorganized in a similarity matrix(using their distance definition in our formula for matrix elements). In particular for the Fourier based approach we performed a huge number of tests to assess the validity of the proposed encodings and for the sake of clarity we show in this section the best ones for comparing it. We point out that due to the different nature of our technique (we refer to it in the figures as $Fourier_{2D}$) w.r.t. $TRACLUS$ in [6] (it depends

[3] Available at http://www.rtreeportal.org
[4] http://www.fs.fed.us/pnw/starkey/data/tables/index.shtml

on some parameters) the comparison main goal is to assess the validity of the approaches. The comparison is done using for $TRACLUS$ the best values for parameters ε and $MinLns$ obtained after a severe tuning guided by the insight provided in [6].

3.1 Measuring Effectiveness

School Bus. For this dataset our prior knowledge is the set of trajectories related to the two school buses. We present the results using two classes but we point out that our technique is able to further refine the cluster assignment identifying the micro-clusters represented by common sub-trajectories. As it is easy to see in Figure 2 both the methods perform really well however $Fourier_{2D}$ clearly outperforms $TRACLUS$.

$TRACLUS$	Bus 1	Bus 2
Bus 1	0.9790	0.8528
Bus 2	0.8528	0.9915

$Fourier_{2D}$	Bus 1	Bus 2
Bus 1	1	0.6250
Bus 2	0.6250	1

Fig. 2. $TRACLUS$ and $Fourier_{2D}$ results for *Bus* dataset

The presence of several turns made by the buses makes the feature of $Fourier_{2D}$ more suited for this dataset since it exactly recognizes the class each trajectory belongs to.

Animals. In this case we considered as a class assignment the different trajectories traversed by elk, mule deer, and cattle. In this case there were 3 main classes as it is shown in Figure 3. In this case $Fourier_{2D}$ still outperforms $TRACLUS$.

$Fourier_{2D}$	elk	mule deer	cattle
elk	0.9885	0.7439	0.7108
mule deer	0.7439	0.9899	0.7223
cattle	0.7108	0.7223	0.9874

$TRACLUS$	elk	mule deer	cattle
elk	0.9986	0.7759	0.7055
mule deer	0.7759	0.9889	0.7566
cattle	0.7055	0.7566	0.9920

Fig. 3. $TRACLUS$ and $Fourier_{2D}$ results for *Animal* dataset

As we can see, differences among the various classes are marked with higher precision by $Fourier_{2D}$. This is mainly due to the fact that our approach is quite discriminative as it considers both modulo and angular distances.

4 Conclusion

In this paper we addressed the problem of detecting clusters in trajectory data. The techniques we have proposed are mainly based on the idea of representing

a trajectory with a smaller version using lifting schemes. Thereby, the similarity between two trajectories can be computed by analyzing their Fourier transforms in the two-dimensional case. Experimental results showed the effectiveness of the approach in detecting common clusters for trajectories.

References

1. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding (2007)
2. Chihara, T.S.: An Introduction to Orthogonal Polynomials. Gordon and Breach (1978)
3. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD 1996 (1996)
4. Giannotti, F., Nanni, M., Pinelli, F., Pedreschi, D.: Trajectory pattern mining. In: KDD 2007, pp. 330–339 (2007)
5. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2000)
6. Lee, J.-G., Han, J., Whang, K.-Y.: Trajectory clustering: a partition-and-group framework. In: SIGMOD 2007 (2007)
7. Lee, J.G., Han, J., Li, X.: Trajectory outlier detection: A partition-and-detect framework. In: ICDE 2008, pp. 140–149 (2008)
8. Lee, J.G., Han, J., Li, X., Gonzalez, H.: TraClass: trajectory classication using hierarchical region-based and trajectory-based clustering. PVLDB, 1(1) (2008)
9. Liu, Y., Chen, L., Pei, J., Chen, Q., Zhao, Y.: Mining frequent trajectory patterns for activity monitoring using radio frequency tag arrays. In: PerCom, pp. 37–46 (2007)
10. Lloyd, S.: Least squares quantization in pcm. IEEE TOIT, 28 (1982)
11. Press, W.H., et al.: Numerical Recipes in C++. Cambridge University Press, Cambridge (2001)
12. Puschel, M., Rotteler, M.: Fourier transform for the directed quincunx lattice. In: ICASSP (2005)
13. Taubman, D., Secker, A.: Lifting-based invertible motion adaptive transform (limat) framework for highly scalable video compression. IEEE Transactions on Image Processing 12(12), 1530–1542 (2003)
14. Wang, W., Yang, J., Muntz, R.R.: Sting: A statistical information grid approach to spatial data mining. In: VLDB 1997, pp. 186–195 (1997)
15. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: An efficient data clustering method for very large databases. In: SIGMOD 1996, pp. 103–114 (1996)

Knowledge Discovery about Web Performance with Geostatistical Turning Bands Method

Leszek Borzemski and Anna Kamińska-Chuchmała

Institute of Informatics, Wroclaw Uniwersity of Technology,
Wroclaw, Poland
{leszek.borzemski,anna.kaminska-chuchmala}@pwr.wroc.pl

Abstract. The work presents a proposal of the application of the geostatistical simulation - the Turning Bands method, to acquire knowledge about network throughput. The data base was created on the basis of Multiagent Internet Measurement System MWING. In the research the connections between an agent in Wrocław and European servers were considered. The preliminary structural analysis of the data which are necessary to use the Turning Bands method was conducted. Next a spatial forecast of the total time of downloading data from Web servers with a two-week time advance was calculated. The analysis of server activity on a particular week day was conducted for a period of a few weeks in selected time intervals.

Keywords: Knowledge discovery, Web performance, server performance prediction, spatial forecast, geostatistics, Turning Bands method.

1 Introduction

In Knowledge Engineering it is very important to use the most effective information which can contribute to Internet performance increase. The authors made an attempt to use the spatial forecast of Internet performance with a two-week time advance. They used the geostatistical simulation Turning Bands for this purpose. So far these methods have been used in many disciplines, for example: geology to assess the size of deposits [8], natural environment protection to assess degree of pollution [13], oceanography to assess phytoplankton dissemination in oceans [5] and even in power engineering where one of the authors tested the electric performance of distribution and transmission networks [9], [10], [11], [12]. A great advantage of these methods is the possibility to make area-time forecasts in which the minimum amount of input information is required and at the same time it takes into account the geographical location of Web servers and the total download time of a given resource. Such forecast information is required, e.g. when one has to download information resource which is available on many Internet nodes at various geographical locations. Knowing the throughput forecast and transfer capacity from these nodes to our localisation will allow to choose the node from which one can receive the required resource in the shortest time [2].

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 581–590, 2011.

2 Turning Bands Method

The first step in geostatistical simulation is modelling the variables process and next the simulation of these variables using an elementary grid. The Turning Bands method created by Matheron is a stereologic tool used for the reduction of multidimensional simulation to a one-dimensional one [15], [16]. The idea of the Turning Bands method is the reduction of random simulation of a Gaussian function with covariance C to the simulation of an independent stochastic process with covariance C_θ.
Let $(\theta_n, n \in \mathbb{N})$ be a sequence of directions $\mathbb{S}_d^+$ and let $(X_n, n \in \mathbb{N})$ be a sequence independent stochastic process of covariance C_θ. Random function:

$$Y^{(n)}(x) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k(< x, \theta_k >), \qquad x \in \mathbb{R}^d, \tag{1}$$

assumes covariance equal to:

$$C^{(n)}(h) = \frac{1}{n} \sum_{k=1}^{n} C_{\theta_k}(< h, \theta_k >). \tag{2}$$

Algorithm of the Turning Bands Method

1. Input data transformation using Gaussian anamorphosis.
2. Selection of directions $\theta_1, ..., \theta_n$ so that $\frac{1}{n}\sum_{k=1}^{n} \delta_{\theta_k} \approx \varpi$.
3. Generation of standard, independent stochastic processes $X_1, ..., X_n$ with covariance functions $C_{\theta_1}, ..., C_{\theta_n}$.
4. Calculating $Y^{(n)}(x) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k(< x, \theta_k >)$ for each $x \in D$.
5. Kriging estimation $y^*(x) = \sum_c \lambda_c(x) y(c)$ for each $x \in D$.
6. Simulation of a Gaussian random function with average value equal to 0, covariance C in domain D on condition points. Let $(z(x), x \in D)$ and $(z(c), c \in C)$ be the obtained values.
7. Kriging estimation $z^*(x) = \sum_c \lambda_c(x) z(c)$ for each $x \in D$.
8. Generating the result $(y^*(x) + z(x) - z^*(x), x \in D)$.
9. Return to original data with Gaussian anamorphosis.

Below there is more information about particular steps of algorithm, while the whole Turning Bands method and conditional simulations are discussed in more detail in [14].

Step 1. In the transformation of Gaussian type variable y into new variable z of arbitrary distribution the anmorphosis function is used. Empirical Gaussian anamorphosis is represented by infinite series of Hermite polynomials [17]:

$$\varphi(y) = \sum_{k=0}^{\infty} \frac{\varphi_k}{k!} H_k(y). \tag{3}$$

Finally the limited anamorphosis is used in the random diagram of Gaussian curve to obtain a random graph of the original random variable.

Step 2. Simulation of covariance C_3 is obtained by summing up covariance C_1 with simulation projection of a given number of lines, covariances C_1. Particular lines are called *turning bands*.

Step 3. The basic steps in the determination of one-dimensional covariances is determining spectral measure χ from C, and next spectral measure χ_θ from C_θ and taking its Fourier transform.

Because of the fact that covariance C is isotropic, it is possible to significantly simplify the calculations where:

$$C(h) = C_d(|h|), \tag{4}$$

for certain scalar function C_d defined on $\mathbb{R}^+$.

In this case all C_θ are equal to individual covariance function, e.g. C_1. Matheron obtained the dependence between C_1 and C_d:

$$C_d(r) = 2\frac{(d-1)\omega_{d-1}}{d\omega_d}\int_0^1 (1-t^2)^{\frac{d-3}{2}} C_1(tr)dt, \tag{5}$$

where ω_d stands for d-dimensional volume of the unit ball in $\mathbb{R}^d$. If $d = 3$, the formula is reduced to:

$$C_3(r) = \int_0^1 C_1(tr)dt, \tag{6}$$

or equivalently:

$$C_1(r) = \frac{d}{dr}(rC_3(r)). \tag{7}$$

Knowing variogram and variance *C(0)* for a stationary random function Z is tantamount to knowing its covariance. Thus to determine a covariance function for each of the variables, a variogram function is calculated, it is defined for any vector h as:

$$\gamma(h) = \frac{1}{2}E((Z(x+h) - Z(x))^2). \tag{8}$$

During the forecasting of Web performance with time advance the direction variogram is determined. In this case the direction along the time axis (for 90° direction) is selected, in this direction calculations are made for various distances $|h|$.

Step 4. There is a wide variety of simulation methods for stochastic processes with a given covariance function C_1. The most common methods are [14]: spectral, dilution, migration.

Step 5. Let Y be a stationary Gaussian random function $\in \mathbb{R}^d$ with average value 0, variance equal to 1 and covariance function C. The goal is to conducting simulation Y satisfying the conditions: $Y(c) = y(c)$ for each c in finite subset $C \in \mathbb{R}^d$.

Simple kriging estimation (linear regression) $Y(x)$ to $Y(c)$ is a linear combination:

$$Y^*(x) = \sum_{c \in C} \lambda_c(x) Y(c), \tag{9}$$

which minimises the mean square error $E(Y^*(x) - Y(x))^2$.

Coefficients λ_c are the solutions of the system of linear equations:

$$\sum_{c' \in C} \lambda_{c'} C(c, c') = C(c, x) \quad \forall c \in C. \tag{10}$$

The mean square error, i.e. the kriging variance is then:

$$E(Y^*(x) - Y(x))^2 = 1 - \sum_{c \in C} \lambda_c C(c, x). \tag{11}$$

For the calculations of the system of linear equations in kriging we need to know estimated neighbourhood point. There are two types of neighbourhood: unique and moving.

A characteristic feature of unique neighbourhood is single calculation of the inverse kriging matrix. It remains the same for all considered points. Thus the type of unique neighbourhood can be computed much more quickly than moving neighbourhood.

In case of moving neighbourhood if the location of point x is known, the closest point in a sphere or ellipsoid are selected to be correlated with the considered point x. The selected points are subject to a sequence algorithm which takes into account such criteria as: rotation, search ellipsoid (it defines the maximum distance along the main axes U, V, W after rotation), anisotropy, minimum number of points considered in the range search, etc. In the case when the neighbourhood distance is established as anisotropic, the search ellipsoid is selected:

$$d = \sqrt{(a_u d_u)^2 + (a_v d_v)^2 + (a_w d_w)^2}, \tag{12}$$

where d_u, d_v and d_w correspond with distance along the axis of the new coordinate system and:

$$a_u = \frac{maximum\ distance\ along\ u}{d_{max}}, \tag{13}$$

$$a_v = \frac{maximum\ distance\ along\ v}{d_{max}}, \tag{14}$$

$$a_w = \frac{maximum\ distance\ along\ w}{d_{max}}, \tag{15}$$

where d_{max} is the largest maximum distance along axis u, v and w.

Step 6. Simulation of a Gaussian random function with average value equal to 0, covariance C in domain D on condition points. Values obtained as a result of calculations: $(z(x), x \in D)$ and $(z(c), c \in C)$.

Step 7. Kriging estimation for each $x \in D$:

$$z^*(x) = \sum_c \lambda_c(x) z(c). \tag{16}$$

Step 8. Finally as a result of a conditional simulation a random function is obtained:

$$W(x) = y^*(x) + z(x) - z^*(x), \tag{17}$$

where $x \in D$.

Step 9. Return to original data with Gaussian anamorphosis.

3 Preliminary Data Analysis

To create the discussed database, we used measurement data multiagent system MWING [3], [4] obtained from the agent located in Wrocław who was the main node belonging to academic campuses. The measured parameters referred to downloading a copy of a text document from many www servers located in Europe.

The measurements encompassed the period between 1st and 28th February 2009 and they were taken once a week on Monday at fixed times: 06:00 a.m., 12:00 a.m. and 6:00 p.m. The input database necessary for calculations contains the information about server (node) geographical location with which the Wrocław agent connected, the total downloading time and the time of taking the measurement.

Table 1. Elementary statistical parameters of server performance on the Internet between 1-28.02.2009 at 6:00 a.m

Minimum value X_{min} [s]	**Maximum value X_{max}** [s]	**Average value X** [s]	**Standard deviation S** [s]	**Variability coefficient V** [%]	**Skewness coefficient G**	**Kurtosis coefficient K**
0.12	7.68	0.77	1.11	144.16	4.37	24.85

Elementary statistics of web performance on the considered servers are presented in table 1. Taking into account the minimum and maximum values, a rather large data range is observed. Moreover the high value of standard deviation and the coefficient of variation which is 144% confirms the process variation. The high value of both the skewness coefficient and kurtosis indicate big right side asymmetry of performance distribution on the Internet.

Parameters referring to the elementary performance statistics in the Internet are presented in table 2. The minimum value is 0.16 s and the maximum value is 5.78 s which confirms the wide range of performance values. Moreover the high value of the coefficient of variation and standard deviation indicates significant differentiation and dispersion of the considered data. However, the skewness

Table 2. Elementary statistical parameters of server performance on the Internet between 1-28.02.2009 at 12:00 a.m

Minimum value X_{min} [s]	Maximum value X_{max} [s]	Average value X [s]	Standard deviation S [s]	Variability coefficient V [%]	Skewness coefficient G	Kurtosis coefficient K
0.16	5.78	1.18	1.33	112.71	1.98	6.12

Table 3. Elementary statistical parameters of server performance on the Internet between 1-28.02.2009 at 6:00 p.m

Minimum value X_{min} [s]	Maximum value X_{max} [s]	Average value X [s]	Standard deviation S [s]	Variability coefficient V [%]	Skewness coefficient G	Kurtosis coefficient K
0.13	16.75	1.04	2.12	203.85	6.03	43.34

coefficient and kurtosis values indicate that the distribution of the considered web performances should show similarity to a symmetrical distribution.

Table 3 presents elementary statistics of web performances on the considered servers. The minimum and maximum values indicate a big dispersion of data values. Additionally, the high value of the standard deviation and the coefficient of variation, above 200%, confirms high variation of the considered process. The high values of both the skewness coefficient and kurtosis indicate big right side asymmetry of web performance distribution.

4 Structural Analysis of Data

The next step after the preliminary data analysis and Gaussian anamorphosis calculation (which are not discussed here in detail due to their extensiveness) is modelling a theoretical variogram function. During the variogram model approximation, the nuggets effect function was used to consider web performance on Mondays at 6:00 a.m. A directional variogram was calculated along the time axis (for 90° direction). The distance class for this variogram was 1.31 km. Figure 1 as example presents a directional variogram approximated by the theoretical model of the nuggets effect. The variogram function indicates a gentle falling trend.

The next two directional variograms of web performance were approximated by the theoretical model of the nuggets effect and J-Bessel at 12:00 a.m. and 6:00 p.m.

5 3D Forecasting of Web Server Performance Using the Turning Bands Method

The forecast model used to predict the total time of resource download from the Internet was the above presented variogram models depending on the forecasted

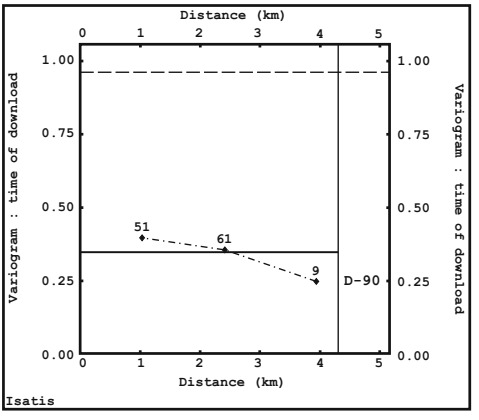


Fig. 1. Directional variogram along the time axis for web performance on Internet nodes at 6:00 a.m. on Mondays, approximated by the theoretical model of the nuggets effect

hour on Monday. The Gaussian anamorphosis was determined for 100 Hermite polynomials. In the simulation the moving neighbourhood type was adopted where the search ellipsoid was 23.15 km for all three directions in the case of web performance at 6:00 a.m., for 12:00 a.m. the ellipsoid was 11.63 km and for 6:00 p.m. the search ellipsoid was 22.92km. The forecast of the download time was determined on the basis of 100 simulation realisations. In the simulation the punctual type was used. 3D forecast was calculated with a two-week time advance, i.e. it encompassed the period between 1st and 14th March 2009. The tables below 4, 5 and 6 present global statistics of the forecasted values of node performance in a computer network for a period of 2 weeks (two subsequent Mondays), the tables present the data for 6:00 a.m., 12:00 a.m. and 6:00 p.m., respectively.

Table 4. Global statistics of the forecasted values of web server performance on the Internet with a two-week time advance, calculated using the Turning Bands simulation method, for 6:00 a.m. on Monday

Geostatistical parameter	Average value Z_s [s]	Maximum value $Z_{s,max}$ [s]	Minimum value $Z_{s,min}$ [s]	Variance S^2 $[s]^2$	Standard deviation S [s]	Variance coefficient V [%]
Average forecasted value Z_s	0.59	2.56	0.26	0.05	0.21	35.59
Forecast standard deviation σ_s	0.35	2.32	0.03	0.08	0.27	77.14
Maximum forecasted value $Z_{s,max}$	2.63	7.68	0.37	3.58	1.89	71.86
Minimum forecasted value $Z_{s,min}$	0.24	0.60	0.12	0.01	0.08	33.33

Table 5. Global statistics of the forecasted values of web server performance on the Internet with a two-week time advance, calculated using the Turning Bands simulation method, for 12:00 a.m. on Monday

Geostatistical parameter	Average value Z_s [s]	Maximum value $Z_{s,max}$ [s]	Minimum value $Z_{s,min}$ [s]	Variance S^2 $[s]^2$	Standard deviation S [s]	Variance coefficient V [%]
Average forecasted value Z_s	1.51	4.47	0.26	0.47	0.69	45.70
Forecast standard deviation σ_s	1.47	2.35	0.06	0.22	0.47	31.97
Maximum forecasted value $Z_{s,max}$	5.52	5.78	0.58	0.42	0.65	11.78
Minimum forecasted value $Z_{s,min}$	0.19	0.96	0.16	0.00	0.04	21.05

Table 6. Global statistics of the forecasted values of web server performance on the Internet with a two-week time advance, calculated using the Turning Bands simulation method, for 6:00 p.m. on Monday

Geostatistical parameter	Average value Z_s [s]	Maximum value $Z_{s,max}$ [s]	Minimum value $Z_{s,min}$ [s]	Variance S^2 $[s]^2$	Standard deviation S [s]	Variance coefficient V [%]
Average forecasted value Z_s	1.83	6.93	0.21	1.17	1.08	59.02
Forecast standard deviation σ_s	2.56	6.87	0.13	2.94	1.71	66.80
Maximum forecasted value $Z_{s,max}$	12.56	16.75	0.67	35.92	5.99	47.69
Minimum forecasted value $Z_{s,min}$	0.17	1.12	0.13	0.01	0.10	58.82

On the basis of the forecasts of web servers performance presented in tables 4, 5 and 6, very interesting phenomena can be observed. Namely the forecasts for 6.00 p.m. have the biggest data range taking into account the disparity between the minimum value and the maximum one, and the standard deviation which is the highest for this time in comparison with the mean value.

However, regardless of these not very favourable values, it turns out that the forecast accuracy is the highest for this time, in the range of 60%. The lowest was obtained for 12:00 a.m. Nevertheless it may be said that this forecast is a quite good representation of some web nodes. The average, absolute percentage, relative forecast error *ex post* for a few selected web nodes at various times is presented below in table 7.

Analyzing the results presented in table 7, one can observe that the forecast prepared using the Turning Bands method represents the download time quite well. Unfortunately due to the big span of Web servers in the examined area and high measurement data dispersion, not all of the results were characterised with

Table 7. Few exemplary results of spatial forecast of Web servers performance calculated with Turning Bands simulation method

URL address	Country/city	Measurement date and time	Download time [s]	Forecasted download time [s]	Forecasted error *ex post* [%]
wigwam.sztaki.hu /rfc/rfc1945.txt	Hungary /Budapest	2 March 2009, 06:00	0.31	0.33	6.71
omega.di.unipi.it/local /home/rfc/rfc1945.txt	Italy /Pisa	9 March 2009, 06:00	0.60	0.65	7.86
www-uxsup.csx.cam.ac.uk /pub/doc/rfc/rfc1945.txt	United Kingdom /Cambridge	2 March 2009, 12:00	0.78	0.71	8.42
ftp.univie.ac.at/netinfo /rfc/rfc1945.txt	Austria /Vienna	9 March 2009, 12:00	0.67	0.72	6.88
curl.nedmirror.nl /rfc/rfc1945.txt	Netherlands /Eindhoven	2 March 2009, 18:00	0.44	0.43	1.97
paginas.fe.up.pt/~jvv /net/rfc1945.txt	Portugal /Porto	9 March 2009, 18:00	0.66	0.69	4.82

good accuracy of prediction results. The final effect of the forecast calculations is presented in figure 2, it is a base map with two-week time advance of the download time from the Internet. The size of the cross on the map corresponds with the download time from a given web server (in the legend the download time is given in seconds).

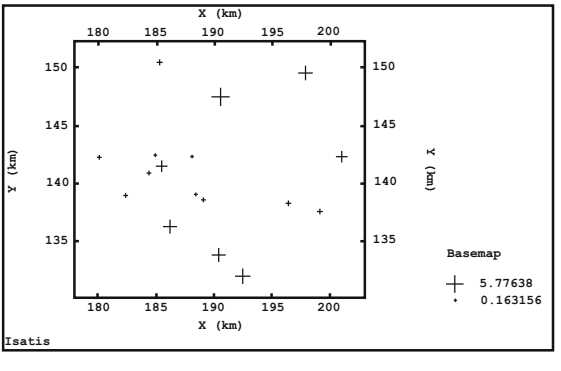


Fig. 2. Base map of download time values from the Internet for 12:00 a.m. on Monday

6 Summary

On the basis of the conducted calculations, which were one of the first ones of this type, referring to discovery knowledge connected with network throughput using the geostatistic simulation method Turning Bands one can conclude that the application of these methods was justified. A realistic possibility of using geostatic methods in yet another, new discipline which is computer science is

outlined. However, there is a need to work on the improvement of forecast accuracy. Internet performance should be analysed also in other areas which were preliminarily discussed in [1]. Moreover various measurement data and prediction lengths should be considered. However, the next step should be an attempt to use other geostatic methods.

References

1. Borzemski, L., Kaminska-Chuchmala, A.: 3D Web Performance Forecasting Using Turning Bands Method. CCIS, vol. 160, pp. 102–113. Springer, Heidelberg (2011)
2. Borzemski, L.: The experimental design for data mining to discover web performance issues in a Wide Area Network. Cybernetics and Systems 41, 31–45 (2010)
3. Borzemski, L., Cichocki, Ł., Kliber, M., Fras, M., Nowak, Z.: MWING: a multiagent system for Web site measurements. In: Nguyen, N.T., Grzech, A., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2007. LNCS (LNAI), vol. 4496, pp. 278–287. Springer, Heidelberg (2007)
4. Borzemski, L., Cichocki, Ł., Kliber, M.: Architecture of multiagent internet measurement system MWING release 2. In: Håkansson, A., Nguyen, N.T., Hartung, R.L., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2009. LNCS, vol. 5559, pp. 410–419. Springer, Heidelberg (2009)
5. Inizan, M.: Geostatistical Validation of a Marine Ecosystem Model Using In Situ Data, Tech. Rep. S-435 Centre de Geostatistique, Ecole des Mines de Paris (2002)
6. Isaaks, E.H., Srivastava, R.M.: An Introduction to Applied Geostatistics. Oxford University Press, New York (1989)
7. Isatis Software Manual, Geovariances & École des Mines de Paris (2004)
8. Journel, A.G., Huijbregts, C.J.: Mining Geostatistics. Academic Press, NY (1978)
9. Kaminska-Chuchmala, A., Wilczynski, A.: Application simulation methods to spatial electric load forecasting. Rynek Energii 1(80), 2–9 (2009) (in Polish)
10. Kaminska-Chuchmala, A., Wilczynski, A.: 3D electric load forecasting using geostatistical simulation method Turning Bands, vol. XVI (series B 215), pp. 41–48. The works of Wrocław Scientific Society, Wroclaw (2009)
11. Kaminska-Chuchmala, A., Wilczynski, A.: Analysis of different methodological factors on accuracy of spatial electric load forecast performed with Turning Bands method. Rynek Energii 2(87), 54–59 (2010) (in Polish)
12. Kaminska-Chuchmala, A., Wilczynski, A.: Spatial electric load forecasting in transmission networks with Sequential Gaussian Simulation method. Rynek Energii 1(92), 35–40 (2011) (in Polish)
13. Lajaunie, C.: A geostatistical approach to air pollution modelling. In: Verly, G., et al. (eds.) Geostatistics for Natural Resources Characterization. NATO ASI Series C-122, pp. 877–891. Reidel, Dordrecht (1984)
14. Lantuejoul, C.: Geostatistical Simulation. Models and Algorithms. Springer, Heidelberg (2002)
15. Matheron, G.: Quelques aspects de la montée, Int. Rep. N-271, CMM
16. Matheron, G.: The intrinsic random functions and their applications. Adv. Appl. Prob. 5, 439–468 (1973)
17. Wackernagel, H.: Multivariate Geostatistics: an Introduction with Applications. Springer, Berlin (2003)

Context Change Detection for Resource Allocation in Service-Oriented Systems

Piotr Rygielski and Jakub M. Tomczak

Institute of Computer Science, Wrocław University of Technology,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{piotr.rygielski,jakub.tomczak}@pwr.wroc.pl

Abstract. In this paper, the problem of detecting the major changes in the stream of service requests is formulated. The change of stream component varies over time and depends on, e.g., a time of a day. The underlying cause of the change is called a context. Hence, at each moment there exists a probability distribution determining the probability of requesting the system service conditioned by the context. The aim is to find such a moment in which the distributions change. To solve that problem dissimilarity measures between two probability distributions are given. Nevertheless, detecting every change is not interesting but only long-lasting changes because of the costs of the service system resources reallocation. Therefore, in the proposed algorithm an issue of sensitivity to temporary changes detection is considered.

Keywords: datastream, dissimilarity measure, sliding window estimation.

1 Introduction

Recently, an increasing interest of service-oriented systems (SOA) can be noticed [6]. SOA consists of elements called *services* that are network-distributed on different computational nodes [8,9]. However, service-oriented architectures cause new problems connected with resource sharing and efficient resource management [10]. In the real-life situations in which streams of requests are time-varying and dependent on circumstances connected with e.g. time of a day, or changing number of users of the system, resources given to services have to be managed in an efficient way. It means that resources should be allocated in such a way that all operations and consequently – services are accomplished successfully in an appropriate time. However, due to the changes of the requests' streams the resources have to be reallocated [12] during the system life. Nevertheless, reallocation process consumes an additional time and operational resources and that is why it cannot be carried out at any change of the stream of requests. Thus, we deal with a tradeoff between rare reallocations but worst resource usage and often reallocation but higher costs associated with the resource management.

In this work we address the problem of the context change detection in service-oriented systems. We state the problem and present an algorithm for change detection in streams of requests. The main idea is to apply a dissimilarity measure

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 591–600, 2011.

between two probability distributions. However, the key challenge is to detect only long-lasting changes and no temporary changes because we would like to avoid too often resource reallocation operations. Therefore, we provide a detailed description of the change detection algorithm and present results of the simulation study.

2 Related Works

In the literature there are different methods for detecting changes. They could be divided into heuristics and theoretical approaches. The first group includes methods that are based on classifiers [15,21]. If the classification accuracy of the classifier drops significantly a context change is reported. It was used e.g. in the FLORA-family of algorithms [21]. Moreover, the weighted forgetting and window size management methods [11,15] could be included to this group as well.

The second group uses the theoretical aspects, especially taken from statistics. These approaches compare two probability distributions to detect a change. However, they differ in assumptions. For example in computational theory-based approaches [4,15] it is assumed that changes are rather permanent and gradual. Then an upper bound of the shift could be given and a number of recent observations for the processing may be calculated. Unfortunately, in practice sometimes we deal with abrupt changes and such methodology fails (e.g. in the anomaly detection [3,16]). Then a dissimilarity measure of two probability distributions to detect a change could be applied. In many applications the Kullback-Leibler divergence is used [3,16,19], or entropy-based measures [19,20]. Sometimes more theoretical measures are proposed [14].

3 Problem Statement

3.1 Problem Background

The service-oriented systems are systems that consist of network distributed components called *services* [8]. The single service is assumed to deliver atomic functionality and has its inputs and outputs precisely defined. We distinguish an atomic and complex service which delivers single functionality and the functionality composed of more than one functionality, respectively. Moreover, the service-oriented system is able to compose the complex service in order to deliver the service demanded by the user [8].

An atomic service is assumed to consume computational resources. Moreover, we assume that each atomic service is located within a computational node which is able to manage its resources using widely known virtualization techniques [18]. In this paper we consider the resource management task on a single computational node with atomic services installed on it.

Furthermore, there is a stream of service requests arriving to the service-oriented system. A request demands certain functionality. In order to compose

required service a composition procedure is executed which has been described in [8]. The most important result of composition is fact that proper atomic services are executed according to an execution scenario.

In order to optimize the performance of each atomic service the computational resources should be managed properly. The ratio of the available amount of the resource should be assigned to the proper virtual machine depending on the intensity of the requests stream and the complexity of the atomic service itself. The way how the complexity of the service is measured is not a concern of this paper and will be considered in future works. Denote the intensity of request stream arriving to the computational node at the moment t as $\mathbf{x}_t$ where $\mathbf{x}_t$ consists of substreams $\mathbf{x}_t = (x_t^{(1)}\ x_t^{(2)} \dots x_t^{(N)})^T$, each $x_t^{(n)}$ is dedicated to the proper atomic service installed on the considered computational node, and n denotes a number of the service, $n = 1, 2, \dots, N$.

We assume that intensities of the request stream components $(x_t^{(1)} \dots x_t^{(N)})^T$ can vary over time e.g. depending on the time of a day. Therefore, the resources should be periodically reallocated in order to maximize the quality of service delivered to the system users.

In order to determine allocation different algorithms could be applied e.g. well-known optimization algorithm called *interior-point* method [2]. Unfortunately, each operation of resource reallocation introduces a slight delay in computational node response time [12]. Thus, in order not to deteriorate the quality resource reallocation should not be performed too often.

3.2 Context Change Detection Problem Statement

In this work we assume that the stream of requests could vary in time because of a changing *context* [11,21]. By the context we understand external, unobservable circumstances e.g. a time of a day, or an increasing number of users, that affect the intensity of requested services. Formally, let us introduce a random variable that is time-depndent and that describes the requested service, r_t. It takes values in $\{1, 2, \dots, n \dots, N\}$ where n denotes the n^{th} service that was requested.

Moreover, the context is also a stochastic process, c_t, but we assume that it is a latent variable that is constant for some periods of time (see Fig. 1.b). However, the context affects the probability of requested services and thus the probability of the n^{th} requested service is denoted by $p(r_t|c_t)$. In real-life situations context changes in a *non-stationary* way [7,11]. Therefore, we deal with the non-stationary stochastic processes r_t, and c_t.

Besides, it could be said that we consider a *Hidden Markov Model* [1] (see Fig. 1.a) in which the context evolves in an abrupt way (see Fig. 1.b). We assume that we are able to observe only the service that was requested and there is no information about the context. We can only try to determine moments of change of the probability distribution, $\tau = \{t : p(r_t|c_t) \neq p(r_{t+1}|c_{t+1})\}$.

However, for our purpose, we are not interested in finding *any* change. For example detecting temporary changes leads to frequent reallocation of resources which implies high costs connected with that process. Therefore, the change should be reported only if estimated probability distributions differ significantly.

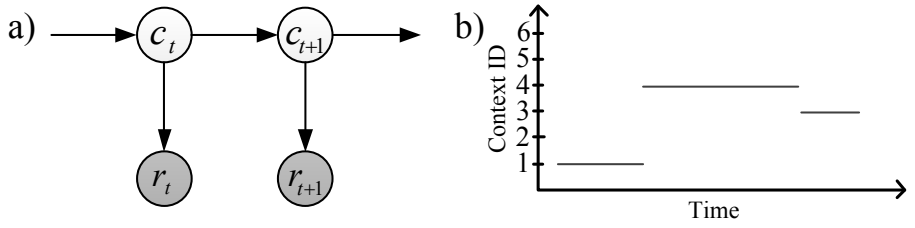


Fig. 1. (a) A probabilistic graphical model for the requesting n^{th} atomic service conditioned by a context. (b) An abrupt change of a context.

The change is measured by a certain dissimilarity measure $\mathcal{D}$. Then the change is reported if the dissimilarity value is higher than a *sensitivity factor* $\sigma \in [0, 1]$. Hence, the problem can be stated as follows:

Given:

- probability distribution at any time t, $p(r_t|c_t)$;
- the dissimilarity measure $\mathcal{D}$;
- the value of the sensitivity factor σ.

Find:

- The moments such that:

$$\tau = \Big\{ t : \mathcal{D}\big(p(r_{t-1}|c_{t-1}), p(r_t|c_t)\big) \geq \sigma \Big\}.$$

In the next section we present an algorithm for context change detection. However, to propose the algorithm two issues have to be solved. First, a dissimilarity measure has to be defined. Second, at any time we do not have probability distributions. Thus, a proper estimation technique adequate for stream of data has to be proposed.

4 Context Change Detection Algorithm

4.1 Dissimilarity Measures

In the literature a variety of dissimilarity measures are known e.g. Kullback-Leibler divergence [16], Lin-Wong divergence (modified Kullback-Leibler) [17], Bhattacharrya measure [13], Kolmogorov measure [5], entropy-based measure [20], cosine distance measure [19], and others (more measures could be found in [5]). Obviously, they differ in forms and mathematical properties.

In further considerations let assume that there is a discrete random variable $x \in \mathcal{X}$, $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$, and two probability distributions: $p(x)$, called *reference* distribution, and $q(x)$, called *model* distribution.

