ESTUDIO DE LA WEB COLOMBIANA: TOPOLOGÍA Y VISUALIZACIÓN

MONOGRAFÍA

CARLOS ANDRÉS ARDILA PADILLA
JORGE ALEJANDRO NÚÑEZ GIRALDO

UNIVERSIDAD EAFIT
DEPARTAMENTO DE INFORMÁTICA Y SISTEMAS
Medellín
2007

ESTUDIO DE LA WEB COLOMBIANA: TOPOLOGÍA Y VISUALIZACIÓN

MONOGRAFÍA

CARLOS ANDRÉS ARDILA PADILLA
JORGE ALEJANDRO NÚÑEZ GIRALDO

Trabajo de grado para optar al título de
Ingeniero de Sistemas

Asesor
Hernán Darío Toro
Docente Departamento Informática y Sistemas

UNIVERSIDAD EAFIT
DEPARTAMENTO DE INFORMÁTICA Y SISTEMAS
Medellín
2007

Nota de Aceptación

Presidente del Jurado

Jurado

Jurado

"*A todas aquellas personas que me apoyaron tanto moral como económicamente, principalmente a mi familia: mi abuela Pastora, mi madre Orlinda, mi hermano Juan Pablo y Germán que desde la distancia hicieron todo lo posible para la culminación de esta etapa.*
A Paola por su constante motivación, empeño y preocupación."

Carlos Andrés Ardila

"*A toda mi Familia, en especial mis padres Jorge y Luz Elena, y a mis hermanas, Laura y Diana, quienes a pesar de los tropiezos, mantuvieron la confianza y propiciaron los medios para que esta parte de mi proyecto de vida se llevara a cabo.*'
Por último a Naty, por el ánimo cuando ya no habían fuerzas y por el apoyo a la moción de inversión tecnológica en el apartamento."

Jorge Alejandro Núñez

Agradecimientos

Los autores de este proyecto expresan sus agradecimientos a:

Juan Francisco Cardona McCormick, nuestro Asesor de Investigación durante una fase inicial de análisis del proyecto, quién desde el primer momento en que trabajó con nosotros, mostró un gran interés y compromiso con el cual nos motivó aún más a desarrollar el proyecto de la mejor forma, siguiendo sus valiosas orientaciones.

Paola Andrea Tapias, Diseñadora Gráfica encargada de las imágenes del proyecto y toda la parte visual en cada una de sus etapas. Además, su experiencia en manuales, fue de gran ayuda para la realización del manual de usuario.

Juan Pablo Ramirez y Andrés Cano, Ingenieros del Centro de Informática, por su colaboración en la minuciosa revisión del proyecto.

Índice general

Introducción **X**

1. Marco teórico **1**

1.1. ¿QUÉ ENTENDEMOS POR LA INTERNET? 1

1.1.1. Antecedentes 3

1.1.2. Internet en Colombia 6

1.1.3. Componentes de la Web 16

1.2. BREVE EXPLICACIÓN DE LA TEORÍA DE GRAFOS . . . 20

1.2.1. Grafos Dirigidos 21

1.2.2. Recorridos en grafos dirigidos 23

1.2.3. Clasificación topológica 25

1.3. LA TEORÍA DE LA CONECTIVIDAD DE LA WEB 29

1.3.1. Modificaciones a la teoría de Conectividad de la Web . 31

2. Metodologías y Tecnologías utilizadas **36**

2.1. HERRAMIENTAS DE DESARROLLO 36

2.1.1. Java como lenguaje de Programación 37

2.1.2. Eclipse 38

2.1.3. Estandares de programación J2EE 38
2.2. SISTEMA DE GESTIÓN DE LA BASE DE DATOS 40
2.2.1. ¿Por qué MySql? 40
2.2.2. Características Particulares de MySql 41
2.2.3. Funcionalidades empleadas en el proyecto 42
2.3. COMPONENTES PRINCIPALES DE LA APLICACIÓN . . 45
2.3.1. Consultor 46
2.3.2. Analizador 47
2.3.3. Clasificador 48
2.3.4. Visualizador 50
2.3.5. Base de datos WebCol 51
2.4. EXPRESIONES REGULARES 54
2.4.1. Aplicaciones en el Consultor 56
2.4.2. Aplicaciones en el Analizador 58
2.5. ESTRATEGIAS APLICADAS EN EL DESARROLLO 61
2.5.1. Construcción del Crawler 62