In this paper, based on our preliminary research and because of the discrete stochastic process r_t, we decided on four measures: *Bhattacharyya* measure, *Lin-Wong* divergence, *modified Lin-Wong* divergence, and *Kullback-Leibler* divergence. That is:

The Bhattacharyya measure is defined as follows [13]:

$$\mathcal{D}_B(p,q) = -\ln\Big(\sum_{n=1}^{N}\sqrt{p(x_n)\cdot q(x_n)}\Big). \tag{1}$$

The Lin-Wong divergence is defined as follows [17]:

$$\mathcal{D}_{LW}(p,q) = \sum_{n=1}^{N} p(x_n)\cdot\ln\Big(\frac{p(x_n)}{0.5\cdot p(x_n)+0.5\cdot q(x_n)}\Big). \tag{2}$$

The Kullback-Leibler divergence is defined as follows [16]:

$$\mathcal{D}_{KL}(p,q) = \sum_{n=1}^{N} p(x_n)\cdot\ln\Big(\frac{p(x_n)}{q(x_n)}\Big). \tag{3}$$

The modified Lin-Wong divergence uses the fact that for discrete probability distributions $\mathcal{D}_{LW}(p,q)\in[0,1]$. Then the dissimilarity measure could be defined as follows:

$$\mathcal{D}_{LW2}(p,q) = \Big(\mathcal{D}_{LW}(p,q)\Big)^2. \tag{4}$$

Remark. In the preliminary research it turned out that for the estimation technique used in this paper (see next section) the measure (4) is less sensitive to underestimations. It is easy to explain and it follows from the property of quadratic function and that $\mathcal{D}_{LW}(p,q)\in[0,1]$. The original function $\mathcal{D}_{LW}(p,q)$ is *flattened*, especially near zero, which protects from underestimations.

4.2 Probabilities Estimation Technique

To estimate the reference and model distributions the sliding window technique was applied [14,20]. The sliding window technique uses two windows, one representing L_1 of older and the other representing L_2 of recent observations (light and dark gray areas in Fig. 2) in the stream. The older instances (the older window) are chosen for estimating the reference distribution and the current data (the current window) – for the model distribution.

For the further simplicity, we set $L_1 = L_2 \equiv L$.

4.3 Detection Algorithm

The idea of the algorithm is very similar to the meta-algorithm presented in [14] and the method in [3]. First, data from datastream is taken and divided into

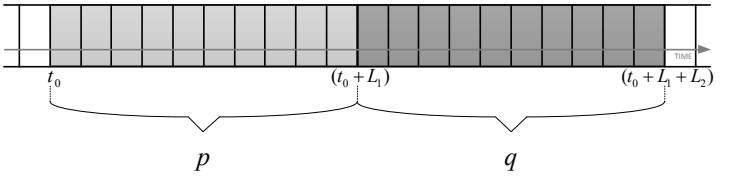


Fig. 2. An illustration of sliding windows for probabilities estimation

two sliding windows. Then, the reference and model distributions are estimated. Later, a value of the dissimilarity measure is calculated. If the value is lower than the given sensitive factor, sliding windows are moved so that all new observations are included. Otherwise the change is reported and sliding windows are set at the same moment and only the window with new observations is moved as long as it does not overlap instances from the older window. After that both windows are moved next to each other. The flow chart of the algorithm is presented in the Fig. 3.

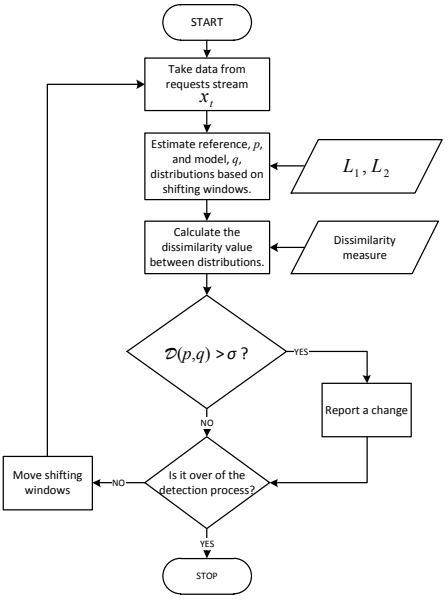


Fig. 3. The flow chart of the change detection algorithm

5 Experimental Study

5.1 Experiment Description

In order to verify the efficiency of the context change detection algorithm the simulation environment has been developed. The OMNeT++ as the simulation

platform was used. We modeled the physical machine with proper amount of virtual machines — each machine was responsible for responding to user's requests. The request stream was generated in such a way that there were three substreams present — each generated by a varying number of users. Each user is modelled as a process that repeatedly formulates a request sized from 1 to 4 kilobytes and after formulating the request the process is idle for a random period of time. One substream was assumed to be a background traffic and had almost constant intensity with small fluctuations during the simulation. The aggregated stream of requests was analyzed and then decomposed and passed to proper virtual machines for execution. Each virtual machine was assigned with initial amount of computational resources. When the context had changed, the resource allocation algorithm was executed.

In the experiment the measures presented in the Section 4.1 were used with following values of sensitivity factor (fixed after several trials): Bhattacharyya measure – $\sigma = 0.01$, Lin-Wong measure – $\sigma = 0.01$, modified Lin-Wong measure – $\sigma = 0.0001$, Kullback-Leibler measure – $\sigma = 0.015$. The sensitivity of modified Lin-Wong measure has to be much more smaller than for Lin-Wong because its value is squared. Besides, because of high sensitivity to even small changes, the value of the sensitivity factor for Kullback-Leibler has to be higher.

The size of the window was set to $L = 1second$ so that the estimation of probabilities was ensured. Moreover, a change is assumed to be correctly detected if the detection took place no later than 1 second from the actual change (no longer than the size of the current window). Otherwise it was not counted as a proper detection. To compare the algorithm with the different dissimilarity measures, following criteria were defined:

– A percentage of properly detected changes:

$$Q_g = \frac{\#\text{ of properly detected changes}}{\#\text{ of real changes}}$$

– A percentage of wrongly detected changes among all detected changes:

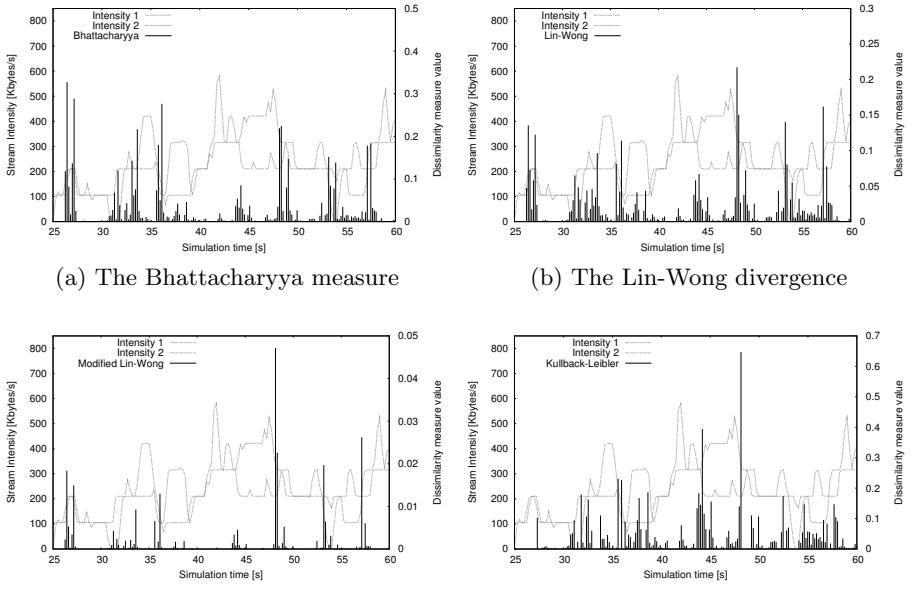
$$Q_b = \frac{\#\text{ of wrongly detected changes}}{\#\text{ of all detected changes}}$$

5.2 Results and Discussion

The simulator was run 10 times with about 25 changes of the context. Each simulation run lasted 100 seconds what was enough to process about 10 Mbytes of requests data. Results are shown in the Table 1. The best results are bold font.

Moreover, single exemplary run of the simulation environment for different dissimilarity measures are presented in Fig.s 4a, 4b, 4c, and 4d. Grey lines represent varying intensities of two classes (third class was fixed to be constant), and black bars – values of the measures.

The results show that Bhattacharyya, Lin-Wong, and modified Lin-Wong measures performed very similarly. However, they outperformed Kullback-Leibler



(a) The Bhattacharyya measure

(b) The Lin-Wong divergence

(c) Modified Lin-Wong divergence

(d) Kullback-Leibler divergence

Fig. 4. Exemplary results for four measures in a chosen simulation period

Table 1. Results for the different dissimilarity measures and considered criteria Q_g and Q_b

	Q_g				Q_b			
Batch No.	$\mathcal{D}_B$	$\mathcal{D}_{LW}$	$\mathcal{D}_{LW2}$	$\mathcal{D}_{KL}$	$\mathcal{D}_B$	$\mathcal{D}_{LW}$	$\mathcal{D}_{LW2}$	$\mathcal{D}_{KL}$
#1 (25 changes)	1	1	0.96	0.64	0.14	0.17	0.17	0.44
#2 (24 changes)	0.96	0.92	0.96	0.75	0.18	0.29	0.26	0.5
#3 (25 changes)	0.96	0.92	0.92	0.68	0.2	0.23	0.18	0.54
#4 (25 changes)	1	0.97	1	0.71	0.06	0.12	0.09	0.41
#5 (27 changes)	0.89	0.93	0.93	0.78	0.2	0.22	0.19	0.46
#6 (24 changes)	0.96	1	0.96	0.67	0.21	0.2	0.21	0.53
#7 (27 changes)	1	0.96	0.96	0.67	0.13	0.13	0.13	0.54
#8 (29 changes)	0.89	0.93	0.97	0.69	0.04	0.04	0.03	0.43
#9 (26 changes)	0.96	0.96	1	0.65	0.04	0.07	0.07	0.54
#10 (29 changes)	0.93	0.93	0.93	0.66	0.04	0.07	0.04	0.42
mean:	**0.96**	0.95	**0.96**	0.69	**0.12**	0.15	0.14	0.48
std dev:	0.04	0.03	0.03	0.045	0.07	0.08	0.08	0.05
worst case:	0.89	**0.92**	**0.92**	0.64	**0.21**	0.29	0.26	0.54

divergence. Moreover, it could be said that Bhattacharyya measure and modified Lin-Wong measure are slightly better than Lin-Wong measure. Because of lack of space no statistical proof is given here. However, after applying t-Student test it turned out that mentioned remark holds true.

Besides, the results of the KL measure seem to be quite odd and puzzling. After analyzing the raw results it turned out that KL divergence detected almost all of real changes but very often too late. Nevertheless, very weak performance according to the *bad* criterion follows from the definition of that measure. Even very small changes in probability distributions affect in quite big values of the KL divergence. Therefore, it is very sensitive to underestimations.

Furthermore, it can be stated that the proposed modification of the Lin-Wong measure is very interesting and promising. Because of the *flattening* less bad detections, in comparison to Lin-Wong measure, were made. The effect of the *flattening* could be easily seen in Fig.s 4b and 4c. However, more theoretical and experimental proofs are needed.

6 Final Remarks

In this paper a problem of the context change detection was stated and the algorithm for detection was presented. Moreover, four dissimilarity measures were given including one proposed for the first time. At the end the experimental study was carried out. The simulation environment implemented in OMNeT++ was briefly described and conclusions were drawn.

The outcomes of this work are planned to be used in a complex algorithm for decision making about resources management in the service-oriented systems. However, the problem of the context change detection is broader and the algorithm given here could be easily implemented in other domains e.g. learning algorithms. Nevertheless, our goal i.e. implementation of presented approach in the real-life service-oriented system is the priority of our future works.

Acknowledgments. The research presented in this paper has been partially supported by the European Union within the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

The research has been also partially co-financed by European Union within European Social Fund.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Singapore (2006)
2. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, New York (2009)
3. D'Alconzo, A., Coluccia, A., Ricciato, F., Romirer-Maierhofer, P.: A Distribution-Based Approach to Anomaly Detection and Application to 3G Mobile Traffic. In: IEEE Global Telecommunications Conference, Honolulu, pp. 1–8 (2009)
4. Dries, A., Rückert, U.: Adaptive Concept Drift Detection. Statistical Analy Data Mining 2(5-6), 311–327 (2009)
5. Dragomir, S.S., Sunde, J., Buse, C.: New Inequalities for Jeffreys Divergence Measure. Journal of Mathematical Sciences 16(2), 295–309 (2000)

6. European Commission: From Grids to Service-Oriented Knowledge Utilities. A critical infrastructure for business and the citizen in the knowledge society (2006), ftp://ftp.cordis.europa.eu/pub/ist/docs/grids/soku-brochure_en.pdf
7. Grzech, A.: Teletraffic Control in Teleinformatics Networks. Oficyna Wydawnicza PWr, Wrocław (2002) (in Polish)
8. Grzech, A., Świątek, P., Rygielski, P.: Translations of Service Level Agreement in Systems Based on Service Oriented Architectures. Cybernetics and Systems 41(8), 610–627 (2010)
9. Grzech, A., Świątek, P.: Modeling and optimization of complex services in service-based systems. Cybernetics and Systems 40(8), 706–723 (2009)
10. Grzech, A., Świątek, P.: Parallel processing of connection streams in nodes of packet-switched computer communication networks. Cybernetics and Systems 39(2), 155–170 (2008)
11. Harries, M.B., Sammut, C., Horn, K.: Extracting Hidden Context. Machine Learning 32, 101–126 (1998)
12. Huu, T.T., Montagnat, J.: Virtual Resources Allocation for Workflow-Based Applications Distribution on a Cloud Infrastructure. In: 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, pp. 612–617 (2010)
13. Kailath, T.: The divergence and Bhattacharyya distance measures in signal selection. IEEE Trans. on Comm. Tech. 15(1), 52–60 (1967)
14. Kifer, D., Ben-David, S., Gehrke, J.: Detecting Change in Data Streams. In: Proceedings of the 30th VLDB Conference, Toronto, Canada, pp. 180–191 (2004)
15. Klinkenberg, R.: Predicting Phases in Business Cycles Under Concept Drift. In: Proc. of Tagungsband der GI-Workshop-Woche, pp. 3–10 (2003)
16. Lee, W., Xiang, D.: Information-theoretic measures for anomaly detection. In: Proceedings of IEEE Symposium on Security and Privacy, pp. 130–143 (2001)
17. Lin, J.: Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory 37(1), 145–151 (1991)
18. Rose, R.: Survey of system virtualization techniques, Technical report (2004), http://citeseer.ist.psu.edu/720518.html
19. Sebastião, R., Gama, J.: Change Detection in Learning Histograms from Data Streams. In: Neves, J., Santos, M.F., Machado, J.M. (eds.) EPIA 2007. LNCS (LNAI), vol. 4874, pp. 112–123. Springer, Heidelberg (2007)
20. Vorburger, P., Bernstein, A.: Entropy-based Concept Shift Detection. In: Proceedings of the Sixth International Conference on Data Mining, pp. 1113–1118 (2006)
21. Widmer, G., Kubat, M.: Learning in the Presence of Concept Drift and Hidden Contexts. Machine Learning 23(1), 69–101 (1996)

A Multi-Agent Approach for Engineering Design Knowledge Modelling

Ricardo Mejía-Gutiérrez, Alejandro Cálad-Álvarez, and Santiago Ruiz-Arenas

Grupo de Investigación en Ingeniería de Diseño (GRID),
Universidad EAFIT
Cra. 49 No. 7 Sur 50,
Medellín, Colombia
{rmejiag,acaladal,sruizare}@eafit.edu.co

Abstract. Usually a product design process implies collaborative work between experts located in different geographical locations. This context makes the process of taking design decisions more difficult and slow. Nowadays, definition of variables and constraints is a key issue, that experts related with product design have to face throughout the product lifecycle. A Multi-Agent system is proposed for tutoring experts in a standardized manner for the definition of variables and constraints. This domain specific knowledge will be use to build "Constraint Satisfaction Problem" models to support early stages of product design, specifically embodiment design.

Keywords: Knowledge Representation, Multi-Agent System, Ontology, Product Design, CSP – Constraint Satisfaction Problem.

1 Introduction

Nowadays product design has two key aspects to deal with: design teams work in a global environment with distributed members all around the world, and defining a product that is a knowledge intensive process that requires integration of experts' know-how in early stages of product development. It is desired to extract that knowledge in a faster and accurate way as product perception can vary from one expert to another. If a product arises from an idea and takes shape when experts are able to translate their mental models into real world models[1], it is necessary to develop a technology that allows experts to share their knowledge. This technology should help experts to cooperatively build models that properly represent the original idea and meet every constraint related with customer requirements as well as with constraints associated to any Product Life Cycle (PLC) stage (e.g. design, production, use, distribution and disposal, among others).

The aim is to represent expert's technical product knowledge in a structured way, defining an application specific ontology using JADE[2] as a development platform.

[1] This means expressing their knowledge on an explicit way that can be interpreted and manipulated by other experts or by computer software.

[2] Java Agent Development Framework.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 601–610, 2011.

Knowledge could be processed and shared by using Multi-Agent technology in order to help experts in distributed teams work together in the definition of structured models for analysis, taking into account the PLC information.

Agent Oriented Programming (AOP), specifically Multi-Agent System (MAS), has been proposed as programming paradigm to develop the software prototype because it is based on the capability of agents to communicate and react in a distributed, sociable and adaptable way, facilitating communication between distributed experts that work under a cooperative environment [1]. Due to many technical facilities to develop a complete agent-based application like *have a run-time environment implementing the life-cycle support features required by agent and the core logic of agents themselves* [2], JADE has been chosen as the multi-agent framework for this work.

The solution proposed in this article is a software application. The software joins and guides every actor trough the process of variables identification using a knowledge database and a set of questions that make actors think about different kinds of variables related to product design. Those variables, their domains and the set of relations among them are captured, stored and represented in a model for further design analysis. This paper is organized as follows: In section 2, an overview of mains topics is presented. Section 3, introduces our solution for a design framework under a Multi-Agent approach. In section 4, presents a study case. In section 5, present conclusions and future work are shown.

2 Background

Product design is a process that its main objective is to develop the best product that satisfies customers' specifications and life cycle requirements considering every actor involved in the process. In order to achieve this objective, several methodologies have been proposed. All of them are characterized by a systematic development composed of a series of activities and methods that link experts, information and resources [3].

Usually a design process is composed by five stages: requirement definition, conceptual design, embodiment design, detailed design and production [4]. The result of the first two stages is the product expressed in terms of variables and constraints, which subsequently, are used in the embodiment design. At this stage, several concepts are established and some numerical techniques such as optimization, combinatory research or operation research are used to search for consensus between every option proposed. This research aims to tutor and automate the creation of CSP[3] models. CSP basically is *a problem decomposed of a finite set of variables, each of which is associated with a finite domain, and a set of constraints that restricts the values the variable can simultaneously take* [5]. It is used to reduce solution space, obtaining as a result a subset of feasible solutions that meets customer requirements and constraints previously defined by experts.

A CSP model is traditionally build by an intensive process of communication between experts, who try to reach consensus about their interests and include those interests into the model as constraints. This process is manually and takes a lot of time to be completed.

[3] Constraint Satisfaction Problems.

CSP has been used in many application areas like computer vision [6], operators scheduling [7], collaborative design process [8], design of mechanical systems [9] among others, proving its effectiveness in the representation and solution of problems. Using CSP models together with optimization techniques in preliminary design stages reduces the solution space and supports and accelerates decision making.

Using interdisciplinary inputs from the entire design team enhances the knowledge model and therefore better results are obtained by reducing solution space.

Automatic creation of a unique CSP model has become a problem because each expert involved has different experience and mental models, which of course, reflect the way variables and constraints are defined, e.g. a variable is defined with the name 'l', which represents length. Then, a new variable 'x', also representing length, is defined by another expert. As both variables have the same objective, the CSP model defined, is created with data redundancy, reducing the accuracy of the solution set. To face that problem it is necessary a process of standardization, capitalization and formalization of expert's knowledge about the product.

Based on the CSP definition presented above, some authors define knowledge as *relations among a set of variables and its corresponding domains* [10]. This definition is adequate to capitalize the product design knowledge. General knowledge may represented in different ways, but for this particular case the use of variables and constraints (relations) enables product representation into mathematical terms that could be processed by computer tools, such as combinatory search and optimization [11]. These relations restrict variable values and define future actions in the product design process. Based on the definition, product variables become the most important element of design knowledge representation, followed by the constraints. To represent variables and constraints a new ontology is defined using JADE API.[3][4]

3 Constructing a Design Framework under a MAS Approach

In order to create the software prototype it is proposed a knowledge representation, an ontology definition, a framework architecture and a communication sequence.

3.1 Knowledge Representation

Some researches in knowledge management show that many authors have defined knowledge in different ways [12]. Liebowitz argues that knowledge is represented by ideas, rules, and procedures that guide actions and decisions [13]. According to the product knowledge definition accepted for the authors of this paper, it is possible to identify three elements: variables, domains and constraints. Using these elements in the creation of the CSP model would guide design experts in decisions making. Therefore, the knowledge modelling process should start by eliciting it in terms of the components exposed above, where:

– V is a set of *n* variables that are defined by the experts.

$$V = \{v_i \mid v_i \text{ is a variable from the design problem, with } i = 1, 2, 3, \ldots, n\} \quad (1)$$

– D is a set of *n* domains of each variable

$$D = \{d_i \mid d_i \text{ is the domain of the variable } v_i, \text{ with } i = 1, 2, 3, \ldots, n\} \quad (2)$$

– C is a set of *m* relations between variables call constraints.

$$C = \{c_i \mid c_i \text{ is an equation that represent relations among variables of the set V, with } i = 1, 2, 3, ..., m\} \quad (3)$$

Under a product design point of view, product variables are the key component of multidisciplinary interactions. That is why it is necessary to highlight the importance of variables, which are a central element for the Product Knowledge model proposed. To tackle this, the design engineering context's perspective on variables is described.

To understand the concept it is important to define the term "variable". Variable has been defined in mathematics as a value that may take any value from a set of attributes, where an attribute is a characteristic of an object that can take numerical or symbolic values. For modern Physics, a variable is a measurable parameter like temperature, time, intensity, etc. In programming, variable is a symbolic entity (numeric, string, etc), which its value changes during the program execution.

Definitions presented above, have in common that a variable is something that changes its value according to the circumstances and the application domain where it is originated. This principle is evident in a product design context due to different technical disciplines and multicultural interactions that characterize the process. Variable is then the lower level of knowledge in product design and in such a case it is necessary to be captured and represented carefully in order to have a solid base to create the CSP model.

In order to create a better CSP model each variable must be unique. This becomes a problem as humans think in different ways and experts can define multiple variables to measure the same parameter. To solve this, variables are constituted by a set of attributes that define its characteristics. Looking to assure that only one variable per measure is defined in the model; attributes are used to compare variables.

For example, an instance of a variable *v* will have these properties: (1) *ID*, a unique identification number. (2) *Name*, variable name. (3) *Symbol*, variable repre*sentat*ion in the model. (4) *Type*, that takes values from a set of variable's type. (5) *Data Type*, which describe if variable is Symbolic, Integer or Real. (6) Expert, the identification number of the expert who creates the variable. (7) *Unit*, representing units' dimensional exponent. (8) *Prefix*, which express quantities in *The International System of Units* (SI). (9) *Lifecycle stage*, corresponding to the stage of the PLC. *(10) Discipline*, technical classification of the variable. (11) *Step*, that is a discrete value that specify how much change the variable in the model. (12) *Domain*, which specifies inferior and superior limits of the values that can take the variable.(13) *Information Source*, representing the source where variable was obtained. *(14) Measure parameter* belongs to the set of measured parameters extracted from a review of different domains from physics and technical sciences.

A further stage of this project intends to treat variable's domains and constraints under similar approaches. Domains must be homogenized by consensus of design partners that share a variable. Constraints, as has been noted before, are relations among variables that represent experts' knowledge in specific PLC stage. Those constraints are subject to similar treatments to avoid knowledge redundancy. These two issues are not the objective of current research.

3.2 Framework Architecture

To address the management of variables, a prototype system is proposed. Two agents compose the system: Tutor Agent (TA) and Data Base Agent (DBA). TA interacts directly with the expert through a tutoring process that helps experts to transform tacit knowledge into explicit knowledge needed to create a coherent model. The idea is to identify and to qualify relevant design knowledge trough an organic analysis of the design problem. Each expert has a TA assigned, who provides a series of questions that make experts inquire into his/her experience in order to extract knowledge about his/her discipline. Throughout the process experts discover new constraints and represent their knowledge through them.

Once knowledge is captured, TA acts as an interpreter in a negotiation that takes place between an expert and the DBA. The DBA acts a manager of the data base, receiving every request to add, modify or delete variables and constraints from the knowledge data base, reducing cognitive redundancy and therefore increasing the model's quality.

An Architecture diagram of the prototype is presented in fig. 1.

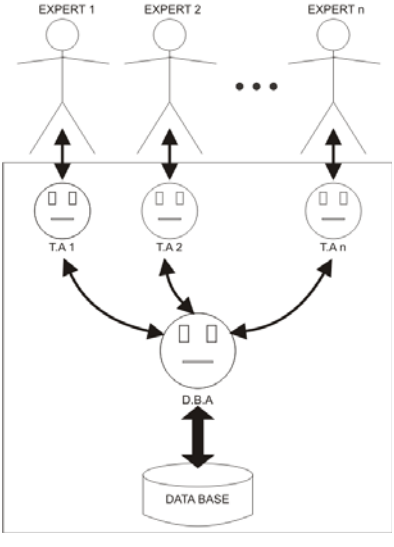


Fig. 1. Prototype architecture. Each expert has a Tutor Agent assigned. Data Base Agent controls interaction with the DB.

The DBA's main task is to compare each expert input data with data from the database. If any redundancy is found, the DBA suggests the expert to use variables that are already defined in the model. To illustrate how the DBA works, the paper continues with the variable definition example used in section 2. As both variables mentioned in the example have the same objective, some of their properties would be equivalent. The DBA compares attributes for the variable 'x' with the attributes of

every variable in the database. If at least 60%[4] of the attributes are the same, the DBA adds the variable, which has been compared in a list of suggested variables, and continues with the next variable until all variables are compared.

3.3 Communication

Communication between agents is illustrated in a UML sequence diagram. Fig. 2 describes the interaction between the TA and the DBA. The communication consists of three stages: Validation, Negotiation and Decision.

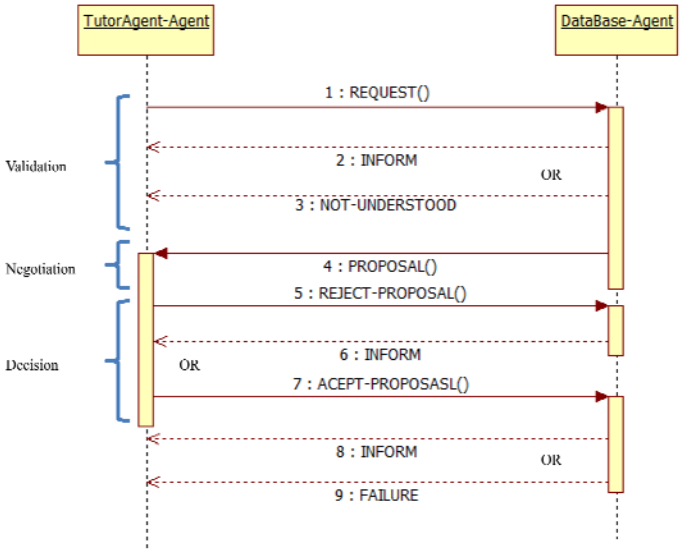


Fig. 2. Sequence diagram between the *tutor agent* and the *data base agent*

Validation starts when the TA sends a `REQUEST` message with the variable to add in the database. The DBA defines if it is possible to add the variable after checking that there are not similar variables in the database by comparing attributes of each variable with the attributes of the requested one. If the DBA finds any similar variable, it makes a suggestion with the analogous variables.

The message is interpreted by the DBA and a response is sent. There are three options of response: The message is not well constructed or an error occurred that produce a `NOT_UNDERSTOOD` response message; in the second case a `INFORM` message is sent to the TA. The message indicates that a new variable has been added and the interaction is finished. In the other case the DBA finds one or more variables that have at least 60%[5] of similarity with the expert's variables, which produces a

[4] This value may modifiable according to problem treated.

[5] This value was defined in order to measure a level of similarity. It can be modified according to the model size.

PROPOSE message, which includes the set of similar variables. With this message the second stage starts.

Negotiation is the process where the expert defines if any of the recommended variables can be used instead of creating a new one. A PROPOSE message contains a list with the similar variables. The interaction continues depending on the expert action that can be: ACCEPT_PROPOSAL, that indicates that this expert becomes a user of the proposed variable or REJECT_PROPOSAL that means the variables proposed do not satisfy the expert's needs. According to the type of message, the DBA can link the expert to the variable that he/she selected and then send a message informing the TA or it can add expert variable to the database and send a message informing that a new variable was created.

3.4 CSP Model Building Ontology

Ontologies are used to describe the elements that agents need to create the content of messages. In this particular case four elements are defined in the ontology: Variable, a concept representing a variable stored in the data base; CreateVariable, an action that represents the action of creating a new variable; Proposal, that represents the action performed by the DBA when it has variables to recommend to an expert; and LinkUser, which represents the action of linking an expert to an existing variable in the model. The ontology can be graphically represented like Fig. 3 where each internal schema can be represented with the same structure. On Fig. 4 a representation of a Variable is depicted.

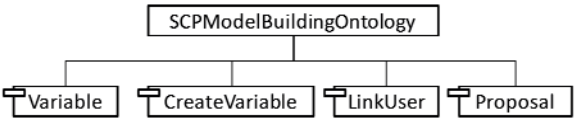


Fig. 3. Ontology used to interchange messages between the TA and the DBA

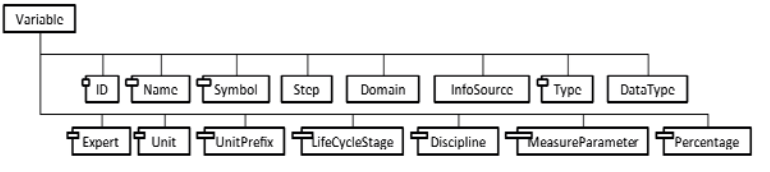


Fig. 4. Variable Ontology. Boxes without the little rectangle are optional values, i.e. that slots do not need to have values when they agent creates an object of that type.

A Variable is composed by PrimitiveSchemas, which means the values of the slots can be primitive data like an Integer, Float or String. As variable, CreateVariable's and LinkUser's slots are composed by PrimitiveSchemas. Otherwise, a Proposal that has only one slot in its internal structure, contains a ConceptSchema, because a Proposal is the set of similar Variables found by the DBA in the comparing process. A Proposal is consider the most important component in the Negotiation stage.

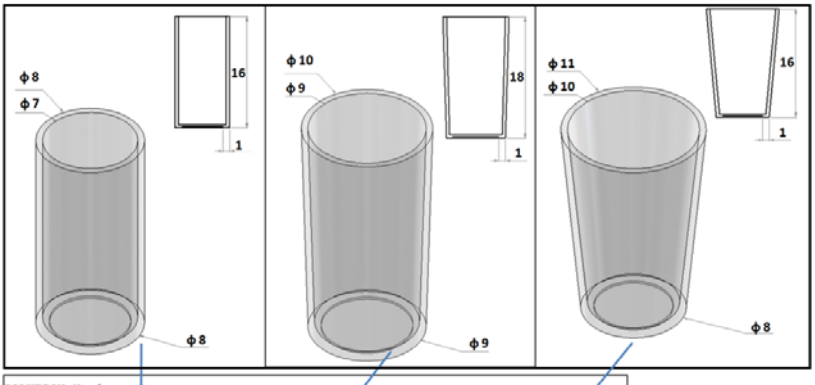
4 Case Study: CSP to Restrict the Solution Space of a Glass

In order to represent the importance of a CSP model in a product design process and to show the application of gathered knowledge in the model creation, a simplified real model of a glass was defined. This model is composed of variables and their corresponding domains, as well as, constraints that represent knowledge gathered.

Construction of a CSP model begins by selecting experts, based on the knowledge needed for the project. Once the experts have been defined, their knowledge is represented through variables, domains and constraints (see Fig. 5) that are gathered

```
####VARIABLES#########                    ####CONSTRAINS######

\vi : h  8...22 ;   #### height           \ci : P1 , x1 = p*y1 ;

\vi : x1 8...11 ; ### Upper external diameter   \ci : P2 , h = 2*y1 ;

\vi : x2 7...10; ### Upper internal diameter    \ci : P3 , y2 = 1 ;

\vi : y1 8...11 ; ### Lower external diameter   \ci : P4 , x2 = x1-1 ;

\vi : y2 1...2 ; ###Low border thickness

\vr : p 0.1 [1,1.4] ; ### Scale factor
```

Fig. 5. CSP model and some solutions resulting from the model's execution




```
SOLUTION No 1
   y2 = 1   x1 = 8   y1 = 8   h = 16   p = 1.000   x2 = 7  sat = 1.000 .
  (trouvee apres 6 instanciations et 4 tests de contraintes)
[...]
SOLUTION No 5
   y2 = 1   x1 = 10   x2 = 9   y1 = 9   p = 1.111   h = 18  sat = 1.000 .
  (trouvee apres 22 instanciations et 14 tests de contraintes)
[...]
SOLUTION No 7
   y2 = 1   x1 = 11   x2 = 10   y1 = 8   p = 1.375   h = 16  sat = 1.000 .
  (trouvee apres 30 instanciations et 19 tests de contraintes)
[...]
############   Fin du Real Full Look Ahead ############
```

Fig. 6. Different solutions obtained from the execution of the CSP model

and defined in the model. Finally the model is executed in Con'flex[6] and a reduced solution space is obtained. Con'flex is an Inference engine for combinatorial calculus developed by INRA[7] free for download. The software has problems in managing real domain calculations due to the presence of discontinuous variables so these kinds of domains were avoided in the model as much as possible.

After executing the example, 10 design options were obtained. These solutions were evaluated and analyzed by the design team which selected three of them to be modeled. Solutions 1, 5 and 7 were modeled in SolidWorks as shown in Fig. 6.

This small CSP Model shows the pertinence of the tool in early stages of the design process, particularly in Embodiment Design where experts are looking to reduce the solution space while ensuring to meet all the constraints related to a specific design problem. Due to this, it is important to properly define all the components of the model and ensure that experts define all constraints related to their knowledge area.

5 Conclusions and Future Research

Design knowledge under a product development approach can be represented as a set of variables and relations among them. It is important to capture and represent domain specific knowledge around PLC stages in order to use it in the construction of CSP Models for a combinatory search of feasible solutions. A knowledge representation and a MAS approach, based on JADE, have been proposed to support the Embodiment Design stage in distributed environments. Current research intends to help reduce design solution space by standardizing variables and constraint definitions that will support the implementation of CSP models. This research contributes to an enhanced modeling approach for Engineering Design knowledge to take into account the different variables and constraints issued from PLC disciplines. As different experts interact during the design process, it enhances the CSP model construction, and proposed methods minimize inconsistencies on combinatory models that may be treated further on by inference engines in a proper manner. As further research, variable´s shared domains will be treated with fuzzy logic to ensure the conflict resolution if two or more experts are sharing variables. Some additional work has to be done with constraints management, for example by checking dimensional units for consistency and redundancy analysis.

References

1. Marik, V., Pechoucek, M.: Multi-agent system for production planning. Applied Artificial Intelligence 14, 727–762 (2000)
2. Bellifemine, F.L., Caire, G., Dominic, G.: Developing multi-agent systems with JADE wileys series in agent technology. Wiley, Chichester (2007)

[6] A Constraint Satisfaction Problem (CSP) solving software. It can be found in: http://www.inra.fr/mia/T/conflex/

[7] Institut National de la Recherche Agronomique.

3. Clarkson, J., Eckert, C.: Design process improvement: a review of current practice. Springer, Heidelberg (2005)
4. Baxter, M.: Product design. A practical guide to systematic methods of new product development. Chapman & Hall, Boca Raton (1995)
5. Tsang, E.: Foundations of Constraint Satisfaction. Academic Pr., London (1993)
6. Yamamoto, K.: Optimization approaches to constraint satisfaction problems in computer vision. Image and Vision Computing 13, 335–340 (1995)
7. Kim, K.H., Kim, K.W., Hwang, H., Ko, C.S.: Operator-scheduling using a constraint satisfaction technique in port container terminals. Computers & Industrial Engineering 46, 373–381 (2004)
8. Yvars, P.-A.: A CSP approach for the network of product lifecycle constraints consistency in a collaborative design context. Engineering Applications of Artificial Intelligence 22, 961–970 (2009)
9. Yvars, P.-A.: Using constraint satisfaction for designing mechanical systems. Int. J. Interact. Des. Manuf. 2, 161–167 (2008)
10. Mejia-Gutierrez, R., Fischer, X., Bennis, F.: A Tutor Agent for supporting distributed knowledge modelling in interactive product design. International Journal of Intelligent Systems Technologies and Applications 4, 399–420 (2008)
11. Giassi, A., Bennis, F., Maisonneuve, J.J.: Multidisciplinary design optimisation and robust design approaches applied to concurrent design. Structural and Multidisciplinary Optimization 28, 356–371 (2004)
12. Fu, Q.Y., Chui, Y.P., Helander, M.G.: Knowledge identification and management in product design. Journal of Knowledge Management 10, 50–63 (2006)
13. Liebowitz, J.: Knowledge management handbook. CRC, Boca Raton (1999)

An Architecture for the Semantic Enhancement of Clinical Decision Support Systems

Eider Sanchez[1], Carlos Toro[1], Eduardo Carrasco[1], Gloria Bueno[2], Carlos Parra[3], Patricia Bonachela[3], Manuel Graña[4], and Frank Guijarro[5]

[1] Vicomtech-IK4 Research Centre, Mikeletegi Pasealekua 57, 20009 San Sebastian, Spain
{esanchez,ctoro,ecarrasco}@vicomtech.org
[2] University of Castilla-La Mancha, VISILAB group, ETSII, Spain
[3] University Hospital Virgen del Rocío, UCAi group, Spain
[4] University of the Basque Country, Computational Intelligence Group, Spain
[5] Bilbomatica, Spain

Abstract. Clinical Decision Support Systems (CDSS) are useful tools that aid physicians during different tasks such as diagnosis, treatment and patient monitoring. Multidisciplinary, heterogeneous and disperse clinical information and decision criteria have to be handled by CDSSs. For such tasks, Knowledge Engineering (KE) techniques and semantic technologies are very suitable, as they support (i) the integration of heterogeneous knowledge, (ii) the expression of rich and well-defined models for knowledge aggregation, and (iii) the application of logic reasoning for the generation of new knowledge.

In this paper we propose a generic architecture of a CDSS based on semantic technologies, which also considers the reutilization and enhancement of former CDSS in an organization. Particularly, an implementation of the proposed architecture is also presented, aiming to support the early diagnosis of AD.

Keywords: Decision support system, architecture, implementation, Alzheimer Disease.

1 Introduction

Clinical Decision Support Systems (CDSS) are active knowledge resources that use patient clinical data to generate case specific advice [1]. In other words, CDSS analyze data and present results to physicians. Such results are used for (*i*) supporting decision making during diagnosis and (*ii*) supporting treatment and patient monitoring. CDSS are massive information systems by nature while at the same time they present an arguably high complexity (in terms of computational resources) in the query construction and retrieval. Such complexity is increased when the variable values needed for decision-making are stored in disperse or heterogeneous repositories [2].

Some reported issues of CDSS are mentioned in the literature. Reported main difficulties are mainly presented while in the process of (a) integrating CDSS into clinical workflows and systems, and (b) transferring successful interventions from one system to another [3].

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 611–620, 2011.

In classical CDSS, the representation of knowledge is static, limiting the type of knowledge that can be represented [3]. Additionally, CDSS definition is specified only through explicit information enumeration (i.e. case-based systems). Hence, arguably no discovery of new knowledge is directly supported.

Another problem in CDSS is the fact that useful information for diagnosis is highly changeable. The aforementioned fact is due to the natural evolution of medical research, where new findings and advances are being continuously made. For instance, a biomarker could be rendered irrelevant, by a new discovery that supersedes it. Thus, variables and criteria of the CDSS should be often updated and for this reason, the maintainability of the system could be a critical problem, i.e. for decision support systems integrated to clinical systems [3]. Terminological interoperability is also an important matter that classical approaches in CDSS do not solve appropriately [3]. Two different CDSS may not understand each other, even if their domain and purpose is the same, because they can adopt different terminologies or, in extreme cases, due to the inertia related to monolithic and legacy system architectures.

Knowledge Engineering (KE) techniques can arguably face efficiently the aforementioned problems (which are in essence, Knowledge handling problems) because, by definition, their underlying models support (i) the integration of heterogeneous knowledge, (ii) the expression of rich and well-defined models for knowledge aggregation and (iii) the application of logic reasoning for the generation of new knowledge [4]. In particular, semantic technologies have been described in the literature as a promising approach to solve knowledge handling problems in medical domain, as shown by Gnanambal *et al.* [5] and by Yu *et al.* [6].

In this paper we propose a generic architecture for the semantic enhancement of CDSS, which also considers the reutilization of knowledge embedded in a CDSS, in order to provide the enhancement of such CDSS taking into account and lessening if not solving the main problems mentioned before. An implementation of our proposed architecture is also presented; this implementation deals with the specific domain of early diagnosis of Alzheimer Disease (AD).

This paper is arranged as follows: in section two we present briefly the related work which is relevant for our approach; in section three we introduce the architecture of a generic knowledge-based Clinical Decision Support System; in section four we present an implementation of the proposed architecture for the early detection of AD, and lastly, we present future work and conclusions in section five.

2 Related Work

In this section we present a short overview of the previous work mentioning briefly concepts related Clinical Decision Support Systems (CDSS) and the possible benefits of the application of semantic technologies in this domain.

2.1 Architectures in Clinical Decision Support Systems

According to Wright *et al.* [3] the evolution of architectures for CDSS has followed four phases: standalone CDSS, CDSS integrated to clinical systems, standards-based systems, and service models.

Standalone CDSS run separately from any other system, such as clinical systems containing the clinical information from patients and cases. Thus, a physician has to intentionally enter the required information and ask for the aid. Time is consumed during this process. Usually the system is not proactive when supporting decision making. On the bright side, these CDSS are very easy to share [3]. On the other hand, CDSS integrated into clinical systems behave just the opposite.

Standards-based systems aim to the standardization of the computerized representation, encoding, storing and sharing of clinical knowledge and decision support content [3]. There are several standards offering a different focus, such as Arden Syntax [7] and GELLO [8].

Service models separate clinical information systems and CDSS [3], and integrate them, while using standardized service-based interfaces. The standard interface can be both, located in front of the clinical system, so that any decision support system that understands the standard can make inference (i.e. HL7 vMR [9]), or located in front of the decision support system, in order that any clinical system that understands this standard can ask for aid to a known CDSS (i.e. HL7 DSS [10]).

Additionally, some other CDSS architectures have also been presented in [11],[12],[13].

2.2 Semantic Technologies Applied to Clinical Decision Support Systems

Knowledge Engineering (KE) techniques can face efficiently the aforementioned problems such as terminological interoperability, system maintainability and source heterogeneity and disparity. More precisely, semantic technologies have been described in the literature as a promising approach to solve knowledge handling in medical domain, i.e. in [5], where different approaches using semantic technologies are presented for several research directions in the medical domain.

In particular, ontologies are very promising. Gruber defined ontologies in the computer science domain as the explicit specification of a conceptualization [14]. Ontologies can fulfill efficiently the needs for organized and standardized terminologies and reusability at a structural level [15]. Because of the aforementioned fact, important consequences in the medical domain can be derived. Some results can be applied to solve interoperability issues, as shown by Ghawi *et al.* in [16], where a general interoperability architecture is presented that uses ontologies for explicit description of the semantics of information sources.

Ontologies also deliver interesting benefits, when used for reasoning and inferring new knowledge [6]. For instance, the fast query systems presented by Toro *et al.* [17].

Among the most widely used ontologies within the medical domain, we can mention the Semantic Web Application in Neuromedicine (SWAN) [18] and the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [19].

SWAN represents an effort to provide an integrated scientific knowledge for researchers to share their results within different projects and locations. It is the result of a project intended for developing an integrated scientific knowledge infrastructure using Semantic Web technologies. SWAN has been applied to Alzheimer Disease, but it is not limited to it. The integration with SWAN endorses contents with hypotheses

and publications, as shown by Lam *et al.* [18]. Decision support needs to be documented and SWAN overcomes this task.

SNOMED CT is a common standardized comprehensive clinical terminology that provides clinical content and expressivity for clinical documentation and reporting [19]. SNOMED CT provides the core general clinical terminology for the Electronic Health Record (EHR). It describes different clinical concepts such as diseases and procedures. Mapping our own ontology to a standard ontology such as SNOMED CT provides reusability of the proposed ontology, according to Houshiaryan *et al.* [15]. This fact is very important for CDSS, to overcome the lack of common language problem. Our approach shares some basic ideas of the works presented by Hussain *et al.* [11] and Lindgren [20], related to the benefits and techniques needed for the coexistence of CDSS and semantic technologies. Our main focus is the re-use and standardization of knowledge as well as the user expertise which ultimately generates the production rules that provide the diagnosis support.

3 Proposed Architecture

In this section, we propose a generic architecture of a CDSS based on semantic technologies. The proposed architecture consists of 4 layers: *Data Layer*, *Translation Layer*, *Ontology and Reasoning Layer* and *Application Layer.* This architecture considers the reutilization and enhancement of former CDSS on existence in an organization. Fig. 1 depicts an overview of our architecture.

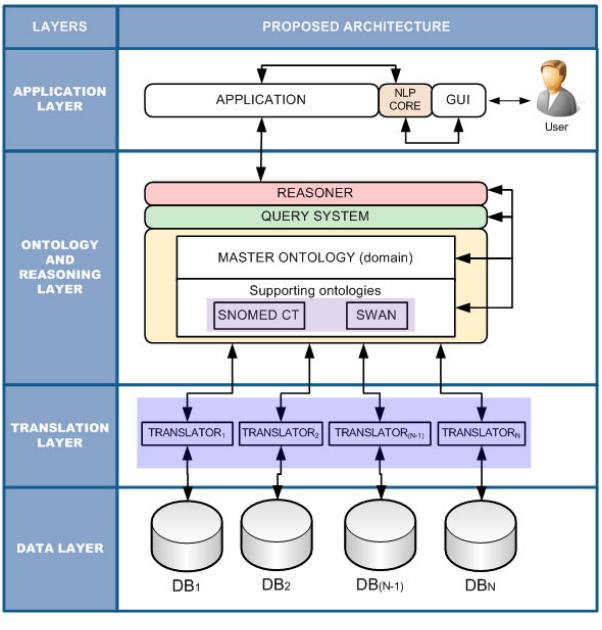


Fig. 1. Proposed architecture for the Clinical Decision Support System

3.1 Data Layer

This layer contains a collection of accessible Data Bases (DB) that store all needed information from medical systems, patient information management systems, and Picture Archiving and Communication Systems (PACS), amongst others. Is common to find that these DB are heterogeneous, as well as spatially disperse.

3.2 Translation Layer

This layer contains a translator for each DB in the *Data Layer*. Each translation module retrieves the corresponding information from the DB and aligns them into the Knowledge Bases in the *Ontology and Reasoning Layer.* In other words, the translator matches the contained information with the ontologies in the upper layer. Within this approach DB do not need to intercommunicate directly, allowing a decentralized data repository to provide input to a centralized knowledge repository.

3.3 Ontology and Reasoning Layer

The *Ontology and Reasoning Layer* deals with the knowledge embedded in the system and performs reasoning processes in order to provide diagnoses. It consists of three modules: the ontologies module, the query system and the reasoning module.

Three different ontologies giving different approaches and descriptions make up the ontologies module: a master ontology and the supporting ontologies SWAN and SNOMED CT.

SWAN links and endorses the knowledge of the system about the disease with hypotheses and publications that are being held by the medical and scientific community [18]. Therefore, the terms and the criteria in the system can be validated to be current and updated.

SNOMED CT is used to for standardization purposes [15]. This alignment provides our system with interoperability to other CDSS or knowledge sources. SNOMED CT is a general-purpose ontology and may not describe all terms needed by the system. Thus, a master ontology containing those particular terms required by each specific CDSS is proposed in our approach. This master ontology is defined by experts for the specific domain of the CDSS.

The reasoning module performs a semantic reasoning process based on expert-given rules, for the knowledge discovery. It queries the underlying ontologies through a query system. The reasoning process concludes diagnoses which are grounded and are presented to physicians to support them during decision making.

3.4 Application Layer

The interaction between the user and the system will be held by a graphical user interface (GUI). The GUI communicates with a natural language processing core, that converts to machine-language procesable the queries made by the user, and to natural language the answers returned by the system. The *Application Layer* communicates directly with the reasoner in the layer below. In this way, the output given the reasoner is presented to physicians clearly, so that a support is given in order to make decisions.

4 Architecture Implementation

The proposed architecture was implemented under the framework of the project MIND (CENIT-20081013). This project is a multidisciplinary approach to Alzheimer Disease (AD) and it is particularly focused on the early diagnosis of AD. The neurodegenerative process of AD is irreversible and for that reason a prodromal diagnosis is desirable. The common approach to support the diagnosis is based on the analysis of the results of different parameters, regarding neurological tests, neuropsychological tests, genetic studies, metabolomical studies, volumetry analysis and diagnostic image processing, amongst others [21]. During this process physicians have to deal with large amounts of heterogeneous and multidisciplinary variables. Additionally, the state of the art regarding these relevant parameters, biomarkers and procedures to follow in order to carry out a proper diagnosis is changing very fast, so that physicians have to be updated with the last medical findings.

Our architecture can handle efficiently with these tasks. Fig. 2 depicts the implementation of our presented architecture within the domain of the early detection of AD.

The *Data Layer* contains two DB: the ODEI Data Base, storing the results of the clinical tests carried out to patients, and the ODEI PACS, containing the DICOM image studies.

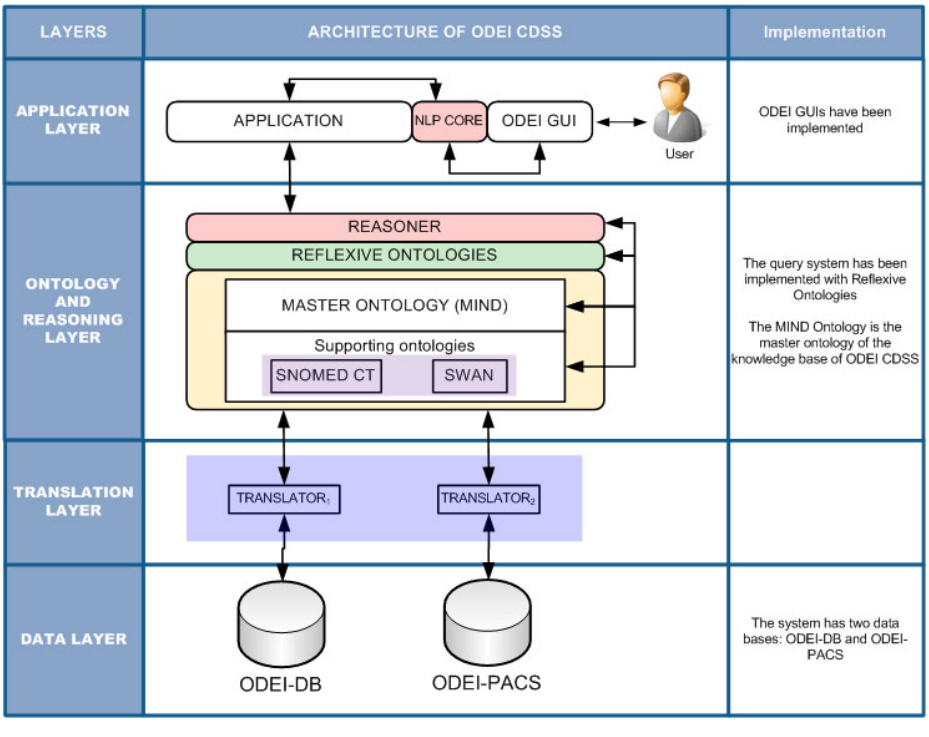


Fig. 2. Architecture implementation for the CDSS of the MIND project

The *Translation Layer* is composed by xml data models, used to align the information stored in DB with the knowledge of the upper later. For every DB an xml schema document is created containing all the elements of the DB. A data xml document is created for every data generated, at the time when data is added, modified or deleted from the DB.

The ontologies module in the *Ontology and Reasoning Layer* contains a master ontology modeling the knowledge concerning the results of the clinical tests and image studies carried out to patients. Such knowledge model is depicted in Fig. 3.

The supporting ontologies are SWAN [18] and SNOMED CT [19]. SWAN contains the description of the domain of the AD and SNOMED CT describes the patient, from a clinical point of view.

The reasoning module performs a semantic reasoning process based on rules given by domain experts, which model the process of diagnosis of AD. The query system implemented is based on Reflexive Ontologies presented by Toro *et al*. [17].