3. Análisis de Resultados 64
3.1. Resumen del Proceso 64
3.2. Resultados Obtenidos 66
3.2.1. Principales Resultados del Consultor y Analizador . . . 66
3.2.2. Principales Resultados del Clasificador 75

Conclusiones 84

Glosario 87

Índice de tablas

2.1. Ejemplo División en componentes en una url 59

3.1. Hosts con más páginas 68
3.2. Principales Dominios según su cantidad de Hosts 69
3.3. Porcentajes de Tipos de Páginas o Extensiones 71
3.4. Porcentajes de los dominios de primer nivel, según el número de sitios que los componen 72
3.5. Porcentaje de Participación de los tipos de archivos o complementarios en la Web colombiana 73
3.6. Porcentaje de Participación de las naturalezas de los archivos complementarios en la Web Colombiana 74
3.7. Porcentaje de participación de los tipos de componentes según el número de hosts o sitios que los componen 76
3.8. Porcentajes de participación de los Dominios de Primer Nivel Territorial, en el "MAIN" de la muestra, según el número de sitios que los componen 77
3.9. Número de hosts que apuntan al sitio relacionado(Anteriores de un Host) 79

3.10. Cantidad de hosts colombianos a los cuales apunta cada sitio relacionado(Siguientes de un Host) 80
3.11. Cantidad de hosts *fuera del dominio colombiano* a los cuales apunta cada sitio relacionado 81
3.12. Cantidad de sitios que hay en los determinados tipos de componentes, de acuerdo a la distancia a la cual se encuentran con respecto a sus tipos de componentes adyacentes 82

Índice de figuras

1.1. Representación del Sistema de Nombre de Dominio (*DNS*) . . 19
1.2. Grafo Dirigido Simple 21
1.3. Representación Clásica de un Grafo Dirigido 24
1.4. Ejemplo de GDA de Componentes Fuertemente Conexos . . . 26
1.5. Resultado del estudio de la Web como un grafo dirigido [2] . . 29
1.6. Subdivisiones adicionadas a la estructura original [20] 32

Introducción

En el mundo de hoy la Internet se ha vuelto una herramienta de trabajo esencial en cada empresa, ya que permite ser utilizado como un medio de comunicación abierto, de divulgación e intercambio de información entre múltiples usuarios desde cualquier lugar del mundo que ofrezca el servicio de conexión. Además, permite publicar y acceder un sin límite de temas, fomentando la investigación y proporcionando soluciones o conocimiento empírico y científico en todas las ramas del saber.

Estas y muchas otras cualidades hacen que la Internet posea una característica que la hace única entre los inventos del último siglo: no es posible controlar su notable crecimiento, ya que cada individuo en todo el planeta que tenga acceso a la Internet tiene la posibilidad de crear otro fragmento de ella. Es por esta razón que su estructura no tiene ninguna forma determinada y existen pocos medios para representarla de una manera gráfica.

El objetivo de este trabajo es ubicar contextualmente al lector en el campo de la Internet de Colombia, explicando sus componentes y características de

cada uno de ellos, para así poder llegar a clasificar la Internet de nuestro país en grupos bien definidos y bastante conocidos entre los expertos del tema a nivel mundial.

Esta monografía está organizada en 3 capítulos: en el capítulo 1 se presenta un breve marco teórico sobre qué es la Internet, su historia tanto a nivel mundial como en nuestro país, cuáles son sus componentes, cómo se relaciona la teoría de grafos con la Internet, en qué consiste la teoría de la conectividad de la Web de Broker - Kumar - Maghoul - Raghavan [2] y las modificaciones a dicha teoría realizadas por Ricardo Baeza Yates de la Universidad de Chile [20]. El capítulo 2 esta constituido por una descripción de las metodologías y tecnologías usadas para el desarrollo del sistema base de la investigación. Finalmente, en el capítulo 3 se explican los resultados entregados por el sistema, evidenciando de alguna forma el estado de madurez de la Internet en nuestro país, permitiendo pensar en un análisis futuro más profundo de los resultados que se obtengan, logrando observar a través de modelos comparativos, en qué nivel estamos frente a los otros países que ya han hecho investigaciones similares.

Se espera que la información contenida en esta monografía sea de un gran aporte para la Universidad EAFIT en el área de investigaciones de redes y telemática y además sea de gran utilidad e interés para la comunidad universitaria, porque ofrece una forma nueva de ver la Internet y sus características desde un punto de vista topológico, brindando una idea del estado de madurez de la Web colombiana con sus debilidades y fortalezas de acuerdo a la teoría y la práctica en que se basa toda esta temática.

Capítulo 1

Marco teórico

1.1. ¿QUÉ ENTENDEMOS POR LA INTERNET?

Cada uno de los usuarios que tiene la Internet a nivel mundial posee una percepción diferente de ella, pero en su concepto más estricto, "la Internet es una red global compuesta por subredes gubernamentales, académicas, organizacionales, comerciales, militares, corporativas y personales que abarcan todo el mundo. La Internet fue desarrollada originalmente por el ejército norteamericano, y poco después se popularizó en la investigación académica y el sector comercial. Los usuarios que tienen acceso a la Internet pueden leer y descargar datos, virtualmente acerca de cualquier tema, desde casi cualquier

parte del mundo". Esta definición, que ofrece la biblioteca virtual Portal Gerona[4], es de bastante interés para este estudio porque desde ya comienzan a relucir las primeras segmentaciones que posee la Internet, gracias a su dominio de primer nivel que puede ser genérico o territorial.

La biblioteca virtual Wikipedia de España ofrece la siguiente definición: "La Internet es una red de redes a escala mundial de millones de computadoras interconectadas con el conjunto de *protocolos TCP/IP* (Transmission Control Protocol / Internet Protocol). También se utiliza este nombre como sustantivo común y por tanto en minúsculas para designar a cualquier red de redes que use las mismas tecnologías que la Internet, independientemente de su extensión o de que sea pública o privada".[26]

Según glosarios técnicos que se encuentran en línea, la Internet es: "red de redes. Sistema mundial de redes de computadoras interconectadas. Fue concebida a fines de la década de 1960 por el Departamento de Defensa de los Estados Unidos; más precisamente, por la *ARPA* (*Advanced Research Projects Agency* o Agencia de Proyectos de Investigación Avanzada). Se la llamó primero ARPANET (*Advanced Research Projects Agency NETwork*) y fue pensada para cumplir funciones de investigación. Su uso se popularizó a partir de la creación de la *World Wide Web*. Actualmente, es un espacio público utilizado por millones de personas en todo el mundo como herramienta de comunicación e información". [22]

1.1.1. Antecedentes

A mediados de la década de 1960, en la cúspide de Guerra Fría, el DoD (Department of Defense) estaba buscando una forma de mantener las comunicaciones vitales del país, en el posible caso de una Guerra Nuclear. Este hecho marcó profundamente su evolución, ya que aún ahora los rasgos fundamentales del proyecto se hallan presentes en lo que hoy conocemos como la Internet. En primer lugar, el proyecto contemplaba la eliminación de cualquier "autoridad central", ya que sería el primer blanco en caso de un ataque; en este sentido, se pensó en una red descentralizada y diseñada para operar en situaciones difíciles. Cada máquina conectada debería tener el mismo status y la misma capacidad para enviar y recibir información.

En 1968, el Laboratorio Nacional de Física de la Gran Bretaña estableció la primer red experimental. Al año siguiente, el Pentágono de los Estados Unidos decidió financiar su propio proyecto, y en 1969 se establece la primer red en la Universidad de California (UCLA) y poco después aparecen tres redes adicionales. Nacía así ARPANET, antecedente de la actual Internet. Gracias a ARPANET, científicos e investigadores pudieron compartir recursos informáticos en forma remota; esta era una gran ayuda porque hay que recordar que en los años 70's el tiempo de procesamiento por computadora era un recurso realmente escaso. ARPANET en sí misma también creció y para 1972 agrupaba a 37 redes.