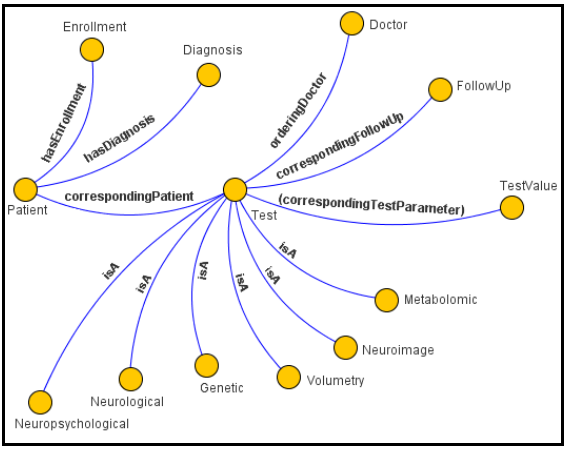


Fig. 3. Knowledge model of the results of the clinical tests and image studies

The production rules follow a classical if/then/else structure and are, on the one side, weighted depending on an importance hierarchy, and on the other, endorsed by the corresponding bibliographic source via a link given by the mapping of the MIND ontology and SWAN. Fig. 4 depicts an example of one of our production rules.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<RuleSet>
    <LoadRule>
        <Rule>if ( ( CLASS Neurological with the PROPERTY Neur
        <weight>0.6</weight>
        <AccordingTo>doi: 10.1016/S0028-3932(01)00055-0</Accor
    </LoadRule>
</RuleSet>
```

Fig. 4. Production rule example

In the *Application Layer* the ODEI Graphical User Interface (GUI) has been implemented, as well as a Natural Language Processing core, which translates the rules and their output, for the reasoner and the physician respectively. Fig. 5 shows a diagnosis for a patient supported by the ODEI system and the output of the reasoner.

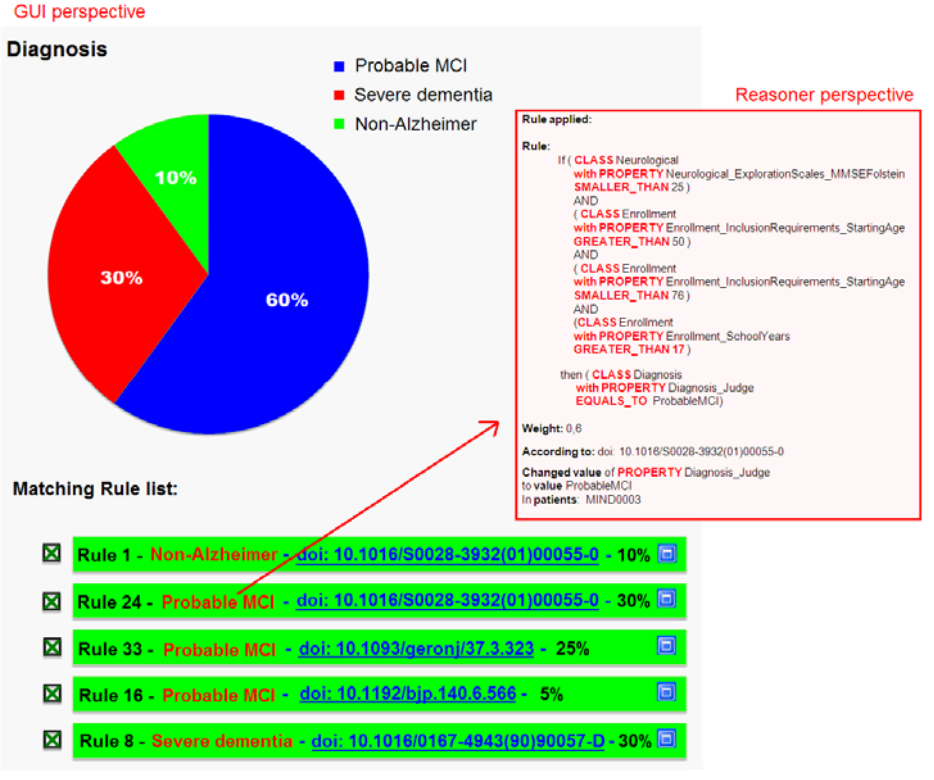


Fig. 5. Diagnosis supported by the ODEI system and output of the reasoner

5 Conclusions and Future Work

In this paper we have presented a generic architecture for the semantic enhancement of CDSS, based on semantic technologies. Our architecture allows the reutilization and enhancement of existent CDSS. In particular, this architecture has been implemented for the early diagnosis of AD. Semantics has been identified as a valuable asset technology for CDSS. The gaps bridged with the application of our methodology are related to interoperability support and discovery of new knowledge.

Our approach is horizontal and as a result the re-utilization of existent knowledge embedded in an actual CDSS is advantaged. We conclude that using standardization efforts (in the shape of existent ontologies) is beneficial as reduplication of the knowledge base is lessened and the resultant semantic layer is strengthened. In this work we used a triple ontology approach that consists of SWAN, SNOMED CT and a

master ontology that contains the domain specifics that are not included directly in the aforementioned supporting ontologies.

As future work, we will explore the Set of Experience Knowledge Structure (SOEKS) [22], in order to support an experience-based reasoning that will provide an experience modeling and re-use on the production rules.

References

1. Joseph, L., Wyatt, J.C., Altman, D.G.: Decision tools in health care: focus on the problem, not the solution. BMC Medical Informatics and Decision Making 6(4) (2006)
2. Vaquero, J., Toro, C., Palenzuela, C., Azpeitia, E.: Using Semantics to Bridge the Information and Knowledge Sharing Gaps in Virtual Engineering. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010. LNCS, vol. 6277, pp. 495–504. Springer, Heidelberg (2010)
3. Wright, A., Sittig, D.F.: A four-phase model of the evolution of clinical decision support architectures. International Journal of Medical Informatics 77(10), 641–649 (2008)
4. Dumontier, M., Andersson, B., Batchelor, C., Denney, C., Domarew, C., Jentzsch, A., Luciano, J., Pichler, E., Prud'hommeaux, E., Whetzel, P., Bodenreider, O., Clark, T., Harland, L., Kashyap, V., Kos, P., Kozlovsky, J., McGurk, J., Ogbuji, C., Samwald, M., Schriml, L., Tonellato, P.J., Zhao, J., Stephens, S.: The Translational Medicine Ontology: Driving personalized medicine by bridging the gap from bedside to bench. In: Proceedings of the 13th Annual Bio-Ontologies Meeting, Bio-Ontologies, Boston, USA (2010)
5. Gnanambal, S., Thangaraj, M.: Research Directions in Semantic Web on Healthcare. Journal of Computer Science 1, 449–453 (2010)
6. Yu, W.D., Jonnalagadda, S.R.: Semantic Web and Mining in Healthcare. In: e-Health Networking, Applications and Services, pp. 198–201 (2006)
7. Zhou, Q., Wang, W.: The Automatic Inference of Arden Medical Logic Modules. In: Proceedings of the International Conference on BioMedical Engineering and Informatics (2008)
8. Peleg, M., Rubin, D.L.: Querying Radiology Appropriateness Criteria from a virtual Medical Record using GELLO. In: Proceedings of the Workshop on Knowledge Representation for Health-Care: Patient Data, Processes and Guidelines, in conjunction with Artificial Intelligence in Medicine Europe (2009)
9. Johnson, P.D., Tu, S.W., Musen, M.A., Purves, I.: A Virtual Medical Record for Guideline-Based Decision Support. Medinfo, 294–298 (2001)
10. Kawamoto, K., Lobach, D.F.: Proposal for Fulfilling Strategic Objectives of the U.S. Roadmap for National Action on Decision Support through a Service-oriented Architecture Leveraging HL7 Services. Journal of the American Medical Informatics Association 14(2), 146–155 (2007)
11. Hussain, S., Raza Abidi, S., Raza Abidi, S.S.: Semantic Web Framework for Knowledge-Centric Clinical Decision Support Systems. In: Bellazzi, R., Abu-Hanna, A., Hunter, J. (eds.) AIME 2007. LNCS (LNAI), vol. 4594, pp. 451–455. Springer, Heidelberg (2007)
12. Wright, A., Sittig, D.F.: SANDS: A service-oriented architecture for clinical decision support in a National Health Information Networkstar, open. Journal of Biomedical Informatics 41(6), 962–981 (2008)
13. Michalowski, W., Slowinski, R., Wilk, S., Farion, K.J., Pike, J., Rubin, S.: Design and development of a mobile system for supporting emergency triage. Methods of Information in Medicine 44(1), 14–24 (2005)

14. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies 43(5–6), 907–928 (1995)
15. Houshiaryan, k., Kim, H.S., Kim, H.H., Tung, T., Kim, I.K., Kwak, Y.S., Cho, H.: Customized Ontology-Based Intelligent Medical Terminology Tool. In: Proceedings of 7th International Workshop on Enterprise Networking and Computing in Healthcare Industry, pp. 320–324 (2005)
16. Ghawi, R., Cullot, N.: Database-to-Ontology Mapping Generation for Semantic Interoperability. In: Third International Workshop on Database Interoperability (InterDB 2007), held in conjunction with VLDB 2007, Vienna, Austria (2007)
17. Toro, C., Sanín, C., Szczerbicki, E., Posada, J.: Reflexive Ontologies: Enhancing Ontologies with self-contained queries. Cybernetics and Systems: An International Journal 39, 171–189 (2008)
18. Lam, H.Y.K., Marenco, L., Clark, T., Gao, Y., Kinoshita, J., Shepherd, G., Miller, P., Wu, E., Wong, G., Liu, N., Crasto, C., Morse, T., Stephens, S., Cheung, K.H.: Semantic Web Meets e-Neuroscience: An RDF Use Case. In: Proceedings of the ASWC International Workshop on Semantic e-Science, pp. 158–170 (2006)
19. Nyström, M., Vikström, A., Nilsson, G.H., Ahlfeldt, H., Orman, H.: Enriching a primary health care version of ICD-10 using SNOMED CT mapping. Journal of Biomedical Semantics 1(7) (2010)
20. Lindgren, H.: Towards personalized decision support in the dementia domain based on clinical practice guidelines. In: User Modeling and User-Adapted Interaction, pp. 1–30 (2011)
21. Monien, B., Apostolova, L., Bitan, G.: Early diagnostics and therapeutics for Alzheimer's disease – how early can we get there? Expert Review of Neurotherapeutics 6, 1293–1306 (2009)
22. Sanín, C., Szczerbicki, E.: Experience-based knowledge representation: SOEKS. Cybernetics and Systems 40(2), 99–122 (2009)

An Approach to Measure Quality of Knowledge in the e-Decisional Community

Leonardo Mancilla-Amaya, Cesar Sanín, and Edward Szczerbicki

School of Engineering, Faculty of Engineering and Built Environment
The University of Newcastle
University Drive, Callaghan, 2308, NSW, Australia
leonardo.mancilla@uon.edu.au,
{Cesar.Sanin,Edward.Szczerbicki}@newcastle.edu.au

Abstract. Sharing knowledge is an activity that can improve an organization's ability to make better decisions and adapt faster to unexpected situations. As a consequence, using measures to determine the quality of the knowledge that is shared by individuals and enterprises will improve decision-making processes and the accuracy of their outcomes. This paper presents a new approach for measuring quality of knowledge in the e-Decisional Community, an integrated knowledge sharing platform that aims at the creation of markets where knowledge is provided as a service.

Keywords: e-Decisional Community, Software agents, Knowledge Quality, Knowledge Quantity.

1 Introduction

Organizations have acknowledged the importance of quality as a strategic factor that can determine their survival in the industry. Quality has been traditionally oriented towards delivering superior products, increase revenues and guarantee customer satisfaction. In addition, consumers have become more demanding and knowledgeable, and this creates additional pressure for managers, who also need to devise new ways to react quickly to their customers' high expectations. Recently, organizations have also realized about the importance of knowledge in their strategies. By using appropriate knowledge management processes, organizations are able to reuse their experience to make accurate decisions, and as a result they save time, money and are able to provide added value (i.e. higher quality) to their products and services.

Quality and knowledge have become crucial elements for enterprises in order to obtain competitive advantage in today's knowledge-oriented economy. To support decision-making processes, knowledge has to meet some criteria to be useful; in other words, it has to be of high quality; otherwise, it may lead to inaccurate decisions. Consequently, many researchers have explored ways of measuring quality of knowledge as presented in [6], [16] and [19]. However, this is still an area of research that allows for further improvement due to the difficulty of measuring an intangible asset such as knowledge.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 621–630, 2011.

In this paper, a different approach to measure explicit knowledge quality is presented. This proposal aims at creating a semi-automatic way of assessing quality, with the purpose of increasing the effectiveness of the decisions made by organizations. The concepts described in this paper are applied in the e-Decisional Community [10], an agent-based platform for sharing experiential knowledge across different organizational levels. The e-Decisional Community uses a standard knowledge representation called Set of Experience Knowledge Structure (SOEKS), which comprises Decisional DNA (DDNA) [13]. Decisional DNA is proposed as a unique and single structure for capturing, storing, improving and reusing decisional experience. Its name is a metaphor related to human DNA, and the way it transmits genetic information among individuals through time.

The structure of this paper is as follows: section 2 presents the conceptual background on quality, knowledge measurement and quality of knowledge. Section 3 describes the proposal on how to measure quality in the e-Decisional Community. Section 4 presents the experiments performed and their results. Finally, section 5 presents the conclusions and future work.

2 Background

Measuring knowledge quality is an area of research that has attracted the efforts of several researchers in the last years. It is a topic that presents many challenges, since there is no exact way to measure knowledge, based on the fact that measurement criteria are not precise [19]. This section presents an overview of the topics of quality, knowledge measurement, and quality of knowledge.

2.1 Quality

There is no consent in literature about the meaning of quality. There are several descriptions of quality, and all of them seem to be based on the specific context on which they are used. For instance, Seawright and Young [14] present a variety of definitions about quality and the relationships between them, and classified them in seven categories as follows: strategic, transcendent, multidimensional, manufacturing-based, value-based, product-based and user-based. These quality definitions influence each other, and according to [14], the understanding of these associations can help an organization compete in a better way.

The proposal presented in this paper, sees quality from the value point of view. For the e-Decisional Community, knowledge and experience are assets that provide a company with the means to adapt, and respond rapidly and appropriately to changes in the environment. This can be seen also as providing an organization with added value from its day-to-day operations. Following the definition presented by Seawright and Young [14], value-based quality is an extension of user-based quality, in which a product satisfies users' needs. More precisely, value-based quality is defined as "excellence" or "fit for use". In the e-Decisional Community, knowledge must be fit to solve an organizational problem with the best possible result, in order to be considered of good quality.

2.2 Knowledge Measurement

Measuring knowledge has been regarded as a complex task by some researchers. For instance, Steedman [15] questions the different proposals on knowledge measurement from the perspective of economics. He argues that knowledge might not be cardinally measured, and that many existing proposals in literature about the subject lack a solid conceptual foundation. Only when theory produces clear indicators, it will be possible to identify magnitudes and measure knowledge accurately [15].

In spite of this limitation, many other proposals have been developed aiming at solving the question of how to measure knowledge. Bontis [1] presents a literature review on the different models that have been used to measure intellectual capital. List, Schiefer, and Bruckner [8] present a workflow-based approach to measure knowledge based on the premise that knowledge is embedded in organizational procedures and daily practices, or develops over time throughout experience and action. Darroch [2] proposes a measure for knowledge behaviors and practices in organizations. In addition, Hunt [5] defines the concept of personal knowledge and presents a method to measure it. This measure addresses the shortcomings of existing multiple-choice tests, including elements like sureness and misinformation as part of the final scores, to produce more meaningful results.

The approaches described previously represent a step towards formalizing the process of knowledge measurement. However, most of these efforts do not provide standardized indicators to assess knowledge, mainly because they are highly coupled with the context in which they are used. Therefore, the approach presented in this paper aims at measuring knowledge in such a way that it can be used in to support decision-making processes in different domains.

2.3 Knowledge Quality Measurement

A number of research efforts have addressed the issue of measuring quality of knowledge. Some of the existing proposals focus on knowledge quality; others integrate process-oriented views for quality assurance, or define quality guidelines for knowledge-based systems.

For example, Tongchuay and Praneetpolgrang [19] and Lee et al. [6] present a set of indicators for knowledge management systems. Both approaches are based on elements from existing data and information quality research. Additionally, Supekar et al. [16] tackle quality of knowledge in the Semantic Web, allowing software agents and knowledge engineers to accurately judge the quality of ontologies based on the ranking of different knowledge sources. Other research efforts ([7] and [11]) have studied the integration of quality management practices with knowledge management. In this way, organizations are able to obtain indicators about knowledge quality from a process-oriented point of view.

It is apparent that there are many research efforts that concentrate on quality of knowledge. However, none of them deal explicitly with experiential knowledge, and most do not provide an automated solution for quality measurement. In many cases, a high degree of human intervention is required in order to define and evaluate different quality indicators.

3 Quality of Knowledge in the e-Decisional Community

This section presents the proposed approach for explicit knowledge quality measurement in the e-Decisional Community.

3.1 Knowledge Quality Attributes

Measuring an entity's knowledge is only possible if there is way to explicitly represent it and quantify it. Consequently, knowledge quality measurement in the e-Decisional Community is based on a set of nine attributes, taken from existing literature on data and information quality (see Table 1). The main reason to base the proposal presented in this section on data and information, is that they play an important role in the creation of knowledge, as described by Davenport and Prusak [3]. Also, since the process of quality assessment is meant to be semi-automatic, the best approach is to define a set of items that the agents participating in the e-Decisional Community can measure.

The proposed quality attributes were selected based on their number of appearances in literature. This indicates that there is a consensus about role in quality measurement. The selection process was based on a Pareto analysis.

After obtaining the preliminary attributes, an individual analysis of each one of them was performed to refine the list, and determine the viability of their implementation in a knowledge-oriented context using Decisional DNA and SOEKS. Table 1 presents the final attributes that were selected after the depuration process, along with their definitions from literature.

Table 1. Knowledge Quality Attributes

Indicator	Definition
Accuracy	Degree of closeness of its value v to some value v', considered correct for an entity and an attribute. (Sometimes v' is referred to as the standard.) [4].
Timeliness	The extent to which the data is up-to-date for the task at hand [12].
Completeness	Data is sufficient and not missing in order to complete a task[12].
Relevance	Relevance is concerned with whether acquired knowledge can be applied in a user's task [6].
Understandability	The level of expressiveness that allows for the meaning of information to be understood easily [6].
Reputation	Data highly regarded in terms of its source or content [12].
Believability	The extent to which data is regarded as true or credible [4].
Objectivity	Data is unbiased [12].
Amount	-The level of appropriateness for quantity of provided information to be used in current affairs [6]. -The extend to which the volume of data is appropriate for the task at hand [12].

The amount of knowledge was included as a key element in this proposal for a main reason: measuring the amount of knowledge in the e-Decisional Community will allow the platform to provide an estimate on the depth of an agent's knowledge. Research on how to measure quantity of knowledge is an ongoing task of the Knowledge Engineering Research Team, at The University of Newcastle.

3.2 Obtaining Values for the Quality Attributes

Attributes were grouped depending on how their values can be obtained; three different ways were identified: user, agent, or the Smart Knowledge Management System - SKMS (presented in [13]). All the values have a range from zero to one ([0, 1]). The grouping is as follows:

- User Category (Values obtained from user feedback):
  - Timeliness: Indicates if an agent's knowledge is updated according to the user's needs. Its default value is 1, because knowledge is assumed to initially meet this criterion.
  - Relevance: Indicates if a solution proposed by an agent is relevant to the problem at hand. The default value is 0. 5; an intermediate value is assigned as default because the system cannot determine beforehand how relevant an experience is.
  - Understandability: Refers to the way a solution is presented to the user, if it makes sense and can be understood. It is the responsibility of the agent to provide solutions to the application layer in a human-readable format. The default value is 0.
  - Objectivity: Evaluates if knowledge is unbiased. User feedback based on a personal perspective might influence knowledge. Therefore, when a solution is shared, users have the opportunity to evaluate its impartiality. The default value is 1; knowledge is assumed to be objective when a new SOEKS is created.
- Agent Category (Values automatically calculated by agents):
  - Amount: The amount of knowledge in a certain area. To be defined in future work.
  - Completeness: Indicates if knowledge is sufficient to perform certain tasks. The default value is 1. When a new experience is created, it is assumed that it has been, and still is, sufficient to solve similar problems.
  - Reputation: The reputation of a knowledge source based on previous knowledge interactions. Default value when an agent joins the system is 0.5. See [9] for more information.
- SKMS Category (Values extracted from the SOEKS):
  - Believability: It is the truth value of the SOEKS. The default is value defined by the Prognosis Macro-Process. See [13] for more information.
  - Accuracy: It is the precision value of the SOEKS. The default value is defined in the same way as Believability. See [13] for more information.

The values that can be automatically calculated by the agents contribute to the automation of the measurement process. For instance, if an agent is not able to respond to a user query, it means that knowledge is incomplete and the completeness attribute is modified automatically. On the other hand, the user category contains the attributes that are left for the final user to evaluate. This approach allows the system to receive feedback from the real world and adjust its behavior accordingly.

3.3 Calculating Knowledge Quality in the e-Decisional Community

First of all, it is important to recall that the knowledge measurements hereby presented are approximations of the actual quality of knowledge held by individuals and agents. Based on the elements presented throughout sections 3.1 and 3.2, quality inside the e-Decisional Community is measured in two steps: firstly, the quality of each individual experience belonging to an agent is calculated. Secondly, the quality of all the experiences of an entity is calculated.

Quality for individual SOEKS is defined as the average of all the quality attributes values. All attributes have the same weight because it is considered that an agent's expertise is as important as user feedback. Reputation is used in the final stage of the process, which is described in the next section. Let us define:

$$A = \{a_1, a_2, ..., a_n\}: \text{The set of agents in the system, where } n \text{ is the total number of agents.} \tag{1}$$

$$S(a_i) = \{S(a_i)_1 S(a_i)_2, ..., S(a_i)_m\}: \text{The set of SOEKS of agent } a_i \in A, \text{ where } m \geq 0 \text{ is the total number of SOEKS for that agent.} \tag{2}$$

$$QA(S(a_i)_j) = \begin{Bmatrix} QA_{accuracy}, QA_{timeliness}, QA_{complete}, QA_{relevance}, QA_{understand}, \\ QA_{believe}, QA_{objectivity}, QA_{amount} \end{Bmatrix}: \tag{3}$$

The set of quality attributes defined in section 3.1, for the *j*-th SOEKS $S(a_i)_j \in S(a_i)$, where attributes have values between 0 and 1.

Therefore, the quality measure Q of an individual SOEKS in the e-Decisional Community is defined as:

$$Q(S(a_i)_j) = \frac{1}{8} \sum_{1}^{k} QA(S(a_i)_j)_k \; ; 1 \leq k \leq 8 \tag{4}$$

The values of quality measures for individual SOEKS can be distributed in several ways; there is not a standard model that can be used to predict a specific behavior in the values. Therefore, overall quality calculations are performed using regression analysis, offering the possibility to discover the equation that best fits a set of data samples (i.e. individual quality measures). Consequently, the total quality can be understood as the area under the best fitting curve: as the area increases so does the final quality.The overall quality for an agent in the system, $Q_{Overall}(a_i)$, is obtained by integrating the best fit equation, as follows:

$$Q_{Overall}(a_i) = \int_1^n fit(x) \cdot dx \tag{5}$$

Where: *n* is the total number of individual SOEKS quality measures, and *fit(x)* is the best fit equation for the data samples.

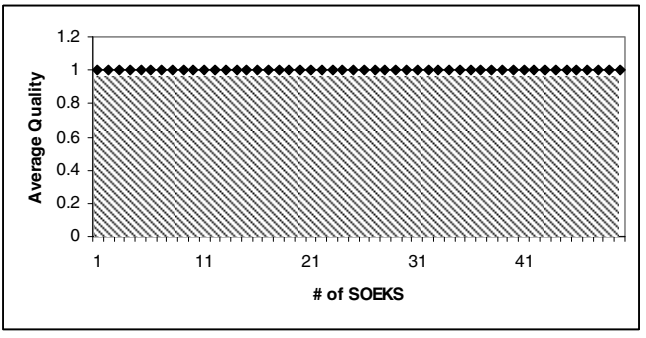


Fig. 1. Example quality for an agent when individual SOEKS quality is 1

example, let us assume that an agent has a total of fifty experiences in its knowledge repository. In this scenario, the agent is a "guru" and the individual knowledge quality for each one of its SOEKS is 1. This is the optimal case, which is illustrated by figure 1.

In the optimal case, the best fit equation is a line in the form $y = mx + b$, with m=0 and b=1. Consequently, the area under the curve is given by the area of the rectangle with length=50 (the total number of samples) and width=1. Then, we have an area of 50, which is equivalent to a 100% knowledge quality.

The overall quality measure is used by agents in the e-Decisional Community as a way to evaluate other peers at the time of conforming Knowledge-Based Virtual Organizations (refer to [9] for more information). Quality values are used as a way to determine which agents are more adequate for engaging in cooperative tasks.

As described in section 3.1, reputation of the knowledge source is an important quality attribute in the e-Decisional Community. When an agent wants to determine which of its fellow entities is more likely to deliver a good result, reputation is taken into account along with overall quality. Therefore, the final criterion to decide if an agent is suitable to participate in a new group is given by the following formula:

$$Q_{Total}(a_i) = R(a_i) \cdot Q_{Overall}(a_i) \tag{6}$$

Where $R(a_i)$ is the reputation of an agent, as defined in [9]. This approach helps in the process of deciding which agent to select when there are several agents with relatively similar $Q_{Overall}$ values.

4 Experiment Design and Results

A prototype implementing the functionality described in the previous section was developed, using Java SE 6, JADE 4.0.1 [18], Symja 0.0.7a [17] and Statcato 0.9.2 [20]. Quality features for individual SOEKS quality measurement were implemented in the Individual Management Layer (IML) package. For more information on the Individual Management layer see [10].

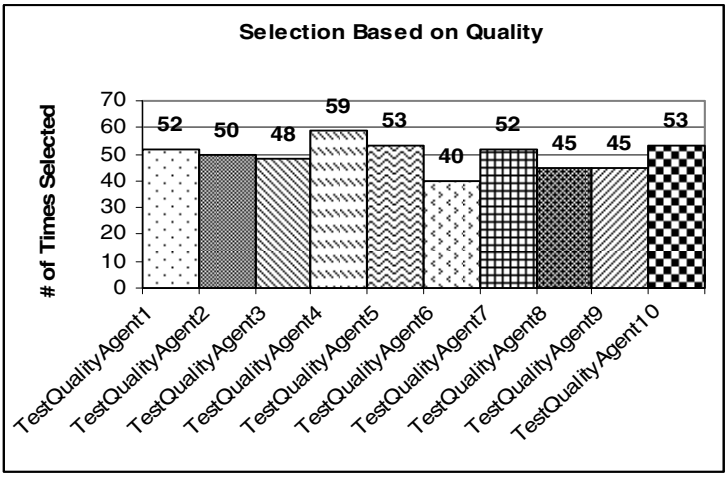


Fig. 2. Selection of agents based on overall knowledge quality

The agent system is comprised of ten working agents, a reputation manager agent, and a quality manager agent. In the experiments, a worker sends a request for the creation of a group; then the system evaluates the request and the manager agents calculate overall quality and reputation values, to finally return a list of the "most knowledgeable" agents. Then, the process is finalized and the experiment is run again. A total of one hundred trials were executed. Quality measures for individual SOEKS were generated using a random number generator; these values are changed between iterations of the experiment. Also, once the final list of agents is returned to the requester, overall quality values are recomputed to simulate the process of user feedback through time and its effects on quality. Each agent has two hundred SOEKS in their respective repositories.

Figure 2 shows the experimental results obtained after measuring the number of times agents are selected as part of a knowledge-based virtual organization, based only on the quality of their knowledge Q_{Total}. Because of the random values generated between trials, all the agents were selected the same number of times on average. This situation helps illustrate a scenario where all of the organizational members are knowledgeable in a particular topic, and can contribute and learn as equals. This is a desired state in the system, given that it will reduce the dependency on "expert" agents, and if some entities leave others can take their place in the knowledge sharing process.

Figure 3 presents a detailed analysis of the quality assessment process for agent 6; it represents the worst case scenario in figure 2. The graphic shows the relationship between overall knowledge, reputation and the total quality for the agent.

During the experiments, it can be observed that the agent has individual quality values around 50% for its SOEKS. Also, reputation for agent 6 had values between 30% and 40%. Reputation has a great impact on the final quality values, and it is an effective method for dealing with duality in results, and for making informed decisions about other agents. Also, the use of reputation enhances the platform with a

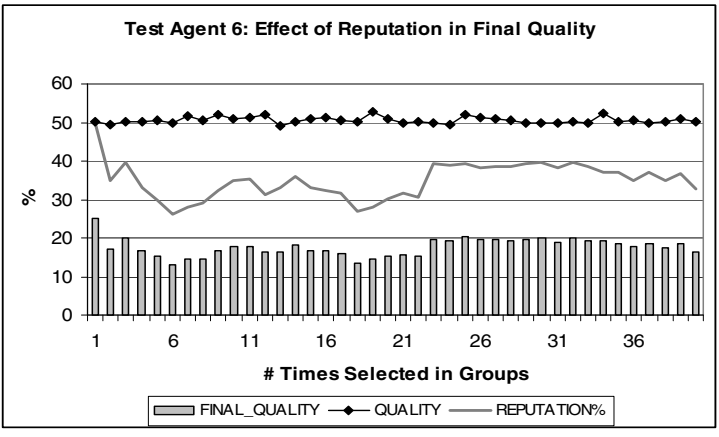


Fig. 3. Effect of reputation over quality for Agent 6

basic ability to resemble social human behavior in group environments. People usually ask for advice from others who can provide new knowledge, but also, individuals who are highly regarded inside groups are preferred for this purpose.

Based on the reputation values for agent 6, the experiments illustrate a possible situation that can be encountered in real life: misinformation. According to Hunt [5], people can strongly believe that they are correct, even when they are not. In the e-Decisional Community, this would mean that a user gives a high score to its SOEKS' quality attributes, based on an erroneous personal conviction. Therefore, an agent will have high quality SOEKS and will be most likely selected for cooperation. After several interactions, other agents/users might realize that the actual knowledge provided by the misinformed agent is not good enough; then, trust levels towards that agent are decreased affecting its reputation inside the community.

5 Conclusion

The model presented in this paper, represents another effort towards providing more accuracy in the process of quality measurement. The assessments obtained from this evaluation, can help users and agents to increase the effectiveness of decision making processes.

Quality measures in the e-Decisional Community can be used as a way to enforce Service Level Agreements (SLAs). Before engaging in a SLA, organizations can evaluate their possible peers based on their reputation and overall quality. However, the proposed technique has not been evaluated in an operational context. In the future, feedback from different industrial and academic sectors is required, in order to refine the approach presented in this paper.

Future work on includes the development of historical measures, as a way to keep track of an agent's knowledge quality over time. Also, the development of a metric to quantify the knowledge possessed by an agent needs to be implemented.

References

1. Bontis, N.: Assessing knowledge assets: a review of the models used to measure intellectual capital. International Journal of Management Reviews 3(1), 41–60 (2001)
2. Darroch, J.: Developing a measure of knowledge management behaviors and practices. Journal of Knowledge Management 7(5), 41–54 (2003)
3. Davenport, T.H., Prusak, L.: Working knowledge: How organizations manage what they know. Harvard Business Press, Boston (1998)
4. Fox, C., Levitin, A., Thomas, R.: The notion of data and its quality dimensions. Inf. Process. Manage. 30(1), 9–19 (1994)
5. Hunt, D.P.: The concept of knowledge and how to measure it. Journal of Intellectual Capital 4(1), 100–113 (2003)
6. Lee, J., Lee, Y., Ryu, Y., Kang, T.H.: Information Quality Drivers of KMS. In: Proceedings of the International Conference on Convergence Information Technology, 2007, pp. 1494–1499 (2007)
7. Linderman, K., Schroeder, R.G., Zaheer, S., Liedtke, C., Choo, A.S.: Integrating quality management practices with knowledge creation processes. Journal of Operations Management 22(6), 589–607 (2004)
8. List, B., Schiefer, J., Bruckner, R.M.: Measuring knowledge with workflow management systems. In: Proceedings of the 12th International Workshop on Database and Expert Systems Applications, pp. 467–471 (2001)
9. Mancilla-Amaya, L., Sanín, C., Szczerbicki, E.: Knowledge-Based Virtual Organizations for the E-Decisional Community. In: Setchi, R., Jordanov, I., Howlett, R., Jain, L. (eds.) KES 2010. LNCS, vol. 6277, pp. 553–562. Springer, Heidelberg (2010)
10. Mancilla-Amaya, L., Sanín, C., Szczerbicki, E.: Smart Knowledge-Sharing Platform For E-Decisional Community. Cybernetics and Systems: An International Journal 41(1), 17–30 (2010)
11. Molina, L.M., Lloréns-Montes, J., Ruiz-Moreno, A.: Relationship between quality management practices and knowledge transfer. Journal of Operations Management 25(3), 682–701 (2007)
12. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Commun. ACM 45(4), 211–218 (2002)
13. Sanin, C.: Decisional DNA and the Smart Knowledge Management System: Knowledge Engineering and Knowledge Management applied to an Intelligent Platform. LAP Lambert Academic Publishing, Berlin (2010)
14. Seawright, K.W., Young, S.T.: A Quality Definition Continuum. Interfaces 26(3), 107–113 (1996)
15. Steedman, I.: On 'Measuring' Knowledge in New (Endogenous) Growth Theory. Old and New Growth Theories: An Assessment, pp. 127–133 (2003)
16. Supekar, K., Patel, C., Lee, Y.: Characterizing quality of knowledge on semantic web. In: Proceedings of the AAAI Florida AI Research Symposium (FLAIRS 2004), Miami Beach, Florida, pp. 220–228 (2004)
17. symja: A Java computer algebra system, `http://code.google.com/p/symja/` (last accessed: April 5, 2011)
18. Java Agent DEvelopment Framework, `http://jade.tilab.com/` (last accessed: April 4, 2011)
19. Tongchuay, C., Praneetpolgrang, P.: Knowledge Quality and Quality Metrics in Knowledge Management Systems. Special Issue of the International Journal of the Computer, the Internet and Management 16(SP3), 21.21–21.26 (2008)
20. Statcato: Free Software for Elementary Statistics, `http://www.statcato.org/` (last accessed: April 5, 2011)

Application of Decisional DNA in Web Data Mining

Peng Wang[1], Cesar Sanin[1], and Edward Szczerbicki[2]

[1] School of Engineering, Faculty of Engineering and Built Environment
The University of Newcastle, Australia
[2] Gdansk University of Technology, Gdansk, Poland
peng.wang@uon.edu.au, Cesar.Sanin@newcastle.edu.au,
Edward.Szczerbicki@zie.pg.gda.pl

Abstract. Web data mining techniques are becoming popular and valuable components of web data analysis systems. It assists website's owners to estimate their website's performance and make explicit and precise business strategies. The main features of Decisional DNA are related to knowledge representation structures. They are dealing with noisy and incomplete data, learning from experience, making precise decision, and predicting. This paper presents a proposal for development of web data mining techniques with Decisional DNA at its core. Integrating Decisional DNA with web data mining techniques involves retrieval, clustering, storage, sharing, and transporting of knowledge and day-to-day explicit experience in a new structure. A set of experiments is also included in this paper to illustrate usage of Decisional DNA applied to decisional domain in web data mining as well as re-usage of such knowledge to facilitate decision making process.

Keywords: Decisional DNA, Set of Experience Knowledge Structure, web data mining, web content mining.

1 Introduction

Today, with the rapid development of computer networks and multimedia technology, Internet has become the major supplier of huge amounts of information. It has been changing human lifestyle on every day basis. In fact, the Internet is a massive information source which has a lot of useless spam, and only a small part of this information is useful [7]. Furthermore, it is difficult to quickly and easily find desired information from websites because of unstructured and dynamical changes in information presentation and content. Though web search engine can help in resource discovery on the web, it is far from satisfying due to its poor precision. On this basis, web data mining research has become a hot spot in the high technology domain. Nowadays, it faces extraction of useful knowledge in order to guide the decision-making from web-based data [7]. This paper introduces a novel and explicit way, integrating web crawler and Decisional DNA [8, 9], which has the ability to easily capture and reuse different structured knowledge from the web.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 631–639, 2011.

2 Background

2.1 Web Data Mining

Web Data Mining is the process of discovering and extracting useful information or knowledge from the Web including web hyperlink structure, page content and usage data [4]. It is an inclusive technology in which several domains are involved, such as Web, data mining, computational linguistics, statistics information standard and other fields of science. In other words, web data mining techniques can be used to analyse the content of documents, the use of available resources, to find effective, potential, valuable, understandable and explicit patterns of knowledge by combining methods of statistics and artificial intelligence with database management [7, 13]. According to different mining tasks, there are three important aspects of web data mining: web usage mining, web structure mining and content mining. Their detailed structure is illustrated after [7] as follows (see Fig.1):

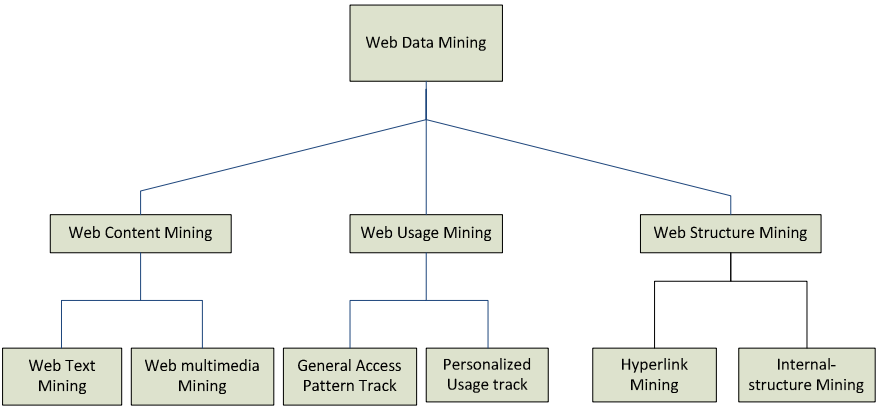


Fig. 1. Classification of Web Data Mining

Web Content Mining. Web content mining comprises the process of discovering useful information from text, image, audio or video data in the web [4]. Furthermore, it is also a data mining technique which is different from traditional data mining techniques because it is primarily for a variety of unstructured data such as text, voice, video and graphic image and so on. The current study of Web content mining is mostly based on document data.

Web content mining has mainly two kinds of text mining objects. Firstly, the text document mining which is composed by the text format, HTML tags, XML tags or semi-structured data and unstructured text among others. Thus, we can use Web document tags, such as `<Title>` `<Heading>` and other additional information, to improve the performance of Web text mining. Secondly, multimedia document mining consisting of image, audio, and other media types [7].

Web Usage Mining. Web usage mining is the ability to automatically discover web user access patterns from web server log files which record the users' access to data.

General access pattern track for user groups and personalized use record track for single user are used to analyse users' browsing patterns. Generally, servers log data should be aimed to mine. These data includes a client-IP, server-side data, authoritative page and data-side proxy. Generally, it uses server log files to find interesting patterns of visiting web sites, which helps on understanding users' behaviour. And, in consequence, it supports websites improvements or personalization of users' service.

Web Structure Mining. Web Structure Mining is the process of analysing node and connection structures of a website by using graph theory. In other words, it acquires knowledge from the organizational structure of a website and the relations among the links. For instance, web structure mining techniques can be used to index pages and seek the most useful pages among them. Web structure data mining is composed of two kinds. One is extracting hyperlinks between web pages or documents and the other one is mining the internal document by analysing the page structure's tree link structure [7].

2.2 Web Crawler

Web crawlers are tools to automatically gather webpages from certain web sites with an orderly pattern. The process of crawling is to use seed URLs to download web pages related to these URLs. Then, they recursively extract and download web pages according to any hyperlinks identified from the URLs. One core component of web search engine is the Web Crawler. Thus, it can be used to assemble the web pages sorted by the search engine. Therefore, many applications apply it to deal with large numbers of web pages including web data mining, comparison shopping engines among others. Major engineering challenges have been bright worth by implementing high-performance web crawler, though its principle is simple [6].

The Heritrix. The Heritrix is an instance of web crawlers. It is an extensible, web-scale and archival-quality open source. It is divided into three striking aspects such as the Scope, the Frontier and the Processor Chains [2]. It brings initial information to creating the Scope with seeds. The seeds contain initial URIs which can be consulted by the Frontier. The Frontier is responsible for which URIs should be ordered to be visited according to the Scope seeds. It maintains a series of internal queues of URIs, ensuring URIs to be not already-scheduled and only choosing the URIs scheduled to be collected.

2.3 Set of Experience Knowledge Structure (SOEKS) and Decisional DNA

Web Data Mining is currently working with different types of knowledge. The idea behind it is to store and manage knowledge in some manners. In other words, mining web data is the process of storing, retrieving, distributing and sharing knowledge. However, web information is mostly unstructured or semi-structured in huge quantities. Thus, a technology which can be used to capture and store formal decisional events as explicit knowledge is necessary. The Set of Experience Knowledge Structure (SOEKS or shortly SOE [12, 11, 14]) as a flexible and independent knowledge representation is a suitable tool for this task. Moreover, it also

has been used to collect and store formal decisional events in an explicit manner [11]. Therefore, the SOEK can be a pattern based on existing and available knowledge offered by a formal decision event with dynamic structure. It can be expressed in XML or OWL as ontology in order to make it shareable and transportable [1, 12, 10].

The SOEKS is composed of variables, functions, constraints and rules [9]. Variables commonly use an attribute-value language to represent knowledge (i.e. by a vector of variables and values) [5]. It is the starting point for the SOEKS and the infrastructure of the SOE because they are the source of other components. Functions are made up of interactions of variables which include dependent variables and a set of input variables. On the other hand, according to the tasks of the decision event, functions are brought to reasoning optimal states. Therefore, this second component of the SOE establishes the relations between variables restricting experience on decision-making. Constraints are another factor of association amongst the variables. Though constraints are another way of functions, they have a different purpose. They limit the performance and possibility of a system and restrict the feasible solutions in a decision problem. Lastly, Rules are another form of expressing links among variables. They condition the relationships that operate the universe of variables. In the other words, they use the statements IF-THEN-ELSE to connect consequence with a condition.

Additionally, the SOEKS is structured in view of some important features of DNA. Firstly, the combination of the four components of the SOE offers distinctiveness, just corresponding to the combination of the four nucleotides of DNA. Moreover, the elements of the SOEKS imitate a gene to connect with each other. In the same way as a gene produces a phenotype, the SOE yields a value of decision with their elements. Each SOE can be categorised and acts as a gene in DNA [11]. A set of SOE in a same category makes up of a decisional chromosome which stores decisional strategies for that category. After this, each module of chromosomes establishes an entire inference tool to offer a blue print of knowledge inside an organization [9].

3 The Decisional DNA-Based Web Crawler

Today, the Internet has been developing very rapidly. There is a huge requirement for sharing, storing, reusing and analysing knowledge among the websites. Using the Web Crawler technique with Decisional DNA is a novel and explicit way for organizations or website owners dealing with their increasingly unstructured number of information. It not only shares knowledge, but also assists in the decision making process.

3.1 Architecture Module Description

SOEKS can be implemented by an architecture that contains four Macro Processes [8, 11] as shown on the top of Fig.2 introducing the architecture of the proposed Decisional DNA based web crawler. Those processes are respectively diagnosis, prognosis, solution and knowledge. This paper describes the necessary key components for a Decisional DNA based Web Crawler by using the above four macro processes. Functions and responsibilities of components are explained as follows.

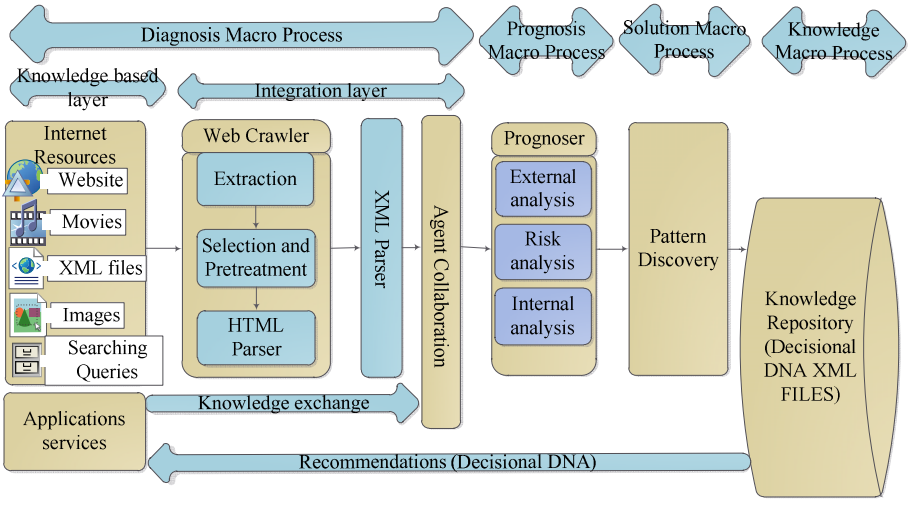


Fig.2. Architecture of Decisional DNA-Based Web Crawler

Internet Resources. The diagnosis Macro-process is composed of knowledge based layer and integration layer. Internet Resources are a component in the knowledge-based layer. Internet contains enormous information. It is useless if information is infinite and non-structured. However, Internet sources, as well, appear in several forms. For example, website documents, emails, metadata, xml files, users' visited logs. Several resources can be related to an organization's knowledge discovery. Thus, the purpose of this component is to define and analyse the scope of website. Then using the Web Crawler component extracts it in the integration layer as shown in Fig.2.

Application Services. Organizations or web owners may need to cooperate with multiple applications inside or outside a company. It is increasingly important to exchange information and knowledge quickly and safely between different applications. Therefore, it needs unified and shareable knowledge. XML is a simple, widely used, transportable and applicable language for sharing knowledge among applications, Decisional DNA formatted in XML language can be exchanged to multiple application services through the Agent Collaboration component (see Fig.2).

Web Crawler. Web Crawler is a component in the integration layer. It extracts certain information from given Internet Resources seeds. Meanwhile, it needs to analyse each given website hyperlinks to find valuable information. Heritrix is used at this stage. It is also responsible for removing useless information like ads, redundant tag format from each page. Afterwards, by using HTML Parsers, it translates related information in order to analyse the pattern of Decisional DNA then passes it to the XML Parser component (see Fig.2).

XML Parser. One purpose of using the XML Parser is to convert gathered information from the Web Crawler to the XML format. In this case, we use it to

transform crawled and selected knowledge into the XML. It is a tool to ensure collected information in order to be formatted as attribute-value mechanism. At next step, that information can be stored through the agent collaboration component.

Agent Collaboration. It can be seen from Fig.2 that the agent collaboration is a container used to collect knowledge from the XML parser or application services. Then it transfers collected information to the prognoser for further extraction. The knowledge in this stage should be formatted as system required.

Prognoser. This process performs a homogenization and unification of information to implement a Multi-objective Evolutionary Algorithm (MOEA) [11]. It generates a holistic group of Sets of experience from which a solution can be chosen. The Prognosis process can be divided into three analyser layers. The Internal analysis layer is responsible for evaluating variables which can be controlled and modified for the website owners. However, the External analysis layer focuses on uncontrollable and unmodified variables. The Risk analysis layer is used to deal with uncertainty, imprecision and incompleteness of the models produced by previous two layers.

Pattern Discovery. The Pattern discovery layer is to find best solution guiding organizations' leader to make decisions. This layer offers a range of variables. The user can choose priorities including value of truth associated, imprecise index, susceptible variables and weights associated with variables.

Knowledge Repository. The knowledge is stored in the Knowledge Repository after pattern discovery. At this stage, information becomes desired knowledge which can be shared and transferred among different applications. Knowledge is stored according to the Decisional DNA structure. In other words, a single set of experience file is a gene of knowledge. A decisional chromosome is composed by many of these genes. And many chromosomes comprise a Decisional DNA. The purpose of the Knowledge Repository is to store and maintain several different Decisional DNAs in order to make them reusable, shareable and transportable among application services.

3.2 Experiment and Case Study

Our plan was to mine the movie website http://www.imdb.com/ and find useful knowledge which can be reused, shared and transported among diverse applications. Three techniques were implemented in this experiment: the Heritrix, the DOM4J Parser and the Set of Experience Knowledge Structure (SOEKS), providing as a result of web mined knowledge extracted and placed in a SOEK form in order to construct a DDNA for movies.

Methodology of Experiment. In the proposed platform, the diagnosis process includes two layers: knowledge based layer and integration layer. The first step starts in the knowledge based layer. The purpose of the experiment is to gather information about the top 250 movies from the website Imdb (http://www.imdb.com). Hyperlinks of the desired web pages are in the web page IMdb Top 250 (http://www.imdb.com/chart/top) (see Fig.3). It can be seen that there are many ads,

Fig. 3. Web page-IMDB Top 250 [3]

tags and pictures not related to our interests in this page. However, we only want to gain access to desire hyperlinks pages inside the red rectangle. Thus, we need to find a pattern for those hyperlinks. If we click any one of those hyperlinks, all pages are from the same URL (http://www.imdb.com/title/). This result in three URLs must be provided to the Heritrix. They are: http://www.imdb.com, http://www.imdb.com/chart/top and http://www.imdb.com/title/. This task can be done by establishing a Frontier class which simply inherits the FrontierScheduler class first. We rewrite the schedule method of this new Frontier to accomplish our goal. When this Frontier class is established, we modified the "Process.option" file in the fold of "conf\modules" in order to be configured by the Heritrix's web interface.

Now, we already got the interest web pages (see Fig.4). The next step is to select and extract necessary information from those pages. The organizations or the website owners may only have interest in information like "description, title, starts, director, genre and ranking" which is showed in Fig.4 in red rectangles. Therefore, this target can be done by using an HTML Parser to extract desired information from tags. For example, when we view the source code of the desired page, the desire information is inside tags such as: "<title>The Godfather (1972) – IMDb</title>". And it can be simply extracted by the HTML parser.

Fig. 4. Fields of interest in the web page [3]

In our experiment, we chose the MirrorWriterProcessor class to store the required files. Therefore, we had to rewrite MirrorWriterProcessor in order to filter the information. In this class, we use a HTML Parser to acquire certain information, and then use DOM4J to translate such information into XML with the required Decisional DNA structure. For our purposes and as an example, the title must be stored as a variable and follows the SOEKS variable's structure [8].

```
<variable>
  <var_name>title</var_name>
  <var_type>CATEGORICAL</var_type>
  <var_cvalue> The Godfather (1972) </var_cvalue>
  <var_evalue> The Godfather (1972) </var_evalue>
  <unit></unit>
  <internal>false</internal>
  <weight>0.0</weight>
  <l_range>0.0</l_range>
  <u_range>0.0</u_range>
  <categories>
    <category></category>
  </categories>
  <priority>0.0</priority>
</variable>
```

There are six variables acquired as discussed above to separately indicate title, description, starts, director, ranking and genre. Those variables are iteratively and automatically stored in a SOEKS-XML file as a gene until crawling is finished. Next step, we use the Prognoser to analyse those variables and find best solutions for different purposes but that is part of a future work. Finally, the Decisional DNA is stored in the Knowledge Repository which can be reused or transported to other applications.

Decisional DNA-Based Web Crawler Experimental Results. This decisional DNA-based web crawler is implemented purely in java on windows 7 operational system. It holistically traversed the website in 2 hours 15 minutes 58 seconds and totally discovered 27645 URIs, each of them consumed 0.2 second to gather the required movies' information. We identify that three factors affect the web crawling speed. They are internet connection, web services' limitations and capability of websites. All those factors occur in any web crawler component. However, the gathered information is effectively and efficiently converted into Decisional DNA structure with minimal time consuming. In consequence, there will be a better performance when those factors are reduced; nevertheless, reducing those factors is not part of our research. Finally, Those 250 movies were converted to decisional DNA-based structure which is able to be reuse for any purposes by multiple applications.

4 Conclusion and Future Work

This paper presents an experimental integration of Decisional DNA into a Web Crawler for web data mining purposes. This integration can be applied to gather and convert unstructured and semi-structured data into knowledge that can assist decision

making process. As Decisional DNA in XML format is shareable and transportable, our proposed architecture can be used by diverse applications.
This research represents an initial stage of combining the Decisional DNA with web data mining techniques. Future research will focus on the following:

- Refinement of the requirements of Decisional DNA for dynamic web data mining. Interaction of web site with Decisional DNA and assistance of decision making needs to be researched in detail.
- Experiencing extraction and inference of multimedia data from web data mining.
- Exploration of Prognosis, Solution and Knowledge processes and findings a way to make websites able to automatically gain knowledge from visitors. Therefore, it can simulate user's behaviour to adapt different clients' needs.

References

1. Nguyen, N.T., Duong T.H., Jo, G.S.: Constructing and Mining: A Semantic-Based Academic Social Network. Journal of Intelligent & Fuzzy Systems 21(3), 197–207 (2010)
2. Mohr, M.K.G., Stack, M., Ranitovic, I.: Introduction to Heritrix, an Archival Quality Web Crawler. In: 4th Intl. Web Archiving Workshop, IWAW 2004 (2004)
3. I. IMDb.com. IMDb Top 250, `http://www.imdb.com/chart/top`
4. Liu, B.: Web Data Mining Exploring Hyperlinks, Contents, and Usage Data. In: Web Data Mining, pp. 1–12. Springer, Heidelberg (2007)
5. Lloyd, J.W.: Learning Comprehensible Theories from Structured Data. In: Mendelson, S., Smola, A. (eds.) Advanced Lectures on Machine Learning. LNCS (LNAI), vol. 2600, pp. 203–225. Springer, Heidelberg (2003)
6. Najork, M.: Web Crawler Architecture. Springer, Heidelberg (September 2009)
7. Mukthyarazam, M.K.K.S., Rasool, S., Jakir Ajam, S.: Web data mining Using XML and Agent Framework. IJCSNS International Journal of Computer Science and Network Security 10(5) (May 2010)
8. Sanin, C.: Decisional DNA and the Smart Knowledge Management System: Knowledge Engineering and Knowledge Management applied to an Intelligent Platform. LAP Lambert Academic Publishing (2010)
9. Sanin, C., Mancilla-Amaya, L., Szczerbicki, E., CayfordHowell, P.: Application of a Multi-domain Knowledge Structure: The Decisional DNA. In: Nguyen, N., Szczerbicki, E. (eds.) Intelligent Systems for Knowledge Management. SCI, vol. 252, pp. 65–86. Springer, Heidelberg (2009)
10. Sanin, C., Szczerbicki, E.: Extending Set Of Experience Knowledge Structure Into a Transportable Language Extensible Markup Language. Cybernetics and Systems: An International Journal 37, 97–117 (2006)
11. Sanin, C., Szczerbicki, E.: Experience-based Knowledge Representation: SOEKS. Cybernetics and Systems: An International Journal 40, 99–122 (2009)
12. Sanin, C., Toro, C., Szczerbicki, E.: An OWL ontology of set of experience knowledge structure. Journal of Universal Computer Science 13(2), 209–223 (2007)
13. Wang, J., Huang, Y., Wu, G., Zhang, F.: Web mining: knowledge discovery on the Web. In: Proceedings of 1999 IEEE International Conference on Systems, Man, and Cybernetics, IEEE SMC 1999, vol. 2, pp. 137–141 (1999)
14. Zhang, H., Sanin, C., Szczerbicki, E.: Decisional DNA applied to robotics. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part II. LNCS, vol. 6277, pp. 563–570. Springer, Heidelberg (2010)

A Concept for Comprehensive Knowledge Management System

Bartosz Kucharski and Edward Szczerbicki

Gdansk University of Technology, Gdansk, Poland

Abstract. In real life organizations, the holistic implementations of knowledge based principles from top strategic level to day to day operations are usually missing. The purpose of this paper is to propose an approach that could bridge this gap. Our approach is illustrated with a case study for a company where we integrate Strategic Management with Business Process Management and introduce a decision support and knowledge management capabilities to process, store, share and exploit the knowledge gathered in routine decisions.

Keywords: Strategic Management, Knowledge management, Decision events, Business Process Management.

1 Introduction and Background

For strategic management systems a scorecard is the means of focusing on critical processes to achieve organisational goals [1]. In the illustrative case presented, it is assumed that an organization which is knowledge oriented will score the initiative of introducing routine decision event knowledge management high enough to be worthy its implementation. When it comes to realization from strategy and mission point of view it is usually the perspective of internal business processes that needs to be considered. Execution of business processes can be supported by Information Technology (IT) tools. In our example those tools should be equipped with knowledge management capabilities which integrate and co-operate with central knowledge management tool for management on corporate level.

1.1 Key Concepts

Strategic management is a field that deals with major initiatives, both intended and emergent, undertaken by general managers on behalf of all stakeholders, involving utilization of all available resources, to enhance the performance of a company in its external environment [2]. It represents the highest level of management in the sense that it is the broadest - applying to all parts of a company - while also incorporating the longest time horizon. It gives direction to corporate values, corporate culture, corporate goals, and corporate missions. Under this broad corporate strategy there are typically business-level competitive strategies and functional unit strategies [3]. This

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 640–649, 2011.

management approach is especially well suited for justifying initiatives from the perspective of how they fulfill high levels achievement factors versus overall costs.

Balanced scorecard supplements traditional financial measures with criteria that measure performance from three additional perspectives – those of customers, internal business processes, and learning and growth [4]. The power of the concept developed by Kaplan and Norton in [3] is evident when it is used for transition of vision and strategy into implementation and achievement at all level of operation [5]. The four perspectives shown in Figure 1 allow companies to track their financial results and monitor the progress in building capabilities and acquiring assets they would need.

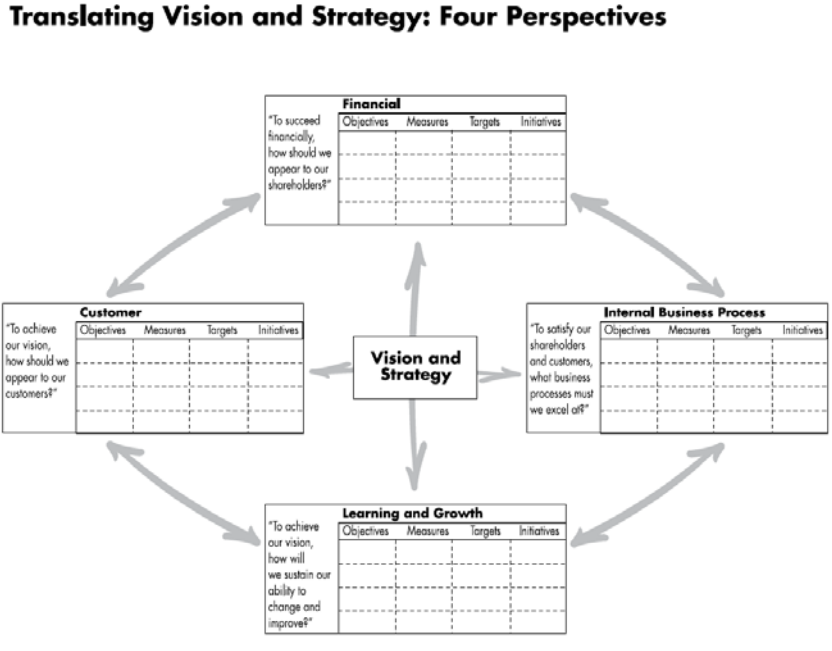


Fig. 1. The four perspectives used in evaluating initiatives [3]

Form knowledge management perspective Figure 1 means taking onboard four important additional management processes:

- *translating the vision* – to build consensus around the organization's vision,
- *communicating and linking* – to communicate the strategy up and down across management levels and link it to objectives,
- *business planning* – to integrate the business and financial plans,
- *feedback and learning* – for being able to adopt, monitor and evaluate.

Business Process Management (BPM) methodology focuses on process related activities and is considered as suitable framework for current process-centric trends in management as it addresses the interplay of people and organization on the one

hand and process-aware software on the other [6]. A Business Process (BP) is a "series or network of value-added activities, performed by their relevant roles or collaborators, to purposefully achieve the common business goal"[7]. BP activities are always present and usually are well-established in the regulations and system specifications for a given company. This philosophy is similar to Total Quality Management (TQM) or Continuous Improvement Process CIP) approaches, however the BPM goes a step further by ensuring that its approach can be supported, or enabled, through technology to ensure the viability of managerial activities in times of organizational stress and change [8].

Business Process Management System or Suite (BPMS) is a set of IT tools for supporting BPM methodology. The underlying idea of BPMS integrates the workflow defined Workflow Management Coalition [9], historical approach to BPMS based on Service Oriented Architecture (SOA), and Business Process Execution Language (BPEL) engine. The term "workflow" is still in use and describes the execution engine for a defined business processes with a human interaction.

Knowledge Management Tool (KMT) is understood as an IT medium for managing knowledge flow at corporate level of a company. It's an aspect specific database with appropriate set of tools including data mining, decision support, and business intelligence. This database should be integrated with a corporate data warehouse and with other business systems to be able to transform data and information into knowledge. The reporting feature is not important here and the main goal of KMT is to keep valuable operational knowledge stored in various notations of business processes. Such tool should provide learning and unlearning capabilities for an organization, and management of metadata should include reasoning aspect and inconsistency administration[10]. For integration abilities, an organizational ontology should be managed not only in modeling aspect of KMT, but also enable bottom-up and top-down tracking from and to a specific business process.

Set of Experience Knowledge Structure (SOE) has been developed to store formal decisions events in an explicit way [11]. SOE is a formal model of experience based knowledge related to every-day decision making events. From KMT perspective it is a notation which has four composite components: variables, functions, constants and rules stored in one entity. A single SOE cannot represent experiential knowledge of a whole system, even in a specific area or category. Therefore, a number of SOEs representing possibly all decisional events related to business processes of a company should be captured and stored into what is called Decisional DNA (DDNA)[12].. In nature, deoxyribonucleic acid (DNA) contains "...the genetic instructions used in the development and functioning of all known living organisms. The main role of DNA molecules is the long-term storage of information. DNA is often compared to a set of blueprints and the DNA segments that carry this genetic information are called genes"[13]. The idea behind DDNA is to develop an artificial system, an architecture that would support discovering, adding, storing improving and sharing information

and knowledge among agents and organizations through experience establishing "genetic instructions" enhancing organizational performance. The use of the analogy between the combination of the four nucleotides of DNA and the four components that characterize decision making actions was coined by Sanin and Szczerbicki[11] about five years ago. Some of the principles for such integration and uniqueness from the perspective of capturing experience and history of decision making in an organization has been mentioned by others but has never evolved into rigorous formalized approach. The DNA metaphor has also been used successfully in terms of formalizing organizational knowledge in stipulations of organization performance (decision rights, information, motivators, and structure). In this sense the proposed Decisional DNA complements and significantly expands these directions of organizational research with formalization of the technological aspects.

2 An Initiative of Introducing Experience Based Decision Support in a Business Process Execution

Every idea that is considered for implementation must be put into some framework for introducing new solutions. For the initial conceptualization of the approach we will use a project charter. A project charter is document that formally authorizes a project or phase and documenting initial requirements that satisfy the stakeholders' needs and expectations. It establishes a partnership between the performing organization (or customer in the case of external projects). The approved project charter formally initiates the project[14]. Before signing the project charter each new idea to be considered is evaluated using the balanced scorecard. After signing the idea it is implemented in corporate environment and BPMS is adopted as the means of experienced based knowledge management using SOE as knowledge representation and structure. For creating the project charter we could use a template from www.projectmanagementdocs.com[15].

2.1 Project Lunch Decision - Justification

Before authorizing the execution of a project there must be a decision making process applied that follows the adopted strategic management approach. In our case for evaluating the Smart Routines project we use the balanced scorecard tool. This is the place where a new initiative must be evaluated against the strategic objectives that come from the mission statement and are provided during strategic analysis process and finally defined in the balanced scorecard. Our assumption is that the illustrative organization is knowledge oriented and has certain strategic objectives shown on a program selection grid of the balanced scorecard template in Tables 1 and 2 [16].

For simple illustration purposes we show in Tables 1 and 2 only strategic objectives that are affected by the initiative. Those strategic goals are not the expected effects of the Smart Routines project, but all objectives that are listed can be partially achieved by introducing the proposed new solution.

Table 1. Assumed strategic objectives evaluated against the Smart Routines project

Balanced Scorecard **Plot Impact of Programs against** **all Strategic Objectives**		**Program**	*SMART ROUTINES*
Perspective	**Strategic Objective**		
Financial	Reduce personnel costs by 10 %		*
	Reduce losses caused by wrong decisions - reduce number of contracts with loss by 10%		*
Customer	Fast service - reduce average awaiting time from 30 to 20 minutes		*
Internal Processes	Lean process - reduce unnecessary tasks or activities		*
	Reduce task execution time by 10%		*
Learning & Growth	Fast deployment of new workers - reduce introduction time in operation department by 40%		*
	Mitigate organizational knowledge los caused by personnel rotation		*
	Learn from the best - score and evaluate best performers		*
	Lunch another branch office in one of developing countries		*

Table 2. Impact of Smart Routines project on the Strategic Objectives

Strategic Objective	**Impact of Smart routines project**
Reduce personnel costs by 10 %	Decision support will reduce average decision time and therefore reduce the workload on decision making role
Reduce losses caused by wrong decisions - reduce number of contracts with loss by 10%	The overall decision quality will be improved especially the decisions made by inexperienced personnel.
Fast service - reduce average awaiting time from 30 to 20 minutes	Due to on-line available suggestion the decision task can be performed faster
Lean process - reduce unnecessary tasks or activities	Storing decision events in unified structure like SOE enables additional optimalization even in some cases reducing the need of making the decision by human
Reduce task execution time by 10%	Due to on-line available suggestion the decision task can be performed faster. Case based reasoning enable intuitive learning by example.

Table 2. (*continued*)

Fast deployment of new workers - reduce introduction time in operation department by 40%	Case based reasoning enable intuitive learning by example.
Mitigate organizational knowledge loss caused by personnel rotation	Stored decision events in formal structure mitigates the effects of attrition
Learn from the best - score and evaluate best performers	By evaluating decision events form performance and quality perspective it is possible to find the best workers.
Lunch another branch office in one of developing countries	Formal decision event structure with formalized processes can be migrate else ware as a startup

3 Solution

3.1 Technical Description

The Smart Routine project will use architecture based on java technology and will follow open standards in the area of defining and exchanging artifacts as decision events, processes, and interface definition. As a platform it uses well establish business process suit from jboss in version 5[17]. The jBPM platform already contains component called Guvnor for storing definition of processes business, rules and other static content. From Smart Routines project perspective there is no proper place for storing data from decision events. The decision events can be easily caught using the WS-HumanTask[18] specification that assumes a task life cycle with 5 main status levels: “Created”, “Ready”, “Reserved”, “InProgress”, and “Completed”. In this case we have to set a listener on leaving the decision task in a process with status “Complete”. This capability is supported in jBPM API in the `ProcessEventListener` interface - in fact there are two suitable methods that can be applied `beforeNodeLeft` and `afterNodeLeft`. One of those must be implemented to store decision event in a experience database. As knowledge and experience representation we use SOE structure and the only thing to ensure the proper implementation is the mapping between the data in the process instance and SOE. The proposed mapping is presented in Table 3.

The mapping could be done using data that already exists in BPMS and other repositories. However, there still exists the need for integration process implementation that enables the exchange between business process instances and experience database. Same integration degree comes from BPMS perspective as an automated task in process definition that performs a specified action. The data exchange could be done using native SOE XML format by introducing specialized services for saving the decision events and for find appropriate experience from the

Table 3. Conceptual mapping between process instance data and SOE

BPMS data	Set of Experience Knowledge Structure (SOE)	Mapping or source
Process instance variables	Variables	1:1
Process definition - internal business rules, evaluation logic, and flow logic	Functions	Extracted from process definition
Process definition – from business rules and task allocation logic; only relations between variables and organizational structure	Constrains	Extracted from process definition and organizational structure from user management system
Process definition – business rule and flow logic – higher level logic – on condition consequence basis	Rules	Extracted from process definition and organizational structure from user management system

past. The saving service is just simple storing operation, but finding it is more complex service with a distance function for picking up the closest case following the case base reasoning decision support model. The output from the decision support mechanism could be used in the same way that the output from business rules already present in the BPMS, there is no need for introduction of any new mechanism for presenting the results. For finding the closest case the FreeCBR[19] java library can be used. This tool is able to both find the closest case for given distance function and return the relevance ratio for this case.

Both distance function and recorded experience is a matter for knowledge management tool - the DDNA Manager software platform. The Manager enables SOE visualization and provides simple operations for experience management. The distance function is a concern because it joins both processes and decision events. Currently there is no support provided for distance function neither in jBPM nor in DDNA Manager but such functionality is planned in the second tool. At this stage of the Smart Routines project it is assumed that distances will be kept as simple functions in plain XML files referencing to a process definition and a decision point.

3.2 Operation Conditions

From post Smart Routines project perspective there are some additional operations to be performed. Experience database must be maintained, the decision events should be evaluated, the processes for learning and unlearning applied. Thus, the Smart Routines assumes introducing a new role in an organization. This role will be assumed by knowledge engineer with responsibilities specified in Table 4.

Table 4. Responsibilities of a knowledge engineer

Responsibility name	Description
Evaluate decisions events	Finding measures for automatic assessment of decision events (i.e. scoring by final case result). In case of manual scoring set an owner for the task and agree on criteria for scoring.
Learning processes	Set a learning process for gained decision events, provide a method for finding representative cases, adop existing decision events to changes in an environment
Unlearning processes	Set a unlearning process, remove irrelevant experience data form decision support, evaluate experience database versus actual situation (Especially changes in processes and internal or external environment)
Optimize decision support	Tune the distance function, measure the quality of decision support, maintain changes that comes from business processes
Share experience	Prepare representative decision events database for learning by example purposes. Consult with business process designers the ability and factors that can affect decision events scoring and case based reasoning decision support. Extract rules that goes from the decision events and help implement them into business processes

3.3 Requirements Fulfillment

From strategic objectives point of view there must be a measure introduced that can directly indicate the level of activity fulfillment. Some of the metrics can be adopted directly from the BPMS. This platform is equipped with Business Activity Monitoring that enables monitoring the task currently performed and task duration time. Expected values of task execution time can be evaluated against data from the past. The quality of decisions is measured indirectly by decision event score and the final outcome is measured as earned value. The knowledge transfer goals can be measured be comparison of time required to introduce a new worker to a given task before and after application of the proposed approach. The process optimization can

be measured in number of new business rules implemented in processes that were discovered by a knowledge engineer. For the objective "lunching another branch" the measures can indicate the number of valid experience records versus all duplicated decision points within processes in a new place.

4 Summary

The briefly introduced approach for applying the idea of a company being knowledge oriented is meant as a general conceptual template that can be tailored to knowledge related initiatives in certain environment. First of all the vision of an organization being knowledge oriented must be decomposed into some strategic objectives that can be addressed very specifically. The presented approach shows that the majority of knowledge management processes can be supported and retrieved from standard basic frames for a business project. Of course introducing completely new idea of gathering and using experience record in BPMS generates some non standard risks, but the odds for a success are very good. With the tools for managing experience fully developed it could even be a commercial product that is very much needed in our knowledge based society and economy.

References

1. Nag, R., Hambrick, D.C., Chen, M.-J.: What is strategic management, really? Inductive derivation of a consensus definition of the field. Strategic Management Journal 28(9), 935–955 (2007)
2. Strategic management, http://en.wikipedia.org/wiki/ (accessed 03.28.2011)
3. Kaplan, R.S., Norton, D.P.: Using the Balanced Scorecard as a Strategic Management System, pp. 75–85. Harvard Business Review (January-February)
4. Rooseman G.E.: Towards a Balanced Scorecard to measure design effectiveness in corporate identity design, p. 4. INHOLLAND University Graduate School (March 2004)
5. Surma, J.: Business Intelligence Systemy wspomagania decyzji biznesowych, p. 59. Wydanictwo Naukowe PWN SA, Warszawa (2009)
6. Kim, Y.G., Park, S.C., Kim, C.Y., Kim, J.H.: An Effective Content Management Methodology for Business Process Management. In: van der Aalst, W.M.P., Benatallah, B., Casati, F., Curbera, F. (eds.) BPM 2005. LNCS, vol. 3649, pp. 416–421. Springer, Heidelberg (2005)
7. Ko, R.K.L.: A computer scientist's introductory guide to business process management (BPM). ACM Crossroads 15(4), 11–18 (2009)
8. Business process management, http://en.wikipedia.org/wiki/ (accessed 03.28.2011)
9. Business process management system, http://www.wfmc.org/ (accessed 03.28.2011)
10. Sliwko, L., Nguyen, N.T.: Using Multi-agent Systems and Consensus Methods for Information Retrieval in Internet. International Journal of Intelligent Information and Database Systems 1(2), 181–198 (2007)
11. Sanin, C., Szczerbicki, E.: Using Set of Experience in the Process of Transforming Information into knowledge. International Journal of Enterprise Information Systems 2(2), 40–55 (2006)

12. Sanin, C., Szczerbicki, E.: An OWL Ontology of Set of Experience Knowledge Structure. Journal of Universal Computer Science 13, 209–223 (2007)
13. Sinden, R.R.: DNA Structure and Function. Academic Press, San Diego (1994)
14. A guide to the project management body of knowledge, 4th edn., p. 45, Project Management Institute, Inc. (2008)
15. Project Charter Template, http://www.projectmanagementdocs.com/templates/Project%20Charter.doc/ (accessed 03.28.2011)
16. ScoreCard Template, http://www.exinfm.com/excel%20files/Balanced_Scorecard_Templates.xls (accessed 03.28.2011)
17. JBPM, http://sourceforge.net/projects/jbpm/files/jBPM%205/jbpm-5.0-Final/ (accessed 03.28.2011)
18. WS-HumanTask, http://download.boulder.ibm.com/ibmdl/pub/software/dw/specs/ws-bpel4people/WS-HumanTask_v1.pdf/ (accessed 03.28.2011)
19. FreeCBR, http://freecbr.sourceforge.net/ (accessed 03.28.2011)

The Role and Concept of Sub-models in the Smart Fuzzy Model of the Internet Mortgage Market

Aleksander Orłowski[1] and Edward Szczerbicki[2]

[1] Gdansk University of Technology,
Faculty of Applied Physics and Mathematics PhD Program
aorlowsk@zie.pg.gda.pl
[2] Gdansk University of Technology,
Faculty of Management and Economics
Edward.Szczerbicki@zie.pg.gda.pl

Abstract. The paper introduces some challenges of the fast growing mortgage market in Poland. One of these challenges is the need for a model development that could be used for various predictions related to this market. At the current stage of the model development process our main goal is to propose and introduce sub-models the role of which would be to describe three different economic environments: stable, fast growing, and recession. After proposing the above three sub-models the paper concludes with directions of further research in this area.

Keywords: fuzzy logics, soft modeling, internet mortgage market, fuzzy model.

1 Introduction

The paper deals with the topic of fast growing and very dynamic part of today's market, i.e. Internet based mortgage market. Polish internet banking market, consists of 4 main stages as shown in Figure 1 [1]. Figure 1 follows the approach of descriptive modelin at conceptual stage of problem formulation and shows the connections that exist on the mortgage market. Starting from the top, banks are institutions that sell mortgage and offer the option to apply for it on their own web pages. Because the market is very substantial in size and is characterised by strong competition, banks allow their partners to sell their financial products on the partners' pages to generate more sold products. These companies are typical brokers that receive the commission for every product that they sell. They promote and sell banks' products on their pages but they also create a network called "partner system" which allows the owners of small web pages to sell the products of the banks on their own pages. The owners of private web pages do not sell enough products to cooperate directly with banks, which is the reason for their association with brokers. For each product sold on a partner web page, its owner receives commission from the broker, who in turn gets his from the bank.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 650–656, 2011.

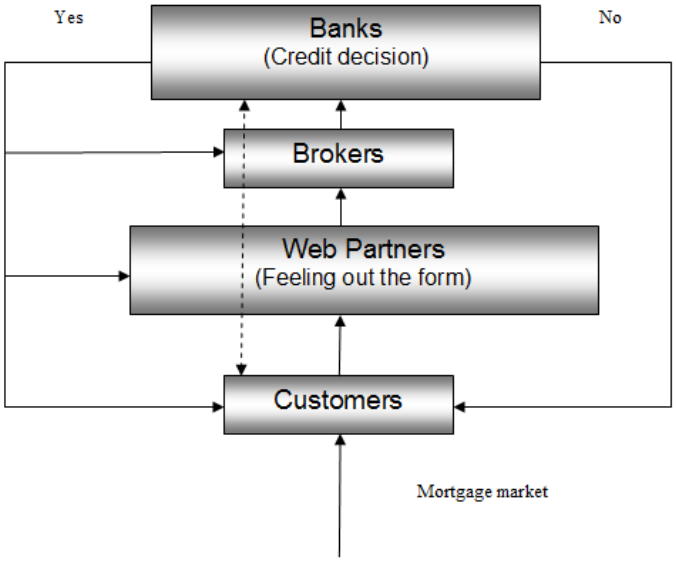


Fig. 1. Formal description of the general mortgage market model [1]

Each level in Figure 1 represents different participants of the market. On the first level there are banks, the second - brokers, level 3 consists of Partner web pages. Customers make up the lowest level. It was the intention of Figure 1 to describe the market from the customers' point of view. The descriptive model presumes numerous customers looking for mortgage using web searchers to find the best offer. Doing so, they reach partner web pages offering different mortgages. It needs to be pointed out that top positions of search results are bought by the owners of private web pages or brokers. A customer chooses one of the web pages (mostly the top one in the given result of the web search) and moves to it. This web page provides information about the credit and often adds calculators that enable the customer to find out his/her credit rating. There is also a special link with the application template for a loan in the chosen bank. While visiting the web page a customer fills out the application for the mortgage, which is subsequently sent to the broker with whom the partner has signed a contract. Next the broker sends this application to the bank with which the broker has signed the contract in the first place. The bank begins the credit procedure, contacts the client, checks the credit rating, and makes a decision.

The main challenge of the strategic nature for this briefly described market is making accurate predictions about the number of sold mortgages, especially in the part of mortgages sold by web partners. Due to the fact that there are several variables of different nature (quantitative and qualitative) influencing the market, traditional statistical methods can't be used here because they don't work properly.

Variables in "hard" mathematics take numerical values, while in "soft" fuzzy logic applications the non-numeric linguistic variables are often used to facilitate the expression of rules and facts [2],[3]. This is an important modelin aspect in this case where a number of variables such as for example "current market feelings" can't be

expressed numerically. As stated in [4] the complete description of a real life system often would require far more detailed data than a human being could ever comprehend simultaneously, process, and understand. To address the above challenge a dedicated rule based fuzzy model for predicting the number of sold mortgages using web partners is proposed and developed. The proposed model is described in the sections that follow.

2 The Process of Model Development and Model's Structure

In the initial approach to the problem of developing a proper prediction mechanism for the Internet mortgage market a dedicated rule based model was created. The model, which was developed and embedded in times of fast growing period of market and economy, did not work properly in the times of financial crisis that came later, so it was necessary to introduce some changes to this model. This need triggered the idea of proposing sub-models that could account for different market situations.

2.1 The First Modelin Generalisation: Rule Based Model

The first version of rule based model was created in the year 2008, and was based on the Internet mortgage market experience of one of the authors encompassing the period of 2003-2008. As such it represented market conditions from that time period (fast growing market with highly positive prospects for the future). The model consisted of 240 rules divided into two scenarios: the positive and the negative one. The general production rule in the model looked as below:

Production Rule:

IF variable_1 is value_1 **AND** variable_2 is value_2 **AND** variable_3 is value_3 **AND** variable_4 is value_4 **AND** variable_5 is value_5 **AND** variable_6 is value_6 **THEN** result will increase value_7 (1)

The rule consists of variables and its values presented below:

- Variable_1 = [Commission]
- Variable_2 = [Interest rates]
- Variable_3 = [Advertising]
- Variable_4 = [WIG]
- Variable_5 = [IbnGR]
- Variable_6 = [WNE]
- Result = [Selling mortgage in the Internet]
- Value_1 = [Small, medium, high]
- Value_2 = [Very small, small, medium, high]
- Value_3 = [Very small, small, medium, high, very high]
- Value_4 = [bad, average, good, very good]
- Value_5 = [bad, average, good, very good]
- Value_6 = [very bad, bad, average, good, very good]
- Value_7 = [Very small, small, medium, high, very high] [5]

After analyzing the model it was first decided that there might be problem with correlation of variables in the model. Generally, the pre-defined initial group of 6 variables had unnecessarily strong representation of very general economic indicators which were highly correlated with each other. The correlation was checked and variables in the model were optimized. After dealing with the correlation issue, the problem with market situation came together with the 2008 financial crisis. A worldwide economic situation has changed rapidly and the deep financial disturbance reached the mortgage market. The model, which was developed and embedded in times of fast growing period of market and economy did not work properly in the times of financial crisis that came later, so it was necessary to introduce changes to the developed model.

As the result of the introduced changes, the number of general economic variables was reduced and the final model includes 4 variables as listed below:

- Variable_1 = [Commission]
- Variable_2 = [Interest rates]
- Variable_3 = [Advertising]
- Variable_4 = [WIG – Warsaw Stock Exchange main indicator]

The above variables were represented a proper balance of general economical variables, market expectations, and the point of the view of the owners of partner web pages.

2.2 Fuzzy Model

The redefined above rule based model did not work as well as expected. The main reason for this was that due to the specific character of the mortgage market there are several variables, which influence o the market, that can't be described by numbers using crisp values needed for hard mathematical modeling. It was necessary to try a different approach and that is why the fuzzy modeling was suggested. The process of building fuzzy model was presented in three steps: fuzzification, fuzzy inference**,** and defuzzification.

In the process of fuzzy model development the production rule base was used which consisted of 81 production rules. This number comes from the number of variables (four variables) in the model and values of these variables (three linguistic values for each variable):

$$3^4 = 81 \text{ [6]}$$

Each rule in the rule base is developed using the IF... THEN logical construct consisting of four variables as in the following example:

Production Rule:

IF Commission is *small* **AND** Interest rates is *small* **AND** Advertising is *small* **AND** WIG is *small* **THEN** Selling mortgage in the Internet is *small*

To develop and train the inference engine for our case it is necessary to specify values of output variables for the existing 81 production rules. As it was not possible to automatically generate these output values, the expert market knowledge was used for

this purpose. For defuzzyfiaction the Height Method (also know as Max-membership principle) was used. [7]

Finally the fully developed fuzzy model was created and basic tests were made.

3 Sub-models and their Role in the General Mortgage Market Model

3.1 The Background

First, as presented in the previous chapter, one general fuzzy model was created which was based on the origin idea of developing a single model to represent the marker reality described in Section 1. The first problems appeared during the process of choosing the variables for the model; it seemed impossible to find variables representing the whole market described by the data from different time periods. In each time period different variables were representing the market conditions. Based on the past results (data from years 2003-2010) it can be concluded that creation of one comprehensive model does not give proper results, especially when the market changes. Due to this fact it was decided to divide the model into three single sub-models which will be suited to different market conditions.

Based on the theory of economic cycles it was decided to compose the general model of three sub-models representing three main market conditions: fast growing market (boom), recession, and moderate growth. These are the most distinct stages we usually go through in any one economic cycle. As it was important to keep the proper balance between the number of sub-models proposed and the level of differences influencing the market, the three sub-models are seen as a proper solution.

3.2 Sub-models Concept

The concept of our idea includes the description of three sub-models with definition of sub-model market conditions, and usage of variables with their importance in the model.

Sub-model nr 1: Stable market

This sub-model is described for market in normal conditions which means stable growth (~2% Gross Domestic Product (GDP) for Central-Eastern European Countries) and is a situation between recession and boom in an economic cycle. This market situation is best illustrated by the data from Polish mortgage market in years 2004 and 2005.

In the above market conditions the impact of non-economical variables is rather small, among from economical and Public Relations (PR) aspects only general and rather stable expectations may be included in the model. Due to these facts it is suggested that the following 4 main variables are this sub-model:

- Variable_1 = [Commission]
- Variable_2 = [Interest rates]
- Variable_3 = [Advertising]
- Variable_4 = [WIG]

Due to the stable situation on the market it is suggested to use the same influence factor of all variables in this sub-model.

Sub-model nr 2: Fast growing market

This sub-model describes the market in the fast growing economic conditions. This is the peak phase of economic cycle in which expectations for the market are extremely high and the number of sold products is also very high but, due to this facts, the risk of market operations is also high. This sub-model might be illustrated by the situation on Polish mortgage market in the year 2007.

The influence of non-economical aspects of the market is significant, especially the role of sociological factors. The following 6 variables are suggested for this sub-model:

1. Variable_1 = [Advertising]
2. Variable_2 = [Commission]
3. Variable_3 = [WIG]
4. Variable_4 = [Customers_market_reviews]
5. Variable_5 = [Number of planning permissions]
6. Variable_6 = [Interest rate]

Sub-model nr3: Recession

This model describes the market in the recession period, the bottom of the economic cycle. So, as in the fast growing market, there is high influence of non-economic aspects on the market but in this situation in a negative way. The aspects and predictions are mainly negative so are the market emotions. The sub-model situation might be represented by the market behaviour in Central-Eastern European (CEE) countries in the last two quarters of the year 2008.

Focusing on the mortgage market, its main players are banks which decide if they offer the product (mortgages) on the market or not. Other economical indicators are less important so as the PR and marketing aspects what is shown in the variables proposed of this sub-model.

1. Variable_1 = [The ability of getting mortgage]
2. Variable_2 = [Interest rate]
3. Variable_3 = [Number of planning permissions]
4. Variable_4 = [WIG]
5. Variable_5 = [Commission]

4 Conclusion and Future Works

The first part of the paper describes the Polish Internet mortgage market with its problems and challenges. One of the challenges is the development of a model that could be used for prediction purposes. Next, we introduce the concept of such model development, starting from the very basic rule-based model and finishing with the fuzzy model. Fuzzy model is described in typical three steps. Finally, the three proposed sub-models of the general model are presented. Each presented model

represents the separate source of information and knowledge related to mortgage market. In the next step we would develop a mechanism for these sub-knowledge sources integration into one working collective architecture. Ideas related to collective intelligence, networks and conflict resolution would be investigated as part of this future research direction [9,10,11].

In the following step we will look for the proper data base and knowledge representation mechanism to store the experiences gathered during model execution. For ultimate verification purposes it is planned to use the 2003-2011 data of one of the biggest Polish Internet brokers (Bankier.pl) that would cover all three basic market situations: recession, moderate growth and economic boom presented in sub-models.

References

1. Orłowski, A.: Knowledge Management in the Internet Mortgage Market. Gdańsk University of Technology, Master Thesis (2008)
2. Zadeh, L.A., et al.: Fuzzy Sets, Fuzzy Logic, Fuzzy Systems. World Scientific Press, Singapore (1996)
3. Zadeh, L.A.: Fuzzy Sets and Information Granularity. In: Selected Papers by Zadeh, L.A., Klir, G.J., Yuan, B. (eds.) Advances in Fuzzy Systems-Applications and Theory. Fuzzy Sets, Fuzzy Logic and Fuzzy Systems, vol. 6, pp. 433–448, 16. World Scientific, Singapore (1996)
4. Zimmermann, H.-J.: Fuzzy Set Theory and Its Applications, 4th edn., p. 3. Kluwer Academic Publishers, Dordrecht (2001)
5. Orłowski, A., Szczerbicki, E.: Polish Internet Mortgage Market: Towars Fully Developed Fuzzy Model. Information Systems Architecture and Technology, 390 (2010)
6. Orłowski, A., Szczerbicki, E.: Conceptual fuzzy model of the polish internet mortgage market. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010. LNCS, vol. 6277, pp. 515–522. Springer, Heidelberg (2010)
7. Orłowski, A., Szczerbicki, E.: The process of fuzzy model development for the case of Polish Internet mortgage market. Journal of Internet Banking and Commerce, 55 (Array 2010)
8. Castells, M.: The Rise of the Network Society. Wiley-Blackwell (2009)
9. Hwang, D., Nguyen, N.T., et al.: A Semantic Wiki Framework for Reconciling Conflict Collaborations Based on Selecting Consensus Choice. Journal of Universal Computer Science 16(7), 1024–1035 (2010)
10. Duong, T.H., Nguyen, N.T., Jo, G.S.: Constructing and Mining: A Semantic-Based Academic Social Network. Journal of Intelligent & Fuzzy Systems 21(3), 197–207 (2010)
11. Nguyen, N.T.: Processing Inconsistency of Knowledge in Determining Knowledge of a Collective. Cybernetics and Systems 40(8), 670–688 (2009)

Knowledge Management Challenges in Collaborative Design of a Virtual Call Centre

Marcin Sikorski[1], Igor Garnik[1], Bohdan Ludwiszewski[1], and Jan Wyrwiński[2]

[1] Gdańsk Technical University, Faculty of Management and Economics,
ul. Narutowicza 11/12, 80-233 Gdańsk, Poland
Marcin.Sikorski@zie.pg.gda.pl
[2] FreecoNet Tlenofon S.A. GPNT, ul. Trzy Lipy 3, 80-172 Gdańsk, Poland
J.Wyrwinski@freeconet.pl

Abstract. This paper presents an analysis of knowledge management issues for a user interface consulting project relevant during the development of a Virtual Call Centre. The experiences gathered by a team of designers and a team of usability consultants have been described and evaluated from the knowledge management viewpoint. A concept of a knowledge-based system, potentially supporting usability consulting in subsequent IT projects has been presented, as well as constraints for its possible use in real settings.

Keywords: knowledge management in IT projects, usability consulting, user interface design.

1 Collaborative Design and Knowledge Management

Developing a successful IT (Information Technology) product requires from its developers a combination of skills and knowledge. This combination is addressed to user characteristics, planned context of use and the way the product is intended to match users' tasks and preferences. Markets of IT products are dynamic in terms of growth and innovation. Therefore IT products available on a competitive market have a high level of embedded knowledge content. The development of IT products can be thus considered as a knowledge-intensive activity; to be effective, IT projects require efficient mechanisms of knowledge transfer among team members as well as between design team and the external world [9].

While an IT company gains experience from successful products, much of the lesson remains captured as valuable pieces of information, converted gradually into the knowledge owned by specific teams within an organization. It is essential to acquire the knowledge about an IT product, its know-how and the development for reuse in next projects, preferably using consciously driven and fully controlled knowledge management processes at the organizational level.

Knowledge management (KM) issues in collaborative design have been observed in many IT projects mainly from the perspective of how difficulties in KM affect project efficiency, quality of the final product, and quality of teamwork [2, 3, 5, 9].

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 657–666, 2011.

Because expertise might be distributed within and across IT product development teams, KM mechanisms should be available for supporting collective understanding of software purpose and its context of use. It should also bring together the existing expertise in order to achieve better quality of a specific IT product. It is especially important when improving usability characteristics of an IT product, because it requires highly specialized knowledge from such areas as human factors, cognitive ergonomics and user-system interaction design. In such cases external consultants are often called up to support a design team and bring their knowledge and experience together with the competence of a specific IT design team [5, 7].

This paper is an attempt to present KM-related experiences gathered by two collaborating design teams working on user interface design in a specific software project. KM-related aspects of this project have been analyzed from the viewpoints of project efficiency and quality of delivered design solutions, including also some flaws which were observed during collaborative work. Finally, authors' considerations have been presented about possible use of a dedicated Knowledge-Based System for supporting next projects planned in the future in the same expertise domain.

2 Method Presentation

2.1 Typical Scope Usability Consulting in IT Projects

In IT projects external consulting has been a frequent practice in all cases when specific domain expertise must be acquired from outside the design team. Typical usability consulting activities in IT projects usually include [1, 4, 7]:

- capturing the context of use and user requirements,
- expert evaluation of prototype user interfaces,
- usability testing of prototype systems with real users,
- gathering users' evaluation data with surveys, interviews and questionnaires,
- analysis of evaluation data and recommending prospective design solutions.

Usability consulting in IT projects is an iterative process, based on communication among designers, consultants and other stakeholders. Collaborative observation of users and their behavior when using an IT product, facilitated by the explanation and interpretation of usability consultants, leads to a better understanding of user requirements by designers - and in principle should lead to creating a better product.

2.2 Virtual Call Centre - Project Description

The product under development was a **Virtual Call Centre (VCC) real time management panel**, developed by an innovation-oriented IT company from Gdansk (Poland). This product was aimed:

- to be used as a mobile, visual, digital, real-time switchboard for phone calls in small and medium size companies on PC's and mobile devices with a touchscreen,
- to serve as a mobile, visual, digital, real-time Virtual Call Centre management panel for small and medium size call centres on PCs and mobile devices with a touchscreen for the supervisor to see the detailed real-time and aggregated statistics of agents, queues, numbers, etc.
- to be used also out of office (for instance at convention centers, trade fairs, exhibitions where sometimes the entire company staff must be allocated for some time or in disaster recovery situations with the staff moved to a new location).

The design team was composed of four IT developers, with various backgrounds, from software engineering to graphic design and information architecture. Two team members were nominated as permanent contact persons to the usability consulting team. They were expected to take part in all activities involving participation of real users (usability testing, interviewing, evaluations etc.). Some parts of the VCC system were developed by external subcontractors (experts in Call Center management issues and statistics). They were also considered as design team members and took part in the usability improvement process.

The consulting team was composed of three usability specialists and HCI researchers from the Gdansk University of Technology, who had previous experience in conducting user-based laboratory usability testing.

The general plan of usability consulting for the VCC project covered:

- consulting conceptual design of the user interface,
- consulting prototype design and interaction solutions,
- developing usability test scenarios,
- testing interactive prototypes in usability tests,
- analysis and interpretation of test data,
- presenting and discussing results with designers,
- supervision upon implementing design decisions.

2.3 Forms of Collaborative Design and Consulting

In the VCC project the following forms of collaboration between designers and consultants were conducted (Fig. 1):

- at the design stage: real time and on-line meetings, user interface design workshops, interactive presentations, reflection, interpretation and discussion, negotiating design decisions,
- at the in evaluation stage: running usability tests with real users as well as discussing and interpreting evaluation results.

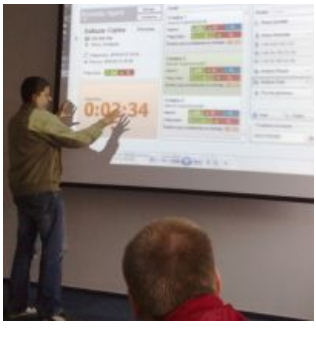


Fig. 1. A scene from a user interface design workshop for the VCC system

Fig. 2 presents a simplified flowchart of usability consulting activities performed for the VCC project. There are two major streams of workflow taking place separately in the design team and in the consulting team, and one placed in a collaborative design space, integrating contributions from both teams across subsequent phases of the project[1].

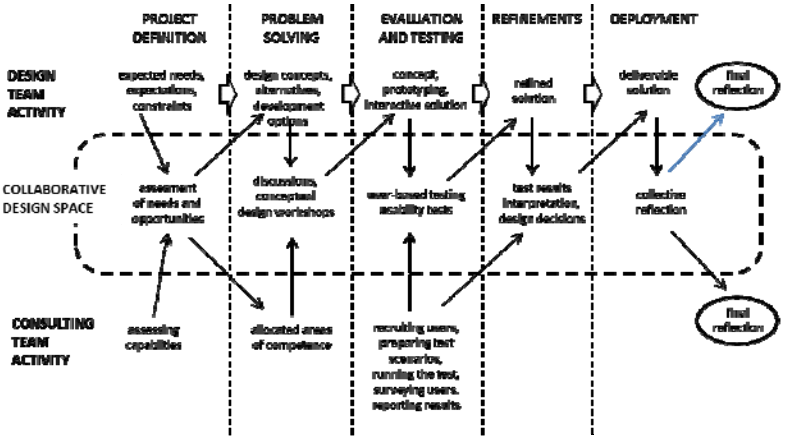


Fig. 2. Collaborative design activities in usability consulting for the VCC project

Major achievements of the VCC project can be classified into three categories:

- application of VCC software, tested with real users and ready for distribution,
- increased potential of each team;
- "final reflection" (shown in Fig. 2), as expanded knowledge within both teams, referring to a better understanding of the nature of user-system interaction in the VCC system; these separate reflections can merge into shared knowledge if two teams decide to undertake a similar project in near future.

2.4 Observations on Collaborative Design and Knowledge Management

During the implementation of this project some interesting observations were made. They are presented below and show relevance to KM practice and to KM-related IT support in usability consulting:

2.4.1 Constructive Reflection

Observing users when they evaluate prototype solutions, and interviewing them afterwards was an interesting experience for designers. This often led to immediate constructive reflection upon the ways in which to improve the product. Further

[1] In order to improve visual clarity, activities presented in Fig. 2 do not include subsequent iterations "concept-design-evaluation-refinement" which took twice for each version of the product. These iterations are crucial in designing and user-based testing of any kind of user interfaces.

discussions and collaborative activity have produced good results as to the product itself and to the quality of teamwork.

2.4.2 Changed Perception of VCC Product in Context

While discussing the design concepts and observing real users using the prototype solutions, as a result of the domain knowledge transfer, both designers and consultants gradually changed their perception of the product in its context:

- designers got closer to theory and users' viewpoint, consultants gained more practice and a better understanding of the project's constraints;
- working in teams supported knowledge transfer from own team to shared team knowledge, but presumably not up to the organizational level.

2.4.3 Evolving Consultants' Roles

As the project progressed, and the information exchange flow increased, a gradual change in consultants' roles could be observed:

- from a "real consultant" to "online consultant", especially while small details had to be frequently consulted online in periods between real-time meetings,
- from "usability inspector" focused on error correction towards a "usability guide", directing the design towards an innovative look of the product, and possibly adding new functionalities suitable for the planned context of use;
- from a "usability consultant" to "outsourcing provider" - inhibiting knowledge absorption and planning its use for consulting services in next projects.

2.4.4 IT Support for Knowledge Creation and Transfer (and the Lack of thereof)

In this project no dedicated IT tools were used to support knowledge management processes, or at least to preserve newly generated knowledge[2]. The main reason for that was the fact that the project was initially considered as a small-scale consultancy. Thus the use of any KM product (or at least dedicated group work software) at that moment apparently seemed to be excessive.

In this project, during usability consulting we observed the usual processes such as socialization, externalization, combination and internalization, regarding transfer of knowledge, knowledge conversion and new knowledge creation [8].

All processes took place among the individuals (inside the teams and across the teams) and between both teams during real-time events like: project meetings, lively discussions, exchange of views, negotiating a design compromise and achieving consensus. Discussions and critique on design concepts and on interactive prototypes, observing users while testing and evaluating interactive prototypes and - finally - discussing evaluation results and planning refinement scope were also performed.

Some of knowledge creation processes were IT-supported in a rather narrow scope - mainly by a reflective analysis of the content available online or delivered by traditional means of electronic communication:

- synchronous: phone calls, internet communicators,
- asynchronous: e-mails with numerous attachments and file transfers.

[2] Except for usual documentation files saved as the project archive.

However, all participants of this project found as most valuable those meetings and workshops which focused on real-time discussions with interactive slideshow presentations to be distributed after the meeting. And - most importantly - the presentations rather than lengthy textual reports were more likely to be read by managers.

As a result of lacking KM mechanism and tools, the knowledge created in both teams of this project had not been moved up to the organizational level[3], and there is a high risk that it will be lost if not used again (and preserved) in next projects.

3 Analysis, Evaluation, Outcomes and Discussion

3.1 An Overview of the Project Outcomes

Table 1 presents a balance of outcomes resulting from after-project evaluations and extracted from "final reflections" of both teams, as shown in Fig. 2.

Table 1. Overview of outcomes from the VCC project

Outcomes	*Designers*	*Consultants*
Gains and benefits	technical: more usable, more innovative, more competitive product organizational: new forms of collaboration, expanding current scope of design teams, changed design habits, acquiring new skills, standardization, patterns and procedures, increased maturity as to customer-orientedness economical: increased efficiency image: user-centered design process as a competitive asset and an element of quality management system	technical: expansion of current usability workbench, new technologies of developing user interfaces organizational: new experience with an innovation-minded client financial: income from a successfully completed service human: expanded skills and new components of professional experience image: possible good/bad references for next usability consulting projects
Efforts and investments	technical: new technologies organizational: external consulting slows down a design pace, necessity to deal with real users and take up their evaluation comments human: necessity to acquire new skills and expand current design perspective financial: costs of external consulting and additional internal activities	technical: customization of usability lab to VCC product testing organizational: time and workload related to consulting as well as to recruiting users for the usability test, underestimation of time and workload for specific fragments of VCC project human: necessity to acquire new skills and new design perspective

3.2 KM-Related Challenges

Despite the evident outcomes in the VCC project, when reviewing the overall results, special attention should be drawn to the difficulties observed during its implementation.

[3] Except of the project repository kept by the designers' company.

There were some trivial organizational challenges, typical of many projects, like:

- difficulties in personal adaptation to collaborative design, which affected individual involvement and contribution to collaborative work,
- difficulties in achieving required technical, usability and other characteristics,
- difficulties in managing teamwork and keeping up with project constraints.

However, a much more interesting part was relevant to KM-related challenges, characterized by the main problems discussed below:

3.2.1 Reusability of KM-Related Outcomes

A number of outcomes, solutions and products were developed:

- solutions created: user interface widgets, patterns, templates, guidelines and personal skills in using them - reusable for next projects;
- designers' skills in involving users into the IT product development - knowledge transferred during usability consultancy, with possible use in next projects when usability consultants are no longer needed;
- the current challenge is how to assure that these distributed resources can be used again and integrated for the benefits of next projects.

3.2.2 Sustainability of Change

Collaborative work in this project resulted in KM-relevant changes in human behavior, habits and attitudes:

- improvement and evolution of decision processes: design decisions related to user interface, initially made entirely within a design team, while the project was being consulted, leading to collective ownership of accepted solutions;
- changing perspective of both designers and consultants, as to the evolution of people and changes in their behavior as to design perspective (user-centeredness), teamwork culture and attitude to collaborative work;
- the current challenge is how to preserve cultural changes in design teams and to provide sustainable impact for the benefits of next projects.

3.2.3 Preserving the Knowledge for Reuse and Distribution

The difficulties occurring during this project stimulated a problem-solving approach:

- difficulties which were overcome are "lessons learned", extremely valuable for the team's future performance;
- innovative solutions worked out in this project and increased the intellectual capital of the company and its competitiveness;
- all experiences from past project form a part of organizational culture, its history, tradition and contribute to the increased design process maturity;
- the current problem is how to preserve all types of the newly gained knowledge and how to transfer it to other projects for their benefits.

3.2.4 Possible IT Support for KM - Needs, Vision and Conceptual Model

KM-related challenges experienced in this project suggest that IT support could be valuable for next projects. Such a support should be considered in two genres:

- a CSCW system[4] - a groupware environment facilitating collaborative design and project management by providing typical functionality, like:
  - individual and group workspaces and repositories for design artifacts,
  - assigning multiple roles to team members,
  - version management and change management,
  - various communication channels,
  - whiteboards for collaborative editing etc.;
- a Knowledge-Based System, which could provide the following functionality:
  - storing design primitives and formal knowledge in an online library,
  - preserving procedures and rules that proved successful in past design problems,
  - formal modeling of knowledge elements which might be applicable for usability improvements,
  - providing multiple mechanisms for knowledge acquisition, preserving, transfer and sharing.

Ideally, both systems should be integrated, used together and appropriately suited for the needs of usability-related projects, with prospect of KM solutions applicable for constant use in subsequent projects.

Fig. 3 presents a model of a future usability consultancy project supported by a hypothetical Knowledge-Based System, conceptualized upon experiences gathered from the VCC project. The VCC product depicted in Fig. 2 represents next versions of VCC and other similar products developed in the same design environment. The potential use of Knowledge-Based System, as shown in this model, could resolve some of the KM-related challenges observed in the VCC project. However, it will not be able to replace entirely human competence and skills as to usability engineering – even if they are allocated within the design team, or acquired from external sources, such as usability consulting [9].

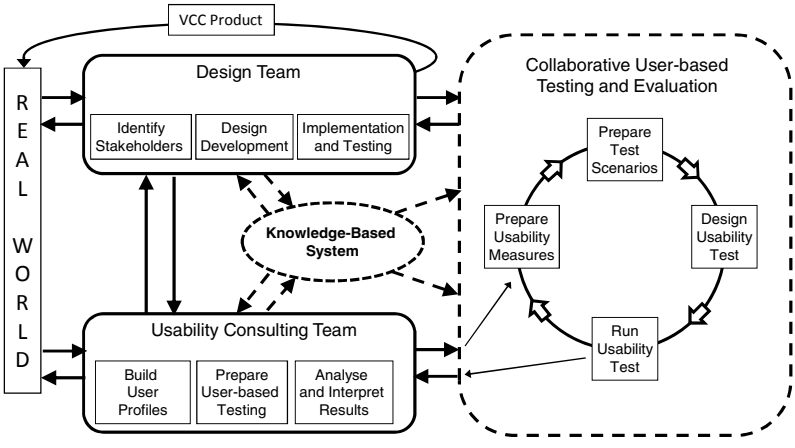


Fig. 3. Collaborative design model for usability consulting, supported by a Knowledge-Based System

[4] Computer Supported Cooperative Work

However, developing such a system - as in the case of many expert systems - may be particularly difficult, not only in an engineering dimension, but paradoxically from the usability point of view. The following factors can make the development of such Knowledge-Based System very problematic:

- in IT projects the professional area of user-relevant competence is sensitive, often exotic for IT developers, context-dependent and thus not always transferrable to other projects;
- usability toolbox is very flexible and full of interpretations, thus developing ontologies for this area may be difficult, because professional knowledge in many aspects is context-specific:
  - few formal models for human-computer interaction design exist,
  - few rule-based procedures and low formalization of user interface design,
  - acceptance of user interface solutions by users remains very subjective.

Literature search for examples illustrating how other Knowledge-Based Systems might help in suability consulting did not bring promising results, producing only little evidence [6] of expert systems supporting usability evaluation; moreover, leaving the conceptual, formative part of usability development within the domain of human experts. This example may serve as a possible explanation why human usability consultants are often needed up to now, and why Knowledge-Based Systems have not yet penetrated the area of usability consultancy in IT projects.

4 Summary and Conclusions

Knowledge management issues in IT projects are often intangible and sometimes even neglected. The description of the VCC project presented in this paper stresses the significance of KM support for transferring, converting and preserving design knowledge: artifacts, skills and experiences.

Potential use of a Knowledge-Based System for usability consultancy project was discussed upon the observations and experiences from recently completed an IT project. A conceptual model of usability consultancy project has been presented with the advantages and limitations of the proposed approach.

Acknowledgement. This project was supported by the Polish National Science Centre under the contract No. 4591/B/H03/2011/40.

References

1. Anderson, J., Fleek, F., Garrity, K., Drake, F.: Integrating Usability Techniques into Software Development. IEEE Software 18(1), 46–53 (2001)
2. Bektas, E., Heintz, J.L., Wamelink, J.W.F.: A Review Of Knowledge Management In Collaborative Design. In: 5th International Conference on Innovation in Architecture, Engineering and Construction, Antalya-Turkey (2008)

3. Ciampi, F.: Management Consulting and Knowledge Creation. SYMPHONYA - Emerging Issues in Management (1) (2007), http://www.unimib.it/upload/gestioneFiles/Symphonya/f2007issue1/ciampieng12007.pdf
4. Henneman, R.L.: Marketing Usability. In: Bias, J., Mayhew, D. (eds.) Cost-Justyfying Usability, pp. 144–163. Elsevier, Amsterdam (2005)
5. Hughes, M.: A Pattern Language Approach to Usability Knowledge Management. Journal of Usability Studies 1(2), 76–90 (2006)
6. Gabriel, I.J.: An Expert System for Usability Evaluations of Business-to-Consumer E-Commerce Sites. In: Proceedings of the 6th Annual ISOnEworld Conference, Las Vegas, NV, April 11-13 (2007)
7. Nieters, J.E., Ivaturi, S., Dworman, G.: The Internal Consultancy Model for Strategic UXD Relevance. In: CHI 2007, San Jose, CA, April 28-May 3, pp. 1825–1832 (2007)
8. Nonaka, I., Takeuchi, H.: The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, New York (1995)
9. Tiwana, A., Ramesh, B.: A Design Knowledge Management System to Support Collaborative Information Product Evolution. Decision Support Systems 31, 241–262 (2001)

Decisional DNA Applied to Digital TV

Haoxi Zhang[1], Cesar Sanin[1], and Edward Szczerbicki[2]

[1] Faculty of Engineering and Built Environment, School of Engineering,
The University of Newcastle, Callaghan, NSW, Australia 2308
[2] Gdansk University of Technology, Gdansk, Poland
haoxi.zhang@uon.edu.au,
Cesar.Sanin@newcastle.edu.au, Edward.Szczerbicki@zie.pg.gda.pl

Abstract. As the booming of digital TV, viewer's TV watch experience could be extremely valuable. Thanks to the fast developing IT techniques and solutions in the digital TV field, now we can run customized applications either inside the digital TV or at the viewer's set-top boxes to do what we want. In this paper, we introduce a new approach called Decisional DNA Digital TV that enables the digital TV to capture, reuse, and share the viewer's TV watch experience and preference; and we present the features, architecture and initial experimental results of our work. Decisional DNA is a domain-independent, flexible, smart knowledge representation structure which allows its domains to acquire, reuse, evolve and share knowledge in an easy and standard way.

Keywords: Decisional DNA, Set of Experience Knowledge Structure, Digital TV, XML.

1 Introduction

As the progress of digitalization and computerization of our daily life, the digital TV is becoming intelligent and interactive, or even becoming a computer. Many companies and organizations have involved into the implementation of such intelligent and interactive TV, like Sun Microsystems, the Digital Video Broadcasting Project (DVB), Google, and Apple. Also, a few solutions have been offered by these companies, such as Java TV [15], Google TV [11], and Multimedia Home Platform [17].

Developers can add interactive functions into Digital TV sets by using these existing solutions. In this paper, we introduce a domain-independent and standard approach, called the Decisional DNA Digital TV (DDNA DTV) to capture, reuse, and share viewers' TV watch experiences. It is based on the Java TV platform, and uses a novel knowledge representation structure – Decisional DNA [4].

This paper is organized as follows: section two describes an academic background on basic concepts related to our work; section three presents the features, architecture and experiments for the DDNA DTV Systems. Finally, in section four, concluding remarks are drawn.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 667–676, 2011.

2 Background

2.1 Digital TV

Digital television (DTV) is the television broadcasting system that uses the digital signals to transmit program contents. DTV not only delivers distortion-free audio and video signals; more importantly, it offers much higher radio spectrum efficiency than analog television does. DTV can also seamlessly integrate with other digital media, computer networks, and communication systems, enabling multimedia interactive services and data transmission [21].

A. Formats and Bandwidth

Digital television supports a range of different picture formats defined by the combination of interlacing, size, and aspect ratio (width to height ratio). With digital terrestrial television broadcasting in the world, the range of formats can be broadly divided into two categories: SDTV and HDTV. These terms by themselves are not very precise, and many subtle intermediate cases exist [18].

Standard definition TV (SDTV), by comparison, may use one of several different formats taking the form of various aspect ratios depending on the technology used in the country of broadcast. For 4:3 aspect-ratio broadcasts, the 640×480 format is used in NTSC countries, while 720×576 is used in PAL countries. For 16:9 broadcasts, the 704×480 format is used in NTSC countries, while 720×576 is used in PAL countries. However, broadcasters may choose to reduce these resolutions to save bandwidth (e.g., many DVB-T channels in the United Kingdom use a horizontal resolution of 544 or 704 pixels per line) [13].

High-definition television (HDTV), one of several different formats that can be transmitted over DTV, uses different formats, amongst which: 1280×720 pixels in progressive scan mode (abbreviated 720p) or 1920×1080 pixels in interlace mode (1080i). Each of these utilizes a 16:9 aspect ratio. (Some televisions are capable of receiving an HD resolution of 1920×1080 at a 60 Hz progressive scan frame rate — known as 1080p.) HDTV cannot be transmitted over current analog channels.

B. Standards

Currently, there are three main DTV standard groups [21]:

1) The Digital Video Broadcasting Project (DVB), a European based standards organization, which developed the DVB series of DTV standards, standardized by the European Telecommunication Standard Institute (ETSI) [9].

2) The Advanced Television Systems Committee (ATSC), a North America based DTV standards organization, which developed the ATSC terrestrial DTV series of standards. In addition, the North American digital cable TV standards now in use were developed separately, based on work done by Cable Television Laboratories (Cable Labs) and largely codified by the Society of Cable Telecommunications Engineers (SCTE) [2].

3) The Integrated Services Digital Broadcasting standards (ISDB), a series of DTV standards developed and standardized by the Association of Radio Industries and

Business (ARIB) and by the Japan Cable Television Engineering Association (JCTEA) [1].

C. Reception

There are various ways to receive digital television. One of the oldest means of receiving DTV (and TV in general) is using an antenna. This way is known as Digital Terrestrial Television (DTT) [19]. With DTT, viewers are limited to whatever channels the antenna picks up. Signal quality will also vary.

Other ways have been devised to receive digital television. Among the most familiar to people are digital cable and digital satellite. In some countries where transmissions of TV signals are normally achieved by microwaves, digital Multichannel Multipoint Distribution Service (MMDS)[12] is used. Other standards, such as Digital Multimedia Broadcasting (DMB) [20] and Digital Video Broadcasting - Handheld (DVB-H) [16], have been devised to allow handheld devices such as mobile phones to receive TV signals. Another way is Internet Protocol TV (IPTV) [3], which is receiving TV via Internet Protocol, relying on Digital Subscriber Line (DSL) or optical cable line.

2.2 Interactive Television

Interactive television (generally known as iTV) describes a number of techniques that allow viewers to interact with television content and services; it is an evolutionary integration of the Internet and DTV [10].

The most exciting thing of an interactive TV is the ability to run applications that have been downloaded as part of the broadcast stream: this is really what makes the difference between a basic digital TV box and an interactive TV system. In order to support and enable interactive applications, the receiver is required to support not only the implementation of APIs needed to run the applications, but also the infrastructure needed to inform the receiver what applications are available and how to run them.

Interactive TV has drawn attention from researchers, organizations, and companies, and there have been a few efforts and solutions offered by them. Java TV and Multimedia Home Platform are the two most popular and vibrant techniques in this field [15] [17].

A. Java TV

The Java TV is a Java-based software framework designed for supporting digital TV platforms from Sun Microsystems. It brings together a number of the common elements that are needed in a digital TV platform. These include the core application model and lifecycle, access to broadcast services (either via Java TV itself or via the Java Media Framework) and access to service information [15].

Most importantly, Java TV is not bound to a specific set of standards for digital TV. Java TV is explicit, pure, and independent. Because of this, it works equally well with many solutions for digital TV, such as ATSC solutions, or OpenCable solutions, or DVB-based systems. It gives Java TV a very strong advantage that applications

written to use Java TV APIs will work on any platform that supports it, rather than being tied to a specific broadcast system [15].

B. Multimedia Home Platform

Multimedia Home Platform (MHP) is an open standard middleware system designed by the DVB Project for enhanced and interactive digital television [17].

The MHP enables the reception and execution of interactive, Java-based applications on a TV-set. Interactive TV applications can be delivered over the broadcast channel, together with video and audio streams. These applications can be, for instance, games, e-mail, information services, interactive voting, shopping or SMS.

MHP defines a generic interface between interactive digital applications and the terminals, which those applications execute on. This interface decouples different applications of a provider from specific hardware and software details of different MHP terminal implementations. It enables digital content providers to address all types of terminals ranging from low-end to high-end set top boxes, integrated digital TV sets and multimedia PCs. The MHP extends the existing DVB open standards for broadcast and interactive services in various broadcasting networks, like satellite, cable or terrestrial networks.

2.3 Set of Experience Knowledge Structure (SOEKS) and Decisional DNA

The Set of Experience Knowledge Structure (SOEKS or shortly SOE) is a domain-independent, flexible and standard knowledge representation structure [14]. It has been developed to acquire and store formal decision events in an explicit way [4]. It is a model based upon available and existing knowledge, which must adapt to the decision event it is built from (i.e. it is a dynamic structure that depends on the information provided by a formal decision event) [8]; besides, it can be represented in XML or OWL as an ontology in order to make it transportable and shareable [5] [6].

SOEKS is composed of variables, functions, constraints and rules associated in a DNA shape permitting the integration of the Decisional DNA of an organization [8]. Variables normally implicate representing knowledge using an attribute-value language (i.e. by a vector of variables and values) [7], and they are the centre root of the structure and the starting point for the SOEKS. Functions represent relationships between a set of input variables and a dependent variable; moreover, functions can be applied for reasoning optimal states. Constraints are another way of associations among the variables. They are restrictions of the feasible solutions, limitations of possibilities in a decision event, and factors that restrict the performance of a system. Finally, rules are relationships between a consequence and a condition linked by the statements IF-THEN-ELSE. They are conditional relationships that control the universe of variables [8].

Additionally, SOEKS is designed similarly to DNA at some important features. First, the combination of the four components of the SOE gives uniqueness, just as the combination of four nucleotides of DNA does. Secondly, the elements of SOEKS are connected with each other in order to imitate a gene, and each SOE can be classified, and acts like a gene in DNA [8]. As the gene produces phenotypes, the

SOE brings values of decisions according to the combined elements. Then a decisional chromosome storing decisional "strategies" for a category is formed by a group of SOE of the same category. Finally, a diverse group of SOE chromosomes comprise what is called the Decisional DNA [4].

In short, as a domain-independent, flexible and standard knowledge representation structure, SOEKS and Decisional DNA provide an ideal approach which can not only be very easily applied to various embedded systems (domain-independent), but also enable standard knowledge communication and sharing among these embedded systems.

3 The Decisional DNA Digital TV

Nowadays, digital TV has being rolling on with full force. Thanks to its capability of transmitting digital data along with the audiovisual contents, the TV program providers can interact with their viewer by offering customized applications, which run at their viewers' set-top boxes or inside the TV. In order to capture, reuse, and share viewers' TV watching experiences, we applied the novel knowledge representation structure – Decisional DNA, to digital TV, called The Decisional DNA Digital TV.

3.1 System Architecture

The DDNA DTV consists of the User Interface, the System I/O, the Integrator, the Prognoser, the XML Parser and the Decisional DNA Repository (see Figure 1).

- User Interface: The User Interface is developed to interact with the user/viewer. In particular, user can control, set and configure the system by using the user interface. Like the user can use remote control to select services, give feedback to a movie and interact with the service provider through the User Interface.
- System I/O: The System I/O allows our Decisional DNA approach to communicate with its domain. The System I/O tells the DTV which service is selected, what movie should play, what feedback was given. Also, it reads the media stream, feedback, system time, service information from its domain.
- Integrator: In our case, we link each experience with a certain scenario. The Integrator is the place where the scenario data is gathered and organized, such as the system time, the name of a selected service, user input and other service information, describing the circumstance under which experience is acquired. Therefore, the scenario data is transformed into a set of experience. The Integrator organizes the scenario data into strings using ID－VALUE format, and send them to the Prognoser for further processing.
- Prognoser: The Prognoser is in charge of sorting, analyzing, organizing, creating and retrieving experience. It sorts data received from the Integrator, and then, it analyzes and organizes the data according to the system configuration. Finally, it interacts with the Decisional DNA Repository and the XML Parser in order to store and reuse experience depending on the purpose of different tasks.

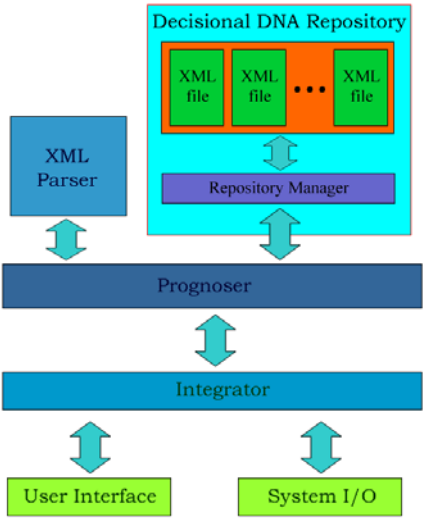


Fig. 1. System Architecture for DDNA DTV

- XML Parser: The XML Parser is the converter that translates knowledge statements generated by the Prognoser into the Decisional DNA experience structure represented in XML format; and interprets the retrieved XML-represented Decisional DNA experience for reusing.
- Decisional DNA Repository: The Decisional DNA Repository is the place where experiences are stored and managed. It uses standard XML to represent experiential knowledge, which makes standard knowledge sharing and communication become easier. It is composed of the Repository Manager and XML files:

a) Repository Manager. The Repository Manager is the interface of the Decisional DNA Repository. It answers operation commands sent by the Prognoser and manages the XML files. There are two main tasks in it: searching experiences and managing XML files.
b) XML files. We use a set of XML tags described in [14] to store Decisional DNA in XML files. In this way, Decisional DNA is explicitly represented, and ready to be shared among different systems.

3.2 Simulation and Experiments

We used the Java TV SDK with NetBeans 6.8 on a DELL Latitude ES400 laptop to test the idea of the DDNA DTV. At this stage, the main purpose of our experiments is to prove that the Decisional DNA can work with Java TV, and our approach can provide its domain with the ability of experience capturing and reusing. Thus, we assume that there are only five types of movies, namely action, adventure, animation, comedy, and crime. And each movie is represented by its type plus an ID number, like Action1, Comedy2; there are 20 movies for each type.

We capture viewer's watching experience by recording seven variables: Movie Name, Director, Watch Date, Watch Time, Ranking, Type, and Viewer. Movie Name

and Director are used to indicate which movie the viewer watched. Watch Day and Watch Time store date and time when this movie is watched. Ranking shows how the viewer likes this movie. Type illustrates what kind of movie it is. Viewer saves the name of user. Those variables are gathered and organized by the Integrator and then send to the Prognoser; finally, they are stored as a SOEKS in XML format [14] (See Figure 2). Once the system have more than ten SOEKS (this number can be set in system), it begins to analyze viewer's watching preference, and gives the viewer better recommendations according to analyze settings.

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<!-- Set of Experience Knowledge Structure -->
- <set_of_experience xmlns:xsi="http://www.w3.org/2001/XMLSchema-instan
    <date>2011-05-11</date>
    <hour>11:05:53</hour>
  - <category>
      <!-- Category encloses this SOE into a determined chromosome of
      <area>Entertainment</area>
      <subarea>Movie Watching Experience</subarea>
      <subject>Movie</subject>
    </category>
  - <set_of_variables>
      <!-- Variables included in the model -->
    - <variable>
        <var_name>Name</var_name>
        <var_type>CATEGORICAL</var_type>
        <var_cvalue>ADVENTURE 7</var_cvalue>
        <var_evalue>ADVENTURE 7</var_evalue>
        <unit />
        <internal>true</internal>
        <weight>0.0</weight>
        <l_range>0.0</l_range>
        <u_range>0.0</u_range>
      - <categories>
          <category>ADVENTURE 7</category>
        </categories>
        <priority>0.2</priority>
      </variable>
    - <variable>
        <var_name>Director</var_name>
        <var_type>CATEGORICAL</var_type>
        <var_cvalue>ADVENTURE 7</var_cvalue>
        <var_evalue>ADVENTURE 7</var_evalue>
        <unit />
        <internal>true</internal>
        <weight>0.0</weight>
        <l_range>0.0</l_range>
        <u_range>0.0</u_range>
      - <categories>
          <category>ADVENTURE 7</category>
        </categories>
        <priority>0.0</priority>
      </variable>
    - <variable>
        <var_name>Watch Day</var_name>
        <var_type>CATEGORICAL</var_type>
        <var_cvalue>WEDNESDAY</var_cvalue>
        <var_evalue>WEDNESDAY</var_evalue>
        <unit />
        <internal>true</internal>
        <weight>0.0</weight>
        <l_range>0.0</l_range>
        <u_range>0.0</u_range>
      - <categories>
          <category>MONDAY</category>
          <category>TUESDAY</category>
          <category>WEDNESDAY</category>
          <category>THURSDAY</category>
          <category>FRIDAY</category>
          <category>SATURDAY</category>
          <category>SUNDAY</category>
        </categories>
        <priority>0.0</priority>
      </variable>
    - <variable>
        <var_name>Watch Time</var_name>
        <var_type>NUMERICAL</var_type>
        <var_cvalue>11:05:53</var_cvalue>
        <var_evalue>11:05:53</var_evalue>
        <unit />
        <internal>true</internal>
        <weight>0.0</weight>
        <l_range>0.0</l_range>
        <u_range>24.0</u_range>
        <priority>0.0</priority>
      </variable>
    - <variable>
        <var_name>Ranking</var_name>
        <var_type>NUMERICAL</var_type>
        <var_cvalue>7</var_cvalue>
        <var_evalue>7</var_evalue>
        <unit />
        <internal>true</internal>
        <weight>0.0</weight>
        <l_range>0.0</l_range>
        <u_range>10.0</u_range>
        <priority>0.4</priority>
      </variable>
    - <variable>
        <var_name>Type</var_name>
        <var_type>CATEGORICAL</var_type>
        <var_cvalue>ADVENTURE</var_cvalue>
        <var_evalue>ADVENTURE</var_evalue>
        <unit />
        <internal>true</internal>
        <weight>0.0</weight>
        <l_range>0.0</l_range>
        <u_range>0.0</u_range>
      - <categories>
          <category>ACTION</category>
          <category>ADVENTURE</category>
          <category>ANIMATION</category>
          <category>COMEDY</category>
          <category>CRIME</category>
        </categories>
        <priority>0.4</priority>
      </variable>
    - <variable>
        <var_name>Viewer</var_name>
        <var_type>CATEGORICAL</var_type>
        <var_cvalue>TOM</var_cvalue>
        <var_evalue>TOM</var_evalue>
        <unit />
        <internal>true</internal>
        <weight>0.0</weight>
        <l_range>0.0</l_range>
        <u_range>0.0</u_range>
      - <categories>
          <category>Tom</category>
        </categories>
        <priority>0.0</priority>
      </variable>
    </set_of_variables>
    <set_of_functions />
    <set_of_constraints />
    <set_of_rules />
</set_of_experience>
```