Y sucedió algo curioso: empezó a verse que la mayor parte del tráfico

estaba constituido por noticias y mensajes personales, y no por procesos informáticos, lo que produjo que ya en 1977, otro tipo de redes no necesariamente vinculadas al proyecto original, empezaran a conectarse. En 1983, el segmento militar de ARPANET decide separarse y formar su propia red que se conoció como MILNET (*Military Network*). Luego en 1984, la Fundación Nacional para la Ciencia *NSF* (National Science Foundation) inicia una nueva "red de redes" vinculando en una primera etapa a los centros de súper computo en los Estados Unidos a través de nuevas y más rápidas conexiones. Esta red se le conoció como NSFNET (*National Science Foundation's Network*) y se caracterizaba por su facilidad de acceso, lo que permitió la conexión de las instituciones educativas con redes más pequeñas. El crecimiento acelerado que experimentó NSFNET así como el incremento continuo de su capacidad de transmisión de datos, determinó que la mayoría de los miembros de ARPANET terminaran conectándose a esta nueva red y en 1989 ARPANET se declara disuelta.

Las redes que conformaban NSFNET escogieron identificarse por su localización geográfica, mientras que los demás integrantes se agruparon bajo seis categorías básicas o dominios: "gov", "mil", "edu", "com", "org" y "net". Los prefijos "gov", "mil" y "edu", se reservaron para instituciones de gobierno, instituciones de carácter militar e instituciones educativas respectivamente. El sufijo "com" empezó a ser utilizado por instituciones comerciales que comenzaron a conectarse a la Internet en forma exponencial, seguidos de cerca por instituciones de carácter no lucrativo, las cuales utilizaron el sufijo "org". Por lo que respecta al sufijo "net", éste se utilizó en un principio para las

computadoras que servían de enlace entre las diferentes sub-redes. En 1988 se agregó el sufijo "int" para instituciones internacionales derivadas de tratados entre gobiernos.[12]

En algún momento de finales de la década de 1980, la gente empezó a ver la aglomeración de redes como una interred, y más tarde como la Internet. El crecimiento continuó en forma exponencial, y para 1990 ya había crecido a 3000 redes y 200.000 computadoras. En 1992, ya contaba con un millón de servidores y para 1995 ya había cientos de redes regionales, decenas de miles de redes de área local (LAN), millones de servidores y decenas de millones de usuarios. El tamaño se duplicaba aproximadamente cada año.[21]

En los años siguientes, se desplegó un increíble crecimiento y desarrollo de la Internet, llegando cada vez a más lugares del mundo y aprovechando cada vez más el desarrollo paralelo que revelaba el hardware a nivel de telecomunicaciones, servidores, capacidad de procesamiento y almacenamiento, entre otros, haciendo cada vez más eficientes y rápidas las conexiones y multiplicando los usos que se le iba dando a la Internet, penetrando prácticamente en todos los mercados hasta llegar al punto de convertirse en una necesidad, y, de tal manera, estar a la vanguardia de las exigencias que se derivan de la globalización a la cual está sometido el mundo, que lo hace tan cambiante y a su vez demandante de recursos.

1.1.2. Internet en Colombia

La Universidad de los Andes fue uno de los principales actores en la conexión de Colombia con el resto del mundo a través de la Internet. Entre 1988 y 1996, quien en ese entonces era el director del centro de cómputo de la Universidad, lideró junto con otros colegas suyos, un gran proyecto de conexión de esa Universidad, y, por ende de todo el país.

Esta parte de la historia comienza en 1988 cuando nace la red de la Universidad de los Andes *RDUA* que conectaba sus edificios de ingeniería y su centro de cómputo, a través de cable coaxial grueso, usando *ETHERNET* (comunicación entre redes LAN's basada en tramas de datos. El nombre viene del concepto físico de *ether*) como protocolo de acceso al medio y utilizando *TCP/IP* como protocolo de comunicación.

En 1990 se "implementa la red de *Macs Local Talk*, usando los cables de backup del conmutador telefónico, permitiendo a la red llegar a todos los edificios de la Universidad y a todos los computadores Macintosh". [10]

Se implementa además una red experimental con configuraciones propias de la época como lo es la "*Token Ring*" (Arquitectura de red con topología lógica en anillo y técnica de acceso de paso de testigo) para iniciar el proceso de conexión entre las estaciones. En este mismo año, la Universidad de los Andes se conectó a la red *COLDAPAQ* de propiedad de Telecom, y se establece conexión con la Biblioteca Luis Ángel Arango del Banco de la República. Por último, el logro más significativo que se llevó a cabo fue la

conexión de la Universidad de los Andes a la red mundial *BITNET* (Because It's There NETwork), siendo administrador del nodo *RUNCOL*.