Fig. 2. A SOEKS of a Watched Movie

We simulated a viewer watching movies on the DDNA DTV. Figure 3 shows a screenshot of the user's TV. As we can see, the viewer's screen is composed by five components: Service Name which shows "Movies" here, Service Information which displays introduction of a selected movie, Ranking, Movie Showcase which shows ten movies recommended by the system, and "Show More…" button, by which viewer can access more movies. At the beginning, DDNA DTV recommends two movies from each movie type. Once the system gets enough experiences, it will recommend movies according to those experiences.

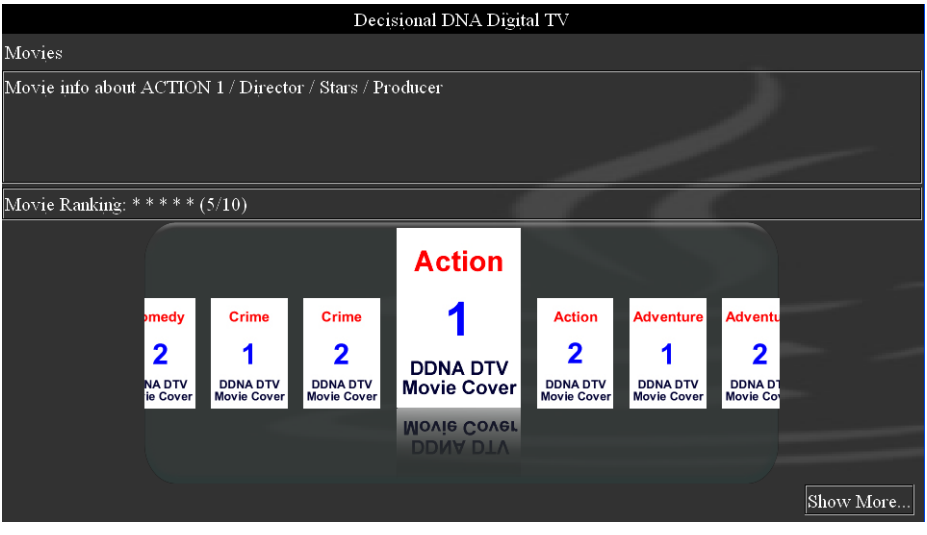


Fig. 2. Screenshot of DDNA DTV

For example, we assume that there is a viewer, Tom, who likes to watch action movies on every Saturday night as shown in the Table 1.