En 1991, se recibió la noticia de que la Universidad de Columbia dejaría de soportar *BITNET* a partir del siguiente año, para pasar a *BITNET II* que operaba sobre TCP/IP. Por esta razón, la Universidad de los Andes adquiere su primer enrutador y el software de emulación llamado *VMNET* (Virtual Machine Network) que implementa TCP/IP bajo *VM* (Virtual Machine). Este cambio de protocolo implicó una liberación de las direcciones que eran propiedad de la Universidad de Columbia y se le solicita a *INTERNIC* (Internet's Network Information Center) que le asigne a la nueva red, direcciones válidas dentro del dominio .co, que fue el definido para Colombia y cuyo administrador sería la Universidad de los Andes, por decisión de la misma entidad y basado en su experiencia administrando BITNET I.

Gracias a esta conexión la Universidad de los Andes tuvo acceso a los servicios de la Internet que ofrecía la Universidad de Columbia, pero los costos estaban siendo asumidos por la Universidad de los Andes y no existía aún una visión lo suficientemente fuerte como para ver la importancia y lo definitivo de esta herramienta para el desarrollo de un país. Por esta razón en ese mismo año se hace una solicitud a *Colciencias* y al comité de RUNCOL para que financien la conexión de Colombia a la Internet, pero dicha petición fue negada. A pesar de esto la Universidad de los Andes siguió adelante, asumiendo los costos hasta lograr madurar la idea en busca del convencimiento de las entidades gubernamentales sobre un patrocinio.

En Río de Janeiro en el año 1992, se realizó una reunión de países latinoamericanos compartiendo el deseo de conexión a la Internet. De allí surgió la necesidad de crear un *backbone* para la Internet en Colombia y esta iniciativa hizo que varias universidades se interesaran, pero finalmente quienes realmente participaron fueron la Universidad del Valle, la Universidad EAFIT y la Universidad de los Andes, todas ellas utilizando como red de transporte a *COLDAPAQ*.

En el último mes de ese mismo año, la Universidad de los Andes, con el patrocinio del ITEC (Instituto Tecnológico de Electrónica y Comunicaciones) de Telecom, presentó nuevamente un proyecto a Colciencias para conectar a nuestro país a la Internet, pero de igual manera que antes, fue negado.

En 1993, a través de un *backbone* (Mecanismo de conectividad primario en un sistema distribuido) ya implementado en Colombia, las universidades miembros de este proyecto desarrollaron localmente servicios de la Internet como lo son *News* (foro de discusión abierta, formado por distintos grupos de noticias temáticos), *Gopher* (acceso a la información a través de menús de forma arborescente), *FTP* (File Transfer Protocol), *Telnet* (protocolo estándar de la Internet que permite la conexión a un terminal remoto) y *Correo*. Paralelamente, en la Universidad de los Andes se realizaron los primeros sitios *Web* y finalmente, luego de muchos esfuerzos, se logró que Colciencias viera la importancia de la Internet y como consecuencia de esto, contrató a la Universidad EAFIT para desarrollar un proyecto con el fin de conectar a Colombia con la Internet, quien uniendo fuerzas con la Universidad de los Andes y el *Instituto Colombiano para el Fomento de la Educación Superior*

(ICFES), deciden presentar un proyecto unificado, reuniendo las propuestas y sugerencias más importantes de cada uno.

"En diciembre de 1993, en una reunión en el ICFES, miembros de este instituto, junto con representantes de Colciencias, la Universidad EAFIT, la Universidad del Valle y la Universidad de los Andes, acordaron la creación de una corporación de derecho privado cuyo fin primordial sería lograr la conexión y el desarrollo de la Internet en el país. Colciencias, como apoyo a la investigación y al desarrollo de la ciencia en el país, asume el costo inicial de la inversión que debe hacer la corporación. Nace así la corporación *InterRed*".