Table 1. Tom's Movie Watching Records

Movie Name	Watch Date	Watch Time	Ranking	Type
Action1	8/01/2011	19:35	7	Action
Action2	22/01/2011	20:02	9	Action
Action3	29/01/2011	20:13	8.5	Action
Action4	12/02/2011	19:42	8.7	Action
Action5	19/02/2011	21:07	8.6	Action

When the Prognoser recommends new movies to the user, it retrieves those stored watching experiences from the Decisional DNA Repository, and analyzes those experiences according to the user's settings. In this experiment, we analyze the movie types the user watched, and what day in a week the user usually watches them. Formula 1 demonstrates how the system calculates the amount of places a movie type should take in the 10-movie recommendation list.

$$N_t = T / D \times 10 . \quad (1)$$

N_t represents how many movies should be recommended from a specific movie type. T represents how many movies of a specific movie type have been watched on a specific week day. D represents the total amount of watched movies on that week day. For example, Tom watched 10 movies in total on Saturdays, and 5 of those movies are action movies. Therefore, there should be five action movies (5 / 10 × 10 = 5) in the next recommendation list for Tom.

As we assumed, during a few weeks capturing experience, the system learns and knows that Tom watched 5 action movies, 1 adventure movie, 1 animation movie, 2 comedy movies, and 1 crime movie on Saturdays so far. As a result, the system will recommend 5 action movies, 1 adventure movie, 1 animation movie, 2 comedy movies, and 1 crime movie for him on next Saturday. Figure 4 shows a screenshot of a newly recommended movie list for Tom.

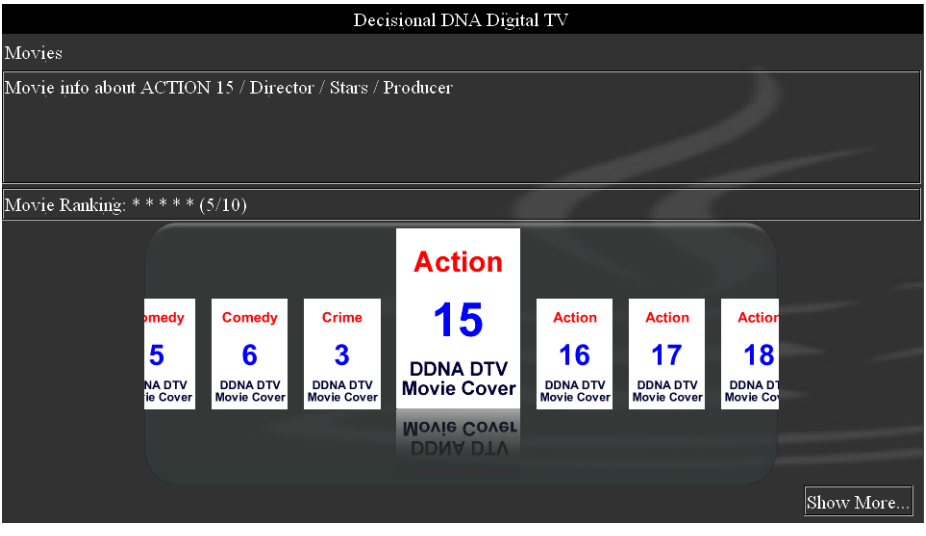


Fig. 4. Newly Recommended Movie List for Tom

4 Conclusion and Future Work

In this paper, we presented the concept, features, and architecture of the Decisional DNA Digital TV, and did some initial tests on a DELL laptop with Java TV SDK. As the result shows, the DDNA DTV can work under the Java TV environment, and it enables its domain to capture, store, and reuse viewers' TV watching experiences by using a novel experiential knowledge representation structure, Decisional DNA. As well as it can provide TV viewers with better user experience.

Since the DDNA DTV research is at its early stage, there is further research and refinement to be done, some of them are:

- Enhancement of the efficiency of Decisional DNA Repository storage and query.
- Further development of the user login system.

- Refinement and further development of algorithm using in the Prognoser.
- Implement better ways to interpret the user experience such as fuzzy logic.

Reference

1. ARIB, the Association of Radio Industries and Business, http://www.arib.or.jp/english/
2. ATSC, the Advanced Television Systems Committee, http://www.atsc.org/cms/
3. Yarali, A., Cherry, A.: Internet Protocol Television (IPTV). In: TENCON 2005 IEEE Region 10, pp. 1-6 (2005) ISBN: 0-7803-9311-2
4. Sanin, C., Szczerbicki, E.: Experience-based Knowledge Representation SOEKS. Cybernetics and Systems 40(2), 99–122 (2009)
5. Sanin, C., Szczerbicki, E.: Extending Set of Experience Knowledge Structure into a Transportable Language Extensible Markup Language. International Journal of Cybernetics and Systems 37(2-3), 97–117 (2006)
6. Sanin, C., Szczerbicki, E.: An OWL ontology of Set of Experience Knowledge Structure. Journal of Universal Computer Science 13(2), 209–223 (2007)
7. Lloyd, J.W.: Logic for Learning: Learning Comprehensible Theories from Structure Data. Springer, Berlin (2003)
8. Sanin, C., Mancilla-Amaya, L., Szczerbicki, E., CayfordHowell, P.: Application of a Multi-domain Knowledge Structure: The Decisional DNA. In: Intel. Sys. For Know. Management. SCI, vol. 252, pp. 65–86 (2009)
9. DVB - The Digital Video Broadcasting Project, http://www.dvb.org/
10. Schwalb, E.: iTV Handbook: Technologies & Standards. ACM Computers in Entertainment 2(2), article 7 (2004) ISBN 0131003127
11. Google - Google TV, http://www.google.com/tv/
12. IEEE, Multichannel Multipoint Distribution Service, http://grouper.ieee.org/groups/802/16/
13. Latest snapshots - Freeview/DTT bitrates, Latest snapshots, http://dtt.me.uk/
14. Maldonado Sanin, C.A.: Smart Knowledge Management System, PhD Thesis, Faculty of Engineering and Built Environment - School of Mechanical Engineering, University of Newcastle, E. Szczerbicki, Doctor of Philosophy Degree, Newcastle (2007)
15. TV Without Borders, Java TV Tutorial, http://www.interactivetvweb.org/tutorials/javatv/
16. Reimers, U.H.: DVB-The Family of International standards for Digital Video broadcasting. Proceedings of the IEEE 94(1), 18–9219 (2006) ISSN: 0018-9219
17. Vrba, V., Cvrk, L., Sykora, M.: Framework for digital TV applications. In: Proceedings of the International Conference on Networking. International Conference on Systems and International Conference on Mobile Communications and Learning, p. 184 (2006) ISBN: 0-7695-2552-0
18. Wikipedia, Digital Television, http://en.wikipedia.org/wiki/Digital_television/
19. Wikipedia, Digital Terrestrial Television, http://en.wikipedia.org/wiki/Digital_terrestrial_television
20. World DMB, Digital Multimedia Broadcasting, http://www.worlddab.org/
21. Wu, Y., Hirakawa, S., Reimers, U., Whitaker, J.: Overview of digital television development worldwide. Proc. IEEE 94(1), 8–21 (2006)

Measurement of the Development of a Learning IT Organization Supported by a Model of Knowledge Acquisition and Processing

Cezary Orłowski and Tomasz Sitek

Gdańsk University of Technology, Department of Management and Economics,
ul. Narutowicza 11/12, Gdansk, Poland
{cezary.orlowski,tomasz.sitek}@zie.pg.gda.pl

Abstract. The paper presents a model of knowledge acquisition and processing for the development of learning organizations. The theory of a learning organization provides neither metrics nor tools to measure its development The authors' studies in this field are based on their experience gathered after projects realized in real IT organizations. The authors have described the construction of the model and the methods of its verification through a series of experiments. It was proven that the model can help an organization in the orderly and formalized acquisition and processing of knowledge about its internal processes, which leads to controlled and permanent development.

Keywords: knowledge base, uncertain knowledge, reasoning, expert system, information technology.

1 Introduction

There is no doubt that today, knowledge is the most valuable business asset. It is knowledge and information that constitute an important factor in competitiveness. The amount of information in the external and internal environment of each organization is growing. Such information noise can induce confusion and make taking rational and risk-free decisions impossible. This phenomenon is accompanied by the rapid obsolescence of knowledge.

It is a natural process in any organization that its members (i.e. employees) gain experience. Such experience can constitute knowledge if it is consciously collected, analyzed and processed. Therefore, managing the knowledge of an organization means managing the knowledge of all its members. Decisions made in organizations which are aware of the value of their knowledge are subject to far less uncertainty and hence the risk of failure is lower.

In the current era, which is known as the epoch of knowledge and information, a company will face new tasks. First of all, it should always have the latest information and current knowledge. Secondly, it should be able to use these resources in order to obtain a competitive advantage and ensure its survival. This leads to the conclusion that knowledge should be the foundation for the development of a modern organization, and all activities associated with it (acquisition, storage, processing) should be structured and formalized.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 677–686, 2011.

1.1 Learning Organizations

Organizations which operate on the basis of accumulated knowledge are called learning organizations. The concept of a learning organization is understood as a management concept based on the sum of the knowledge possessed by individual contributors - knowledge which is constantly enriched and "shared" with the company. An organization which works in this way should constantly explore new ways of thinking and new perspectives. New knowledge should be used in practice to solve emerging problems [1].

The main feature of a learning organization is the fact that it improves its operations through the greater knowledge of its workers and their deeper understanding of the principles and purposes of the functioning of the organization. A learning organization supports the learning process of all its members, whether individual, collective or organizational. At the same time, it constantly transforms itself. Another distinctive feature is the fact that the experiences gained serve to create new knowledge [2].

While an economy based on knowledge is today an undeniable fact, learning organizations have so far been just one of many theories of management [3]. In many businesses issues related to intellectual capital are still unheard of or undervalued. Those organizations which understand that competition is no longer based solely on the product and innovation, have a chance to gain an advantage and to develop. However, for the development of a learning organization to be implemented as planned, it must be monitored and constructive conclusions must be drawn from measurements taken.

2 A Suggestion for the Quantitative Measurement of the Development of Learning Organizations

Implementing the concept of a learning organization requires a lot of self-awareness and knowledge of psychology and sociology. Thus, it is based on "soft" disciplines. This determines its fundamental flaw - all the definitions of both the learning organization and its success factors focus exclusively on qualitative aspects.

Such factors as willingness to improve and willingness to build and shape a vision or a specific organizational culture have been indicated as determinants of learning. Theorists, however, do not create any foundations for a process of consciously and comprehensively monitoring the processes affecting the development of learning organizations. No quantitative measures have yet been created which could accurately describe such an organization. There are no indicators or even reference values which could provide a tool for auditing (internal and external) [4].

2.1 The Need to Quantify the Determinants of Organizational Learning

The currently non-existent methods of quantifying the factors which enable a company to be included in the group of learning organizations do not allow these organizations to set measurable targets for development. It often happens that inconsistent guidelines presented by professionals in this field allow the path of such

organizations to be described only in a fuzzy way. Not having a precise scale, it can only be concluded that *a lot* has already been done, or that certain objectives are *nearly* met. This leads to the conclusion that only the concept of "feeling like" a learning organization can be stated, rather than the fact that they exist, supported by figures.

Therefore, the inability to measure the progress in learning does not allow the diagnosis of the causes of the observed state either. Even if it is known that the intended objectives have not been achieved, it is not always possible to determine why. This situation does not allow the identification of the weaknesses or the opportunities which are not fully used. The inability to employ its full potential leads to an uneven development of such an organization (there may be considerable disparity between the maturity of processes in its different areas). As a result, the learning process is slowed down.

Moreover, the inability of quantitative description in learning organizations prevents a prediction of their future status and behavior. This is not only a simple projection of the shape of the organization, resulting from the extrapolation of current trends related to learning. A far bigger drawback is the lack of tools to determine the precise objectives in this field, both at a strategic and a tactical-operational level. Thus, the absence of such tools hinders a review of potential options for development. It is difficult to make an optimistic and a pessimistic model, and to prepare alternative plans of action on such a basis.

These arguments allow a thesis to be formulated: learning organizations need quantitative measures. There is a need for a coherent and complete standard of evaluation with a set of reference values. With their help, a learning organization would have the opportunity to develop in a conscious and fully controlled way. Having identified a need, the authors propose a model, based on the concept of intelligent systems, for the acquisition and processing of knowledge [5].

3 A Model for the Acquisition and Processing of Knowledge in Managing a Learning Organization

The proposed model of managing knowledge on the development of a learning organization is based on the concept of decision support systems (expert systems). Expert systems are used in many fields of human activity. Most reported implementations carry out decision support in the realities of technical systems.

Technical systems are characterized by the high availability of data, which is processed into knowledge. This data usually comes from measurements or calculations, hence it is objective and unequivocal. The functioning of technical systems is clear and well defined. For the acquired and processed knowledge, data constitutes a complete base relying on rules and/or facts. In contrast to technical systems, social systems are not fully recognized and are therefore difficult to describe. Very often the available knowledge is imperfect. It may not be possible to obtain knowledge which is certain or sufficiently precise. It is rarely complete. Organization management is carried out in mixed conditions. Necessary decisions should take into account both the knowledge resulting from objective sources of information, as well as all the "soft" aspects which are hard to measure. Therefore, in accordance with such classification, an organization (a company) is a socio-technical system.

The authors present a model for decision support in the management of organizational knowledge which takes problems arising from its social nature into account. They identify these problems in two areas: the acquisition and the processing of knowledge. Because of this division, separate sub-models can be distinguished in the model of knowledge acquisition and processing (Fig. 1):

- The Sub-model of Knowledge Acquisition,
- The Sub-model of Knowledge Processing.

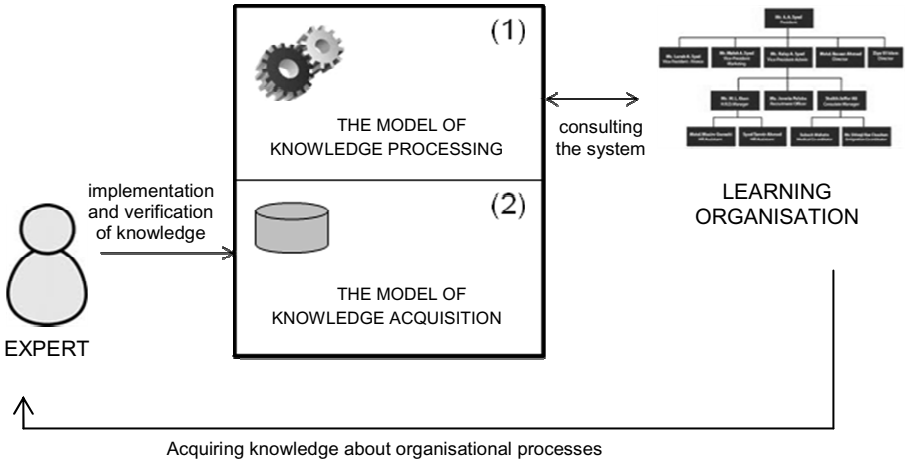


Fig. 1. The general model of knowledge acquisition and processing [5]

3.1 The Sub-model of Knowledge Acquisition

Knowledge in expert systems is usually represented by facts and rules. In technical systems, a complete model of knowledge is required in which the base of rules includes a description of all possible states arising from the number of combinations of input variables. Physically, there are usually two types of structure - the base of facts and the base of rules.

In social systems, such a division of knowledge proves to be too simple. The knowledge of rules may come from different sources, which leads to potential inconsistencies. It may transpire that it is partly uncertain (an expert can express their doubts in certain aspects). According to the criteria used in technical systems, such knowledge should be rejected. In management, however, it may prove to be a valuable reference for a policymaker, therefore, it should be retained.

Nevertheless, it is important to appropriately differentiate between knowledge of different "ranks". Assuming that both types of knowledge will be used, one in the desired form (certain, full), as well as the imperfect one, priorities must be established for both of them. The machine for drawing conclusions should always make use of the certain knowledge first. To determine such a chronology, it is necessary to develop an appropriate logical division of knowledge. It is suggested that several separate databases should be established, as shown in Figure 2.

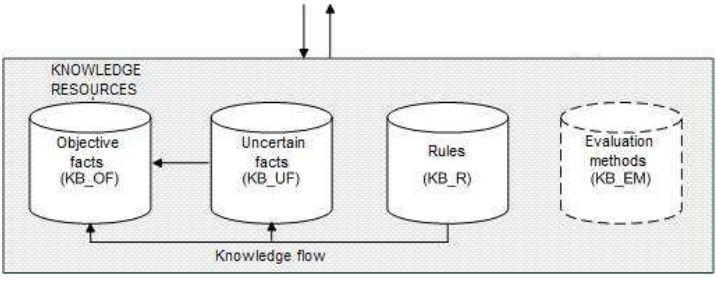


Fig. 2. The structure of knowledge bases in the suggested model [6]

In addition to the knowledge bases, due to its nature (facts, rules), the base of Evaluation Models should also be taken into consideration. It has been assumed that the evaluation model does not contain a typical knowledge base equipped with a set of rules or facts. It has also been assumed that it contains complete models of knowledge processing, within which inference can be carried out. This construction of the knowledge base contains subjective knowledge about evaluation methods (with compelling observations and measurements). Furthermore, it has been assumed that the knowledge accumulated here is a consequence of the different experiences of experts and their conclusions. The use of various evaluation models is described later in this article.

The knowledge gained in organizations is usually imperfect (incomplete, uncertain, vague). The question arises whether this situation is due to the nature of the organization or perhaps it results from the bad selection of experts? Perhaps knowledge on the same subject from another source would be without drawbacks. Therefore, the question about the quality of knowledge turns into a question about the confidence given to a developed model. Should the currently acquired knowledge reach the production bases directly and thus give rise to the implemented inference? It seems that the potential risk associated with bad decisions is too high. The model should establish appropriate mechanisms to verify the acquired knowledge. To minimize the risk of putting "junk" in the knowledge base, a mechanism, a so-called *knowledge buffer*, has been developed (Fig. 3). The initially gathered knowledge is stored in the buffer for verification.

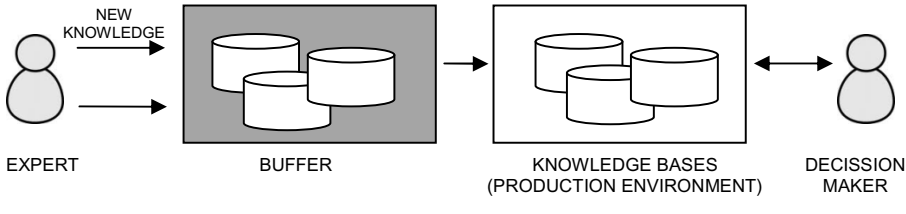


Fig. 3. The concept of knowledge-buffer operation [6]

3.2 The Sub-model of Knowledge Processing

The Sub-model of Knowledge Processing offers the specific design of an inference mechanism, in which varied inference mechanisms will be applied. It has also been assumed that the model of knowledge processing should be used for typical situations in the operation of a learning organization: evaluation and streamlining of operational processes, evaluation of the organization's maturity, etc. Experience shows that such situations are characterized by [7]:

- the uncertainty of knowledge,
- the lack of precision of the acquired knowledge,
- its incomplete probability and incompleteness.

For this reason, first of all, it is assumed that the expert should consciously identify (on an accepted point scale) their level of knowledge for each statement. Certainty factors should be introduced for such situations. It has been assumed, in the suggested model, that such an approach focuses on a base of facts. In the second case, it is assumed that the expert is able to describe the complete set of rules specifying the state of domain knowledge. However, it has been established that the expert has no option (and often need) to note them in a sharp (quantified) way. The model can only be described in a linguistic form (in which, by describing, the expert agrees to lesser precision). The use of inference mechanisms based on fuzzy logic has also been assumed. In the case of incomplete probability and the incompleteness of knowledge, it has been assumed that the knowledge has an incremental character. Initially, it is incomplete and thus the application process should consider the level of probability of the posed hypotheses, as the level increases over time as new data is added (when hypotheses are made more probable by facts).

Given the potential problems with gaining knowledge in the area of management, the construction of an inference mechanism based on a hybrid inference algorithm has also been adopted. In this case, the hybrid character is understood as using more than one of the aforementioned methods of inference for each application process.

4 The Verification of the Concept

For the purpose of the model verification process, two learning organizations were selected. The choice resulted from the belief that learning processes are the overriding processes determining the essence of their functioning.

The verification was carried out by implementing inference using various evaluation models. During a series of experiments, the functioning of the concept was examined, including:

- simple evaluation models based on fuzzy systems
- more advanced models of self-organizing,
- self-adjustable models.
- neural networks.

These experiments were designed to verify the model's behavior in different types of decision problems. The identification and elimination of existing problems led to modifications of the model which would allow the main goal of the research to be carried out in the best way, namely, to provide support for the quantitative assessment of the development of learning organizations.

The basic idea which was adopted was the identification of the concept of organizational learning with a controlled increase in its level of maturity. This results from the authors' experience of working together with IT companies. These companies were used as a target group for the suggested solution. This means, therefore, that IT organizations have been treated as a representative example of a learning organization. The basis for this assumption comes from numerous interviews explaining the principles, priorities and goals of their operation.

4.1 Using COBIT as a Standard Evaluation Model of the Development Processes of IT Organizations

For measuring the development of an IT learning organization, an appropriate evaluation model was needed. COBIT was chosen as a standard for organizational evaluation.

COBIT (*Control Objectives for Information and Related Technology*) is a collection of good practices in IT management (a so-called framework), developed by ISACA and the IT Governance Institute in 1996. COBIT covers all areas of management, administration and operation in IT. COBIT is a comprehensive standard for the implementation of control mechanisms and good practices in IT. Using COBIT does not exclude, nor is it a substitute for other standards. Moreover, in order to build a well-functioning IT department which is transparent in terms of its organization, it is desirable to reach for good practices describing a particular process in a more detailed way. COBIT helps, in a prominent manner, to indicate which mechanisms should be implemented and to what degree, albeit being, in itself, rarely the ultimate solution. Therefore, it is often referred to as an integrator, which "spans" all the other good practices (not just standards and norms, but also the author's organizational solutions in a company) used in a given organization [8].

With the use of COBIT, any IT organization can be described in a quantified form in terms of its maturity. For this reason, it was chosen as an evaluation model for subsequent experiments for verifying the Model of Knowledge Acquisition and Processing, as described in the next part of this article.

4.2 The Process of the Development of an IT Organization Driven by a Model

The verified Model of Knowledge Acquisition and Processing (on the basis of studies conducted in a learning organization environment) was used for the needs of an organizational process of learning guided by a model and carried out by three organizations. The process of organizational learning was supported by sub-models of knowledge acquisition and knowledge processing (MKA and MKP), in which two different evaluation models were employed. In the first case, COBIT was used as the evaluation model to assess the level of organizational processes.

COBIT was used to assess the organizational processes on two levels:

- the assessment of maturity based on the COBIT control objectives in the form of competence questions,
- the classification of control processes in terms of those which have a powerful impact on their level of maturity.

The experiment was divided into two stages. During phase 1, a complete COBIT audit was conducted in the studied IT organization and, as a result, processes with the lowest level of maturity were classified. They were identified, first of all, to improve the functioning of the organization. During phase 2 - after several months - the functioning of the organization was assessed again to see how the changes suggested on the basis of the COBIT model were implemented into the functioning of the organization [9].

Table 1. A repeated COBIT test in an examined organization - results

Measure	Value – study made in 2009	Value – study made in 2010
The number of tested COBIT processes	184	84 (only the processes that obtained weaker ratings in 2009 were measured)
The number of competence questions to be assessed	262	112 (the competence questions evaluated referred only to the processes which obtained weaker ratings in 2009)
The number of competence questions (the processes evaluated higher in the repeated test, in relation to the first test)	47 questions (38 processes)	
The number of competence questions (processes) with a lower assessment in a repeated test, in relation to the first test	0 (0)	
The number of competence questions (processes) with identical assessment in a repeated test, in relation to the first test	65 questions (53 processes)	

The research shows a positive trend in the levels of process maturity, all the process domains, and hence - the entire organization. The study, which was conducted in the organization of an Internet service provider, after 12 months, shows that as many as 47 competence questions (referring to 38 processes) out of the 112 taken into account (in the buffer) were rated higher. It can be observed that each of the four

domains noted an increase in the level of maturity and no investigated process was rated lower than previously.

Conducted in this way (namely, in two stages) the study gave an additional opportunity to show the dynamics of the learning organization's development. Thus, by examining the processes in a similar manner, at regular intervals, there could be attempts to identify both the change trends (to see if the growth is permanent or rapid) and to identify both the processes and all the process areas which increase their level of maturity the slowest.

5 Conclusion

The authors of this article have identified a gap in the theory of knowledge-based organizations. The problem of the inability to describe these organizations in a quantitative manner is reflected in the inability to both diagnose problems and to predict desired states in the future. The lack of such a measuring apparatus leads to a development which is not fully conscious, and therefore inefficient. Based on the experience of working with IT companies, the authors suggest a solution to this problem.

The assessment of a learning organization can be identified by its maturity, namely, by an in-depth analysis of the maturity of its processes. There are complete and proven methods of process evaluation. The most universal of them is COBIT. The authors show that IT organizations meet the conditions enabling them to fall into a group of learning organizations, and suggest that the methodology for measuring IT learning organizations should be based on COBIT. Nonetheless, it cannot constitute an independent solution, inter alia, due to the fact that its character is statistical.

The Model of Knowledge Acquisition and Processing, in the area of knowledge acquisition, creates conditions for entering any knowledge about the state of the organization by dividing knowledge in terms of its type and quality. In the area of knowledge processing, it introduces mechanisms of processing knowledge which is imperfect. Although the model has been prepared to support the development of learning organizations, the experiments, however, indicated the possibility of using it in other domains - both of a technical nature (where knowledge is certain and complete) and social nature (with imperfect knowledge). There are plans to establish whether (after possible modifications) the model can be called a generic one.

References

1. Perechuda, K. (ed.): Zarządzanie wiedzą w przedsiębiorstwie. Wydawnictwo Naukowe PWN, Warszawa (2005)
2. Pedler, M., Burgoyne, J., Boydell, T.: The learning company. A strategy for sustainable development. Mcgraw-Hill, London (1997)
3. Zgrzywa-Ziemiak, A., Kamiński, R.: Rozwój zdolności uczenia się przedsiębiorstwa, Difin, Warszawa (2009)
4. Galata, S.: Strategiczne zarządzanie organizacjami. Wiedza, intuicja, strategie, etyka, Wydawnictwo Difin, Warszawa (2004)

5. Sitek, T.: Model of Knowledge Acquisition and Processing for Management of the Learning Organization, PhD Thesis, Gdańsk (2011)
6. Orłowski, C., Sitek, T.: Supporting Management Decisions with Intelligent Mechanisms of Obtaining and Processing Knowledge. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010. LNCS, vol. 6277, pp. 571–580. Springer, Heidelberg (2010)
7. Orłowski, C., Rybacki, R., Sitek, T.: Methods of Incomplete and Uncertain Knowledge Acquisition in the Knowledge Processing Environment. In: Jędrzejowicz, P., Nguyen, N.T., Howlet, R.J., Jain, L.C. (eds.) KES-AMSTA 2010. LNCS, vol. 6071, pp. 340–350. Springer, Heidelberg (2010)
8. COBIT 4.1 Control Objectives Management Guidelines Maturity Models Framework – documentation. IT Governance Institute (2007)
9. Orłowski, C., Sitek, T., Nalewajko, M.: Badanie Technologii Informatycznych u Dostawcy Usług Internetowych. In: Górski, J., Orłowski, C. (eds.) Inżynieria Oprogramowania W Procesach Integracji Systemów Informatycznych. Pomorskie Wydawnictwo Naukowo-Techniczne PWNT, Gdańsk (2010)

Prediction Based Handovers for Wireless Networks Resources Management

Piotr Rygielski, Paweł Świątek, Krzysztof Juszczyszyn, and Adam Grzech

Institute of Computer Science, Wrocław University of Technology,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{piotr.rygielski,pawel.swiatek,krzysiek,adam.grzech}@pwr.wroc.pl

Abstract. The paper is devoted to the problem of mobility and resources management in heterogeneous wireless networks. It is assumed that in certain area covered by multiple overlapping wireless networks there is certain number of mobile client which consume networks resources by use of available communication services (e.g.: voice or data transmission) delivered by network providers. Moreover it is assumed, that communication services continuity may be assured with use of common handover techniques supporting clients mobility (e.g: MIPv6, IEEE 802.21, etc.). The task of mobility and resources management consists of making decisions concerning the moment and the network to which particular clients should be handed over in order to optimize certain quality criterion (e.g. utilization of network resources). In this paper we show that gathering knowledge about clients movement and prediction of their future position may significantly improve the overall quality of delivered services and network resources utilization.

Keywords: network-assisted handover, user mobility prediction, wireless network management.

1 Introduction

Wireless networking is becoming more and more popular. Personal devices manufacturers equip their laptops, PDAs and smartphones with each and every wireless networking technology available, beginning with short-range bluetooth radios, through mid-range WLAN interfaces, ending on long-range 3G and WiMAX cards. As a result, the network services provider can deliver its services to mobile users 'anywhere and anytime'.

Providing high level of the quality of services (QoS) in mobile wireless environment raises a number of technical problems, which have to be overcome. The main issues here are: a) mobility management in heterogeneous wireless environment which allows for seamless handover between different data transmission technologies, and b) security and trust between different administrative domains allowing for secure transfer of credentials of a mobile device roaming through different networks.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 687–696, 2011.

A lot of research effort has been already made to address aforementioned problems. Intra-technology data link layer handovers are handled with use of media specific signaling procedures introduced as extensions to particular wireless transmission standards, e.g.: 802.11r for WLAN [8], 802.16e for WiMAX [7], etc. Currently a new standard is being developed to manage inter-technology layer 2 handovers. This proposal, 802.21 – Media Independent Handover [9], provides a set of mechanisms which allows to trigger higher layer media independent handover procedures based on unified set of commands and media specific events. Network layer handover is managed by mobility extensions of the IP protocol, i.e. MIPv4 [5] and MIPv6 [11]. In order to provide service continuity and guarantee required quality of service during handovers additional modifications to mobility management protocols has been made, e.g. FMIPv6 [12] and HMIPv6 [2]. However, a breakthrough in mobility management has been made by introduction of Media-Independent Pre-Authentication (MPA) mechanism [4], which allows for seamless and secure handovers between different administrative domains.

Application of above mechanisms in heterogeneous wireless environment allows to provide truly mobile services on *anywhere and anytime* basis. In such scenarios, where users are not bound to use particular wireless network, network selection and handovers do not affect users application performance, a number of network management and optimization tasks can be performed using network-assisted (network-enforced) handovers. These tasks include among others: network load balancing, user allocation, resources allocation, network resources utilization optimization and quality of service provisioning.

In this paper a general concept of network resource assignment optimization assisted by network-enforced handover is proposed. Network optimization tasks which utilize proposed concept are introduced. Moreover, it is shown that application of even simple methods of prediction of user's movement may significantly improve the efficiency of network management and optimization tasks. The influence of the proposed network optimization concept on the network performance is evaluated by means of computer simulation.

The paper is organized as follows. In section 2 we present assumed models of wireless networks and user-mobility. In section 3 we formulate the problem of assignment users to networks and propose the solution. Section 4 is devoted for experiment description and result analysis. Finally in section 5 we summarize the presented work.

2 Network and Mobility Model

2.1 Network Model

Assume that there are N wireless networks net_n $(n = 1, ..., N)$ covering certain area A. Exemplary area A is depicted in figure 1. We assume that each network net_n covers a circular area with a center $(net_n(x), net_n(y))$ and radius r_n. Each network net_n is characterized by maximal amount of available resources U_n. Depending on the amount of free network resources u_n a client may receive

certain amount of network capacity c_n calculated by function $c_n = f_n(u_n, r_n)$ which is specific to each access network net_n.

Assume that area A is divided into $I \times J$ identical square cells. Integers I and J are chosen in such a way, that cells are small enough for the characteristics of each network to be constant across the area of single cell. Each cell $cell_{ij}$ $(i = 1, \ldots, I; j = 1, \ldots, J)$ may contain any number of users and network access points. Coordinates of users and access points lying within particular cell $cell_{ij}$ are assumed to be equal to the coordinates of the cell. Moreover, we assume that the distance between certain user and certain access point is equal to the distance between cells containing user and access point.

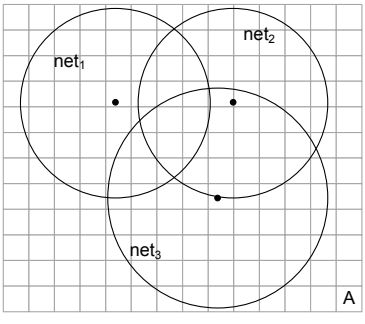


Fig. 1. The exemplary area A divided into 168 cells with networks coverage

The service of the network which can be delivered to multiple clients may have various interpretations in different access networks. In networks based on time-division multiplexing (TDM) medium sharing technique, allocated resource is interpreted as number of time slots in which a client is allowed to transmit data. On the other hand, in networks based on frequency-division multiplexing (FDM) certain amount of bandwidth is allocated to each user.

The form of function $f_n(u_n, r_n)$ depends among others on the type of network it is associated with. Another important factor is the ISO/OSI layer at which network capacity is measured. At the physical layer capacity c_n measured as number of bits send per second is roughly proportional to the amount of allocated resources and does not depend on the distance from the network access point. At the data link layer, where due to transmission errors datagram retransmission may occur, user's distance from the antenna plays important role. At higher layers of the ISO/OSI model protocol specific mechanisms requiring data retransmissions may further decrease delivered network capacity. In general it may be assumed that effective network capacity c assigned to a user is proportional to the amount of allocated resource u and inversely proportional to the distance r from the network access point:

$$c_n = f_n(u_n, r_n) = C_n \cdot \frac{u_n}{U_n} \cdot \frac{r_n - r}{r_n}, \tag{1}$$

where C_n is the maximal achievable effective capacity under assumption that a user is assigned with all available resources $u_n = U_n$ in a near zero distance from the network access point. For such a general model we add some assumptions and formulate mobility management tasks which are presented in section 3.

2.2 User's Mobility Models

The quality of services and performance of wireless networks highly depend on the position and movement trails humans which operate the various types of communication devices. Most of these devices are small, handheld equipment attached to their operators. It is rather difficult to deploy large-scale wireless networks for testing purposes, so various mobility models are used for simulations and performance evaluations. Mobility models, which reproduce the movement patterns of humans, are applied to make their behavior predictable, and support the algorithms used for network management [3]. We consider three mobility models, reflecting various statistical features observed in human activity patterns: Random Walk, Truncated Levy Flight (TLF) and Self-Similar Least Action Walk (SLAW).

The Random Walk model does not require much explanation — we assume stochastic movement with the maximum distance limit. The Truncated Levy Flight (TLF) model was based on the proposal discussed in the work [10], where the applicability of Levy walk model was proved for human mobility patterns. In particular — this result is especially interesting because the model verification in [10] was carried out on the data gathered in mobile telecommunication networks. Typically in experiments Levy exponent for flight length distribution and pause time distribution are equal to 1.5.

Self-Similar Least Action Walk was based on the model proposed in [13]. It is the most advanced approach, which covers several distinctive features observed in human mobility patters. First, it gives truncated scale-free distributions of flights (elementary movement actions) and pause times (time intervals between movements). This is the same feature which is also addressed by the TLF model. Moreover, the SLAW simulates the influence of individual mobility areas, typical for each user. The individual character of movement patterns was confirmed in [1]. The next feature modeled by SLAW are intercontact times. It is assumed, that the movements of individuals are correlated, and people tend to move in spontaneously formed groups which have truncated power law time distribution. The last feature are characteristic movement destination points which have fractal-type geographical distribution — this is used to simulate that some destinations are preferred and visited more often than the others. Following the results from [16] it is a characteristic feature observed in many scenarios, especially in an urban and industrial environments. The geographical space (area A) in our experiments is a 30×30 mesh with reflection boundary (points crossing the boundary turn back instead of returning on the other side of the mesh which is wrap-around boundary). According to the state-of-art we use the SLAW model in our experiments.

3 Problem Formulation

There is a number of mobility management tasks which can be performed to improve delivered quality of service in wireless networks. Each mobility management task can be formulated as an optimization problem which in general is NP-hard. In this section we focus on the problem of maximization of connected users number. Then we consider the simple prediction mechanism in order to minimize the overall handovers number.

3.1 Maximization of the Number of Connected Users

It is assumed that there are M users in the area A, each accessing one of the wireless networks. The task is to find such an assignment of M users to N networks for which the maximum number of users is connected with network.

Let matrix $R_{N\times M}$ ($R_{n,m} \in \{0,1\}$) models the possibilities of connecting users to the networks in the moment t. Value $R_{n,m} = 1$ means that m-th user is located within the range of n-th network, while $R_{n,m} = 0$ means that such a connection is impossible at t-th moment of time. Assuming that all variables are considered in the moment t we can formulate the following optimization problem.

$$\text{Maximize} \quad \sum_{n=1}^{N}\sum_{m=1}^{M} P_{n,m} \tag{2}$$

$$\text{subject to :} \quad \forall_{n=1\ldots N}\ \forall_{m=1\ldots m} P_{n,m} \in \{0,1\} \tag{3}$$

$$R \cdot P = P \tag{4}$$

$$\forall_{m=1\ldots M}\ \exists! P_{n,m} = 1 \tag{5}$$

$$\forall_{n=1\ldots N} \sum_{m=1}^{M} P_{n,m} \leq c_n. \tag{6}$$

It can be shown that the problem formulated above is in general a NP-hard optimization problem. The proof can be shown by transformation to *Multidimensional Multiple-choice Knapsack Problem* (MMKP) [14]. Therefore above formulation cannot be utilized in real-life applications where the number of users and networks may be large. In the next section we simplify the problem by assuming that each user requests the same amount of network resources. For such an assumption we present polynomial time exact algorithm.

3.2 Minimization of Resource Assignment Errors Number

The situation when user is not assigned with desired amount of resources we call the resource assignment error. The method presented in this subsection works if and only if all the users in the considered area requests the same amount of resources (e.g. the same maximum bandwidth). We can identify this with situation, when the mobile operator guarantees the equal throughput say 1Mbps for each mobile device with a contract.

Solution for this task is obtained by transformation of the original problem to the classic assignment task. Assuming that each user demands equal network capacity $c = c_n$ (for $n = 1, \ldots, N$) we create virtual wireless networks in the following way. Each n-th real network with total capacity U_n is divided into $\lfloor \frac{U_n}{c} \rfloor$ virtual networks. After such transformation we know that each of $K = \sum_{n=1}^{N} \lfloor \frac{U_n}{c} \rfloor$ virtual networks can handle exactly one user.

In the next step we build square binary matrix R' of size $max\{K, M\}$ in the following way. If m-th user can be connected (is located within the range) to virtual network k then $R'_{mk} = 1$, otherwise $R'_{mk} = 0$. If $K > M$ we add $K - M$ artificial users with possibility to connect each network. If $M > K$ we add $M - K$ artificial virtual networks where each virtual network can be accessed by any user in the area A.

Having the matrix R' filled with proper values we produce the assignment cost matrix denoted CR which defines the apparent cost of assignment the users to the virtual networks. The size of CR matrix is the same as the size of R'. The cost matrix CR is calculated in the following way. If $R'_{mk} = 1$ then $CR_{mk} = 0$; if $R'_{mk} = 0$ and $m < M$, $k < K$ then $CR_{mk} = b_1$; otherwise $CR_{mk} = b_2$. The b_1 and b_2 are any high numbers such that $b_1 > b_2$. In our experiments we take $b_1 = 1000$ and $b_2 = 500$. Such a cost matrix puts preference of connection of users that are within the range of any real wireless network. If user is not within the range of any network, the assignment cost is very high so the connection is unlikely to be preferred. The cost of assignment of artificial user or the assignment to the artificial network is also high, but lower than in the case when no network is available. This does not impact the overall solution because the artificial users are not going to be connected in real. The connection of real user to artificial network will cause the resource assignment error (in real the user will remain not connected) but will satisfy the general problem constraints (eq. 3–6).

Such a cost matrix is the only parameter of an algorithm solving the stated assignment problem. As the algorithm we use the Hungarian method [15] with cost minimization objective.

3.3 Minimization of Movement Prediction Errors Number

In this subsection we consider the minimization of prediction errors number problem which is a further extension of the resource assignment errors number minimization problem (eq. 2). We assume that the user location can be predicted for the next step thus we want to minimize the overall handovers number over the experiment time.

As the movement prediction error we understand the situation when handover of user occurred when it was not necessary — there existed such an assignment earlier that the handover would not occur now. The handover operation is not costless so we want to minimize handovers number in order to manage the wireless system resources properly.

In order to consider the prediction in the assignment algorithm we introduce the following changes to the cost matrix CR. Consider t-th moment of time. If

m-th user is within the range of k-th virtual network and will remain in the range of this network in the moment $t+1$, the cost equals $CR_{mk} = 5$. The cost equals $CR_{mk} = 10$ if the k-th network will be unavailable in the moment $t+1$ but is available in t. Moreover, if the m-th user is connected to the k-th network in the moment t and this network will be available in the moment $t+1$ the cost is $CR_{mk} = 0$. As the prediction algorithm we use the geometric mean. The solution is obtained again using Hungarian algorithm with cost minimization objective.

4 Experiment and Results Discussion

In order to evaluate the efficiency of proposed methods there has been a simulation environment developed in C++ with use of Qt library. Developed simulator simulates the set of users moving over a square area covered with wireless networks. The area A has been split into 900 (30×30) square cells where defined number of wireless radio stations and users were placed. Each user located within the area A behaves in the following way. Every step the user makes a movement from one cell to another one according to the mobility model. When user changes a cell the decision is being made which network should be connected. After each step the number of unconnected users was counted.

The experiments executed in the simulation study consist of running proper number of simulation runs, each with constant number of networks but with increasing number of users and using various assignment algorithms. The results for each simulation run contain statistics about the following parameters values: connection errors — how many times user's requirements were not satisfied (user was not connected); prediction errors — how many times the handover occurred even if there was earlier such an assignment possible that the handover would not occur.

We examined four algorithms of assignment: $USR1$ — the user makes decision which network to connect depending on the signal strength; $USR2$ — the user chooses network with lowest ping; $HEU1$ — the heuristic algorithm presented in [17]; $OPT1$ — the optimal assignment solution for problem given with equations 2–6.

The results of the first experiment are presented in the figure 2. Each simulation run was executed with different number of users in the area A. There was 600 steps of simulation. In each step every user made a movement according to the mobility model. The heuristic algorithms $USR1$ and $USR2$, where user makes assignment decision, caused the largest number of connection errors. Both USR algorithms were used to show the waste of wireless networks resources when the user is making the decision — the way that is mostly used in practice nowadays. In the heuristic $HEU1$ the networks were making decision which user to connect. This method outperforms the methods $USR1$ and $USR2$. The optimal method $OPT1$ presented in this paper maximizes the number of users which have connectivity to the network.

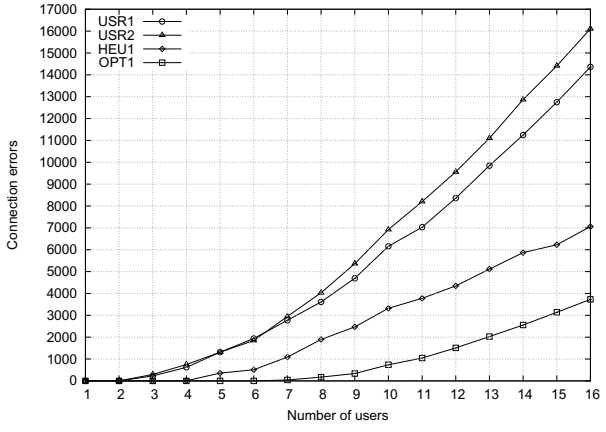


Fig. 2. Number of network resource assignment errors under control of various assignment algorithms for increasing number of users present in the area A

In the second experiment we investigate the quality of exact method by comparing the number of prediction errors when using the following prediction assumptions (in all cases the geometric mean is used as a predictor):

- no prediction of user location,
- low quality predictor — only partial information about user's past movements was available,
- hi quality predictor — only some data about past movements were missing,
- ideal predictor 1 step — we assume that we know exact position of user in the next step
- ideal predictor 2 step — we assume that we know exact position of user in the next two steps.

The results of the second experiment are presented in the figure 3. Each simulation run in this experiment gives equal connection errors number — same as in the first experiment for algorithm $OPT1$. This means that the modifications of algorithm does not change the optimality in sense of optimization criterion (eq. 2) chosen as minimization of connection errors. In this experiment we change the optimization criterion and compare number of prediction errors for the optimal algorithm.

The results show that prediction of future user location is worth the effort. Even the low quality prediction gives the satisfying improvement in the prediction errors number. Using more sophisticated methods of prediction improves the wireless network resources usage, assuming that each handover impacts the load of the network. In this case the high quality predictor gives satisfying results. More interesting observation is that using ideal predictor with longer prediction horizon does not improve significantly the results comparing to ideal predictor with shorter horizon.

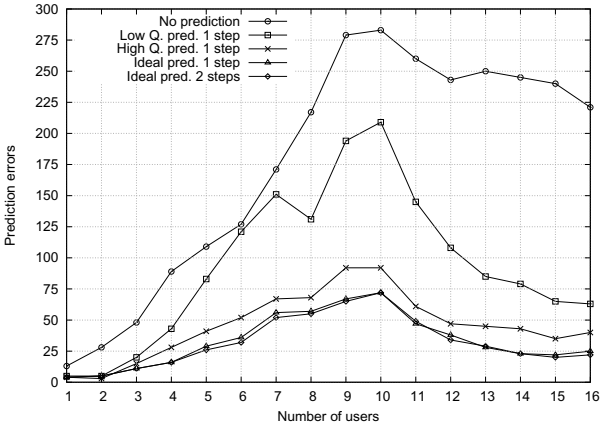


Fig. 3. Number of prediction errors under control of predictors with various quality using exact assignment algorithm for increasing number of users present in the area A

5 Final Remarks

The general problem formulated in this paper concerns a situation when the users are moving through area covered by many wireless networks. Every user uses the network but can be connected to any particular wireless station that is within the range. We have formulated the general problem of assignment users to the networks and the simplify it in order to show that prediction of user's location in the future causes less handovers. Moreover, we point out that sophisticated methods of prediction will rather not improve much the performance of the wireless system. In the future we plan to discard the simplifications introduced in this paper and propose efficient heuristic algorithm for general problem formulated in section 3. Moreover, we plan an application of such a methods in service-oriented [6] sensor data acquisition system.

Acknowledgments. This work was partially supported by the European Union from the European Regional Development Fund within the Innovative Economy Operational Programme project number POIG.01.01.02-00-045/09-00 "Future Internet Engineering". Fellowship co-financed by European Union within European Social Fund.

References

1. Brockmann, D., Hufnagel, L., Geisel, T.: The scaling laws of humantravel. Nature 439, 462–465 (2006)
2. Castelliccia, C.: HMIPv6: A hierarchical mobile IPv6 proposal. SIGMOBILE Mob. Comput. Commun. Rev. 4(1), 48–59 (2000)
3. Song, C., et al.: Limits of Predictability in Human Mobility. Science 327(5968), 1021–1081 (2010)

4. Dutta, A., Famolari, D., Das, S., Ohba, Y., Fajardo, V., Taniucho, K., Lopez, R., Schulzrinne, H.: Media-independent pre-authentication supporting secure interdomain handover optimization. IEEE Wireless Communications 15(2), 55–64 (2008)
5. El Malki, K.: Low-Latency Handoffs in Mobile IPv4. IETF RFC 4881 (June 2007)
6. Grzech, A., Swiatek, P.: Modeling and optimization of complex services in service-based systems. Cybernetics and Systems 40(8), 706–723 (2009)
7. IEEE Standard 802.16e: Air interface for fixed broadband wireless access systems amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands (December 2005)
8. IEEE Standard 802.11r-2008: IEEE Standard for Information Technology-Telecommunications and Information Exchange Between Systems-Local and Metropolitan Area Networks-Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 2: Fast Basic Service Set (BSS), (Amendment to IEEE Std 802.11-2007 as amended by IEEE Std 802.11k-2008), c1–108 (July 2008)
9. IEEE P802.21/D14.0, Draft Standard for Local and Metropolitan Area Networks: Media Independent Handover Services (September 2008)
10. Rhee, I., Shin, M., Hong, S., Lee, K., Chong, S.: On the Levy-walk Nature of Human Mobility. In: INFOCOM, Arizona, USA (2008)
11. Johnson, D.B., Perkins, C.E., Arkko, J.: Mobility Support in IPv6, IETF RFC 3775 (June 2004)
12. Koodli, R., et al.: Fast Handovers for Mobile IPv6, IETF RFC 4068 (July 2005)
13. Lee, K., Hong, S., Kim, S.J., Rhee, I., Chong, S.: SLAW: A Mobility Model for Human Walks. In: The 28th IEEE Conference on Computer Communications (INFOCOM), Rio de Janeiro, Brazil (April 2009)
14. Martello, S., Toth, P.: Heuristic algorithms for the multiple knapsack problem. Computing 27(2), 93–112 (1981)
15. Munkres, J.: Algorithms for the Assignment and Transportation Problems. Journal of the Society for Industrial and Applied Mathematics 5(1), 32–38 (1957)
16. Rhee, I., Lee, K., Hong, S., Kim, S.J., Chong, S.: Demystifying the levy-walk nature of human walks. Techical Report, NCSU (2008), http://netsrv.csc.ncsu.edu/export/DemystifyingLevyWalkPatterns.pdf
17. Swiatek, P., Rygielski, P.: Wireless Network Management Through Network-Enforced Handover. Applications of Systems Science, 227–236 (2010)

Author Index

Abbasi, Alireza II-256
Abbattista, Fabio I-249
Abe, Akinori II-495
Abe, Keiichi III-106
Adachi, Tomoya III-11
Adachi, Yoshinori IV-80, IV-117
Adrian, Benjamin II-420
Ahmadinia, Ali III-453, III-462, III-472
Aimi, Annuar H.B.M. III-415
Akdoğan, Erhan I-271
al Agroudy, Passant II-410
Alamgir Hossain, M. I-151
Albert, Dietrich IV-261
Alghowinem, Sharifa I-377
Alizadeh, Hosein I-21
Ambiah, Norbaitiah III-346
Anderson, Terry I-161
Aoki, Kumiko III-548
Aoki, Shingo IV-242
Aoki, Yuki III-558
Arakawa, Yutaka IV-252
Arasawa, Ryosuke IV-14
Arghir, Stefan I-1, I-72
Argotte, Liliana II-94
Aritsugi, Masayoshi IV-53
Arotaritei, Dragos I-1, I-72
Arroyo-Figueroa, G. II-94
Atteya, Walid Adly I-151
Aude, Aufaure Marie II-41
Aufaure, Marie-Aude II-538
Azzeh, Mohammad II-315

Baba, A. Fevzi I-90
Baba, Norio II-366
Badaracco, Miguel II-124
Bae, Junghyo I-289
Banba, Hideaki III-227
Bao, Yongguang IV-98
Baralis, Elena II-515
Bardis, Georgios I-347
Bardone, Emanuele II-486
Batres, Rafael III-395
Batsakis, Sotiris I-558
Baumann, Stephan I-495
Beigi, Akram I-110
Belanche, Lluís I-100
Beloe, Neil III-483
Ben-Abdallah, Hanene I-407
Ben Ahmed, Mohamed I-516
Benites, Fernando I-579
Ben Romdhane, Nadra I-407
Berns, Karsten IV-167
Bi, Yaxin I-161
Bianchi, Alessandro I-249
Biernacki, Pawel I-418
Boland, Katarina IV-366
Bonachela, Patricia II-611
Bondarenko, Andrey I-62
Boongasame, Laor I-230
Borzemski, Leszek II-581
Bouamama, Sadok II-325
Bouki, Yoshihiko III-153
Bravo, Jose II-296
Breiner, Kai IV-136
Brezovan, Marius I-505
Bridge, David III-483
Brucker, Florian I-579
Brusey, James III-483
Bueno, Gloria II-611
Bui, Len I-436
Burdescu, Dumitru Dan I-505
Burgin, Mark II-524
Burns, Nicola I-161

Cairó, Osvaldo I-316, II-306
Cálad-Álvarez, Alejandro II-601
Carlson, Christoph I-239
Carrasco, Eduardo II-611
Castellano, G. II-84
Ceccarelli, Michele I-568
Ceci, Michelangelo II-559
Cerquitelli, Tania II-515
Chang, Wei-Lun II-285
Chen, Bo-Tsuen II-382
Chen, Chia-Chen II-382
Chen, Hanxiong II-21
Chen, Mu-Yen II-382
Chen, Weiqin I-239, III-558

Chetty, Girija I-436
Chiang, Hsiu-Sen II-382
Chiusano, Silvia II-515
Chookaew, Sasithorn II-134
Chowdhury, Nihad K. II-355
Chu, Yuyi III-237
Constantin, George I-72
Cox, Robert I-377
Coyne, Bob IV-378
Csipkes, D. III-472
Csipkes, G. III-472
Cuzzocrea, Alfredo II-559, II-571

Dahal, Keshav I-151
Davies, Gwyn III-433, IV-425
Decker, Hendrik II-548
De Felice, Fabio I-249
Dengel, Andreas I-397, I-495, IV-187, IV-212, IV-222
de Schryver, Christian IV-177
Deßloch, Stefan IV-126
de Vey Mestdagh, Kees (C.N.J.) II-524
di Bella, Enrico II-392
Dolog, Peter II-505
Doran, Rodica-Elena II-265

Ebert, Sebastian IV-222
Eichhoff, Julian R. I-387
Eklund, Tomas II-186
Endo, Yasunori I-131
Enomoto, Yoshitaro III-246
Eto, Kaoru III-31

Fanelli, A.M. II-84
Faragó, Paul IV-455
Farjami, Sahar III-499
Farkas, Ioana-Iuliana II-265
Feng, Yaokai IV-195
Fernandez-Canque, Hernando III-453, III-462, III-472
Feştilă, Lelia IV-455
Firmansyah, Tatan III-395
Fontecha, Jesús II-296
Fouladgar, Hani II-214
Frank, Anette IV-366
Fruhata, Takashi III-11
Fuchino, Tetsuo III-423
Fujii, Satoru III-86, III-144
Fujita, Tomoki III-77
Fujita, Yoshikatsu III-378
Fujiwara, Minoru IV-288
Fujiwara, Reiko II-447
Fukuda, Akira IV-252
Fukui, Shinji IV-108
Fukumura, Yoshimi III-499, III-548
Fukushima, Taku II-31
Furuhata, Takashi III-1
Furuse, Kazutaka II-21
Futamura, Kazuya IV-117

Gaillourdet, Jean-Marie IV-126
Gălătuş, Ramona III-493
Garnik, Igor II-657
Gasmi, Ghada I-590
Gaura, Elena III-483
Gavrilova, Tatiana A. I-337
Geibel, Peter I-538
Genquan, Ren I-528
Georgieva, Olga I-82
Gesell, Manuel IV-167
Ghezala, Henda Ben II-538
Gill, Balpreet II-440
Goda, Kazumasa II-154
Godehardt, Eicke II-402
Golfinopoulos, Vassilios I-347
Gotoda, Naka III-21, III-520
Graña, Manuel II-611
Grand, Alberto II-515
Grauer, Manfred III-56
Grimaudo, Luigi II-515
Grivas, Stella Gatziu II-51, II-275
Grosvenor, Roger III-433, IV-425
Grundmann, Thomas IV-126
Grzech, Adam II-687
Guardati, Silvia I-316
Guijarro, Frank II-611

Ha, Taehyun I-289
Hajer, Baazaoui II-41
Håkansson, Anne IV-415
Hamaguchi, Takashi III-415
Hamasuna, Yukihiro I-131
Hammami, Mohamed I-407
Han, Lee Chen II-366
Hanabusa, Hisatomo III-596
Hanaue, Koichi IV-14
Hangos, Katalin M. III-385
Hara, Chihiro IV-288
Harada, Kouji IV-308
Haraguchi, Makoto II-457

Hartung, Ronald L. IV-409
Hasegawa, Mikio III-587
Hasegawa, Shinobu I-484
Hashimoto, Kiyota IV-261
Hashimoto, Takako II-73
Hashizume, Ayako III-197
Hattori, Fumio IV-44
Hattori, Masanori III-67
Hayashi, Yuki II-104, III-578, III-637
Heckemann, Karl IV-167
Helard, Maryline III-116
Hellevang, Mathias I-239
Henda, Ben Ghezala II-41
Henmi, Fumiaki III-227
Hernandez, Yasmin II-94
Hervás, Ramón II-296
Hintea, Diana III-483
Hintea, Sorin III-443, III-472, IV-455
Hirokawa, Sachio II-457
Hirosue, Noriaki III-177
Hochin, Teruhisa IV-1
Homenda, Wladyslaw IV-232
Horiguchi, Ryota III-596
Horiuchi, Kensuke III-67
Hossain, Liaquat II-256
Huang, Xu I-436
Hudelot, Céline I-538
Hung, Tzu-Shiang II-285

Ichikawa, Teruhisa III-134
Igarashi, Harukazu I-120
Iida, Takayuki III-67
Iijima, Chie III-246
Iizuka, Kayo III-366
Iizuka, Yasuki III-366
Ikeda, Mitsuru IV-288
Imono, Misako I-367, I-474
Inoue, Etsuko III-153
Inoue, Shuki IV-242
Inuzuka, Nobuhiro IV-89
Ionescu, Florin I-1, I-72
Iribe, Yurie III-548
Ishida, Yoshiteru IV-308, IV-318, IV-328, IV-338, IV-348, IV-357
Ishihara, Seiji I-120
Ishii, Naohiro III-616, IV-73, IV-98
Iswandy, Kuncup II-335, IV-155
Ito, Hideaki IV-34
Ito, Nobuhiro III-616
Itou, Junko III-77, III-126
Ivanciu, Laura III-443
Iwahori, Yuji IV-80, IV-108, IV-117
Iwamura, Masakazu IV-204
Iwashita, Motoi III-256, III-275
Iwata, Kazunori III-616

Jabban, Ahmad III-116
Jevtic, Dragan I-357
Jianmin, Wang I-528
Jimbo, Takashi IV-73
Jin, Ping II-505
Jlaiel, Nahla I-516
Johansen, Bjarte III-558
Jones, Leslie I-377
Jumutc, Vilen I-62
Jung, Matthias IV-177
Juszczyszyn, Krzysztof II-687

Kahl, Gerrit IV-187
Kambayashi, Yasushi I-260, I-280
Kameda, Hisashi III-606
Kamide, Norihiro I-599, II-225, II-235, II-246
Kamińska-Chuchmała, Anna II-581
Kanematsu, Hideyuki III-499
Kanenishi, Kazuhide III-520
Karadgi, Sachin III-56
Kashihara, Akihiro I-484, II-165
Kataoka, Nobuhiro III-207
Katayama, Shigetomo I-280
Katsumata, Yuji IV-328
Kawaguchi, Masashi IV-73
Kawai, Atsuo II-144
Kawai, Hideki II-63
Kawano, Kouji III-423
Kholod, Marina III-304
Kikuchi, Masaaki III-67
Kim, Daekyeong I-289
Kim, Ikno III-237
Kim, Jinseog II-203
Kimura, Naoki III-415
Kise, Koichi I-397, IV-204, IV-212
Kitagawa, Hiroyuki II-21
Kitajima, Teijii III-423
Kitami, Kodai III-285
Kitamura, Akira II-447
Kitani, Tomoya III-134
Kitasuka, Teruaki IV-53
Klawonn, Frank I-82
Klein, Andreas IV-146

Klinkigt, Martin I-397, IV-212
Kohtsuka, Takafumi I-280
Kojima, Masanori III-207
Kojiri, Tomoko II-104, III-578, III-637
Koketsu, Hiroaki III-616
König, Andreas I-424, II-335, IV-155
Köppen, Mario III-177
Korn, Ralf IV-177
Koschel, Arne II-275
Koshimizu, Hiroyasu IV-34
Kostiuk, Anton IV-177
Kotulski, Leszek I-180, I-190
Kouno, Shouji III-275
Kowalczyk, Ryszard I-200
Krömker, Susanne IV-366
Kubo, Masao III-627
Kuboyama, Tetsuji II-73
Kucharski, Bartosz II-640
Kunieda, Kazuo II-63
Kunimune, Hisayoshi III-529
Kurahashi, Setsuya III-356
Kuroda, Chiaki III-405
Kurosawa, Takeshi III-275
Kusztina, Emma III-510, III-568
Kuwabara, Kazuhiro I-326

Lakhal, Lotfi I-590
Laosinchai, Parames II-134
Lee, Chun-Jen II-285
Lee, Gyeyoung II-203
Lee, Hyungoo I-289
Lee, Seongjoon I-289
Lee, Shawn I-260
Leimstoll, Uwe II-51
León, Coromoto I-32
Leray, Philippe II-176
Leshcheva, Irina A. I-337
Leung, Carson K.-S. II-355
L'Huillier, Gaston II-11
Li, Li I-424
Li, Wei III-167
Li, You III-217
Li, Zhang I-528
Lin, Mu Fei III-558
Liu, Kokutan II-366
Liwicki, Marcus IV-187, IV-204, IV-222
Lokman, Gürcan I-90
Lovrek, Ignac I-357
Lu, Chung-Li II-285
Luckner, Marcin IV-435
Ludwiszewski, Bohdan II-657
Lukose, Dickson III-346

Maass, Wolfgang I-387
Madokoro, Hirokazu I-446
Maeda, Keita III-637
Maekawa, Yasuko IV-280
Magnani, Lorenzo II-486
Majima, Yukie IV-280
Makris, Dimitrios I-347
Malerba, Donato II-559
Mamadolimova, Aziza III-346
Mancilla-Amaya, Leonardo II-621
Mannweiler, Christian IV-146
Marmann, Frank II-430
Martínez, Luis II-124
Marxen, Henning IV-177
Masciari, Elio II-571
Massey, Louis II-1
Matsubara, Takashi III-627
Matsuda, Noriyuki III-49
Matsumoto, Chieko III-328
Matsumoto, Hideyuki III-405
Matsumoto, Kazunori III-285, IV-271
Matsuno, Tomoaki III-106
Matsuodani, Tohru III-336
Matsushima, Hiroshi IV-89
Matsuura, Kenji III-21, III-520
Maus, Heiko II-430, IV-212
Meixner, Gerrit IV-136
Mejía-Gutiérrez, Ricardo II-601
Memmel, Martin I-495, IV-126
Methlouthi, Ines II-325
Metz, Daniel III-56
Miaoulis, Georgios I-347
Mihai, Gabriel I-505
Minaei-Bidgoli, Behrouz I-21, I-110, II-214
Minarik, Milos I-11
Mine, Tsunenori II-154
Mineno, Hiroshi III-106, III-227
Mitsuda, Takao II-366
Miura, Hirokazu III-49
Miura, Motoki III-96, III-539
Miyachi, Taizo III-1, III-11
Miyaji, Isao III-86
Miyamoto, Takao IV-271
Mizuno, Shinji III-548
Mizuno, Tadanori III-106, III-207
Mori, Hiroyuki III-405

Mori, Yoko III-126
Morioka, Yuichi IV-98
Morita, Takeshi III-246
Mukai, Naoto III-606
Müller, Ulf III-56
Munemori, Jun III-77, III-126, III-167
Murai, Soichi IV-24
Muramatsu, Kousuke III-529
Mustapha, Nesrine Ben II-538
Mutoh, Kouji II-447
Myriam, Hadjouni II-41

Nagata, Ryo II-144
Nakagawa, Masaru III-153
Nakahara, Takanobu III-295
Nakahira, Katsuko T. III-499
Nakamura, Yu IV-261
NanakoTakata III-548
Nasser, Youssef III-116
Németh, Erzsébet III-385
Nguyen, Hoai-Tuong II-176
Nguyen, Ngoc Thanh I-210
Niimura, Masaaki III-529
Ninn, Kou II-366
Nishide, Tadashi III-77
Nishihara, Yoko II-469, III-265
Nishino, Kazunori III-548
Nishino, Tomoyasu III-40
Noda, Masaru III-415
Nomiya, Hiroki IV-1
Nonaka, Yuki III-587
Nunez Rattia, Rodrigo III-499
Nyu, Takahiro III-96

Oberreuter, Gabriel II-11
Oehlmann, Ruediger II-440
Ogata, Hiroaki III-520
Ohira, Yuki IV-1
Ohmura, Hayato IV-53
Ohmura, Hiroaki II-21
Ohsawa, Yukio II-469
Okada, Yoshihiro IV-63
Okamoto, Masayuki III-67
Okamoto, Ryo II-165
Okamoto, Takeshi IV-298
Oku, Kenta IV-44
Okubo, Yoshiaki II-457
Olarte, Juan Gabriel II-306
Oltean, Gabriel III-443
Omachi, Shinichiro IV-204
Onishi, Rie III-144
Onozato, Taishi I-280
Oosuka, Ryuuji III-106
Orłowski, Aleksander II-650
Orłowski, Cezary II-677
Orozco, Jorge I-100
Osogami, Masahiro I-296
Otsuka, Shinji III-21, III-520
Ouziri, Mourad I-548
Ozaki, Masahiro IV-80

Pagnotta, Stefano M. I-568
Pan, Rong II-505
Panjaburee, Patcharin II-134
Parra, Carlos II-611
Parvin, Hamid I-21, I-110, II-214
Pellier, Damien I-548
Pertiwi, Anggi Putri I-52
Petrakis, Euripides G.M. I-558
Petre, Emil IV-388
Pfister, Thomas IV-167
Pham, Tuan D. I-466
Pichanachon, Akawuth I-230
Pietranik, Marcin I-210
Plemenos, Dimitri I-347
Poetzsch-Heffter, Arnd IV-126
Prickett, Paul III-433, IV-425

Rakus-Andersson, Elisabeth IV-399
Ramirez-Iniguez, Roberto III-453, III-462, III-472
Ramstein, Gérard II-176
Refanidis, Ioannis II-114
Ren, Fuji I-456
Resta, Marina II-372
Ríos, Sebastián A. II-11
Rivera, Fernando II-306
Ro, Kou II-366
Rombach, Dieter IV-136
Rostanin, Oleg II-410
Roth, Michael IV-366
Rouhizadeh, Masoud IV-378
Rousselot, François II-345, IV-445
Różewski, Przemysław III-510, III-568
Ruiz-Arenas, Santiago II-601
Rumyantseva, Maria N. I-337
Rybakov, Vladimir V. I-171, I-306, II-478
Rygielski, Piotr II-591, II-687

Saga, Ryosuke III-285, IV-271
Saito, Muneyoshi III-356
Sakamoto, Yuuta III-86
Sanchez, Eider II-611
Sanín, Cesar II-621, II-631, II-667
Sapozhnikova, Elena I-579
Sarlin, Peter II-186
Sasaki, Kazuma IV-357
Sasaki, Kenta III-67
Sato, Hiroshi III-627
Sato, Kazuhito I-446
Satou, Yuuki III-86
Sauter, Rolf II-275
Schaaf, Marc II-51, II-275
Schäfer, Walter III-56
Schirru, Rafael I-495
Schmidt, Benedikt II-402
Schmidt, Karsten IV-126
Schneider, Jörg IV-146
Schneider, Klaus IV-167
Schotten, Hans D. IV-146
Schuldes, Stephanie IV-366
Schwarz, Sven II-420, II-430
Sedziwy, Adam I-180, I-190
Segredo, Eduardo I-32
Segura, Carlos I-32
Seissler, Marc IV-136
Sekanina, Lukas I-11
Selişteanu, Dan IV-388
Şendrescu, Dorin IV-388
Seta, Kazuhisa III-558, IV-261, IV-288
Shida, Haruki IV-298
Shigeno, Aguri II-31
Shigeyoshi, Hiroki IV-242
Shiizuka, Hisao III-197
Shim, Kyubark II-203
Shimada, Satoshi IV-280
Shimada, Yukiyasu III-423
Shimogawa, Shinsuke III-275
Shintani, Munehiro I-260
Shiraishi, Soma IV-195
Shirota, Yukari II-73
Sikora, Katarzyna III-568
Sikorski, Marcin II-657
Sirola, Miki II-196
Şişman, Zeynep I-271
Sitarek, Tomasz IV-232
Sitek, Tomasz II-677
Sklavakis, Dimitrios II-114
Slimani, Yahya I-590
Soga, Masato III-40
Son, Hongkwan I-289
Söser, Peter IV-455
Sproat, Richard IV-378
Stanescu, Liana I-505
Stefanoiu, Dan I-72
Stratulat, Florin I-72
Stratz, Alex II-275
Stravoskoufos, Kostas I-558
Strube, Michael IV-366
Su, Ja-Hwung II-285
Sugihara, Taro III-539
Sunayama, Wataru III-265
Suyanto I-52
Suzuki, Motoyuki I-456
Suzuki, Nobuo III-378
Świątek, Paweł II-687
Szczerbicki, Edward II-621, II-631, II-640, II-650, II-667
Szpyrka, Marcin I-180, I-190

Tagashira, Shigeaki IV-252
Taguchi, Ryosuke III-499
Takahashi, Masakazu III-320
Takahiro, Masui III-106
Takai, Keiji III-304
Takano, Shigeru IV-63
Takeda, Kazuhiro III-415
Takeda, Yasuchika IV-108
Takeshima, Syujo III-49
Takeuchi, Shin IV-89
Taki, Hirokazu III-40, III-49
Takimoto, Munehiro I-260
Takubo, Yuto III-1
Talonen, Jaakko II-196
Tamano, Keniti IV-242
Tamura, Hitoshi I-280
Tanaka, Hidekazu IV-98
Tanaka, Katsumi II-63
Tanaka, Kouji III-328
Tanaka, Toshio III-21, III-520
Tanida, Akihide I-484
Thieme, Sandra IV-187
Ting, Lan I-528
Todorov, Konstantin I-538
Tokumitsu, Masahiro IV-318
Tomczak, Jakub M. II-591
Topuz, Vedat I-90
Toro, Carlos II-611
Torsello, M.A. II-84

Tran, Dat I-436
Tran, Trong Hieu I-200
Trapp, Mario IV-167
Tschumitschew, Katharina I-82
Tseng, Vincent S. II-285
Tsuchiya, Seiji I-367, I-456, I-474
Tsuda, Kazuhiko III-320, III-328, III-336, III-378
Tsuji, Hiroshi IV-242, IV-261, IV-271
Tung, Ta Son IV-44

Uchida, Seiichi IV-195, IV-204
Ueda, Takuya IV-338
Ueno, Tsuyoshi IV-242
Uetsuki, Keiji III-336
Umano, Motohide IV-288
Unno, Masaru III-310
Uosaki, Katsuji I-296
Ushiama, Taketoshi IV-24
Utsumi, Yuya I-446

Velásquez, Juan D. II-11
Villarreal, Vladimir II-296
Vo, Quoc Bao I-200
Voiculescu, E. III-493
Vukovic, Marin I-357

Wang, Bo III-217
Wang, Hui I-161
Wang, Peng II-631
Wanichsan, Dechawut II-134
Watabe, Hirokazu I-367, I-474
Watada, Junzo III-187, III-217, III-237
Watanabe, Nayuko III-67
Watanabe, Shosuke III-1
Watanabe, Toyohide II-104, III-578, III-637, IV-14
Wathanathamsiri, Sakon I-230
Wehn, Norbert IV-177
Werner-Stark, Ágnes III-385
Wolff, Daniela II-51
Woodham, Robert J. IV-108
Wu, Juiyu III-237
Wyrwiński, Jan II-657

Xu, Guandong II-505
Xu, Hua III-310
Xu, Yanhao I-424

Yaakob, Shamshul Bahar III-187
Yada, Katsutoshi III-295, III-304
Yamada, Keiji II-63
Yamada, Kunihiro III-86, III-207, III-227
Yamagiwa, Shinichi III-21
Yamaguchi, Takahira III-246
Yamanishi, Teruya I-296
Yamano, Takayuki I-220
Yamazaki, Atsuko K. III-31
Yan, Wei II-345, IV-445
Yano, Yoneo III-21, III-520
Yasunaga, Shotaro I-326
Yim, Jaegeol II-203
Yinwen, Zhang I-528
Yoshida, Akira IV-204
Yoshida, Kaori III-177
Yoshida, Kouji III-86, III-144, III-207
Yoshihiro, Takuya III-153
Yoshimura, Eriko I-367, I-474
Yoshino, Takashi I-220, II-31
Yuizono, Takaya III-167
Yusa, Naoki III-227

Zanni-Merk, Cecilia II-345, IV-445
Zatwarnicka, Anna I-141
Zatwarnicki, Krzysztof I-42, I-141
Zghal, Hajer Baazaoui II-538
Zhang, Haoxi II-667
Zhang, Xicen III-578
Zong, Yu II-505
Zühlke, Detlef IV-136