Ya en 1994, la Universidad de los Andes, quien era el único poseedor de la infraestructura necesaria en esa época, realiza el montaje del primer *ISP* (Internet Service Provider) de Colombia, mientras que InterRed gestiona la adquisición de los equipos y el personal indicado para su manejo.

Por recomendaciones de la *NSF*, y como medida de control del tráfico de la Internet que proviene de Latinoamérica, Europa y Asia, cada región debe tener un *enrutador* al cual se conectarían las redes que le correspondan. Por esto dicha entidad dona un enrutador, el cual fue ubicado en Homestead, Florida - USA.

"Durante las negociaciones para contratar el canal satelital con Telecom, éste insiste en utilizar COLDAPAQ como la red de transporte en contra de las recomendaciones de los miembros de InterRed. Al no llegar a un acuerdo

sobre este punto, Telecom decide lanzar su propio servicio de la Internet denominado *SAITEL*". [10]

Con la independización de Telecom en este sentido, se redujeron considerablemente las tarifas de acceso a la Internet ya que inicialmente éstas le quitaban la oportunidad de hacer uso de los beneficios de la Internet a casi todos los planteles educativos de enseñanza superior. Así, por medio de la empresa IMPSAT y gracias a la apertura del sector de las telecomunicaciones se pudo contratar el canal satelital.

Esta conexión tuvo un tropiezo más, que fue la destrucción de las instalaciones de la NSF en Homestead a manos del huracán Andrew, pero finalmente en junio 4 de 1994 se logra la conexión del país a la Internet usando la señal que llega a la Universidad de los Andes, proveniente de la Torre Colpatria, que recibe la señal de las instalaciones de IMPSAT en el Cerro de Suba (Bogotá), y quienes a su vez tienen conexión con Homestead. En julio de ese mismo año se pudo hacer el primer proceso de inscripción en línea de los estudiantes a los cursos, usando la Web. [5]

Luego del conjunto de esfuerzos hechos por parte de todos estos actores, y con el objetivo cumplido de conexión de Colombia a la Internet, se construyó el sendero hacia un gran crecimiento, desarrollo y consolidación de nuestra porción de la *autopista de la información.*

De acuerdo a eso y según estudios realizados, "ese desarrollo de la Internet en Colombia lo sitúa como uno de los países de América Latina y el

Caribe con el mayor crecimiento y, de acuerdo a datos proporcionados por el Ministerio de Comunicaciones, entre 1997 y 1999 el crecimiento del número de usuarios fue del 220 % y la densidad de conexiones (hosts) aumentó a una tasa promedio anual del 231 % entre 1993 y 1998". [5]

"En 1998, debido a la gran demanda de conexión en el país, se decide crear el *NAP Colombiano (Network Acces Point)*, el cual comenzó a funcionar en 1999 y es operado actualmente por la *Cámara Colombiana de Informática y Telecomunicaciones (CCIT)*". [5]

"El NAP es un punto de conexión nacional de las redes de las empresas que proveen el servicio de acceso a la Internet en Colombia, con el cual se logra que el tráfico de la Internet que tiene origen y destino en el dominio colombiano, utilice solamente canales locales o nacionales.

NAP Colombia permite el uso eficiente de la red de telecomunicaciones del país, produce una mejora significativa en el servicio de las empresas que lo conforman y reduce los costos por el uso de enlaces internacionales.

En este país el NAP es administrado por la CCIT, la empresa *INTESACOL* tiene a su cargo la operación mediante un contrato de outsourcing suscrito con la misma CCIT". [6]

Por la inserción de las Tecnologías de Información y Comunicaciones (*TIC's*) han surgido cierta cantidad de proyectos orientados a su desarrollo en Colombia. Uno de esos proyectos fue presentado en marzo del año

2000 en el cual el gobierno presenta la *Agenda de Conectividad – El Salto a Internet*, que sería un conjunto de estrategias desarrolladas a través de programas y proyectos específicos articulados entre sí, con el fin de lograr que el país aproveche el uso de las TIC's para su desarrollo económico, social y político.

El fin es de masificar el uso de las TIC's y con ello aumentar la competitividad del sector productivo, modernizar las instituciones públicas y de gobierno, y socializar el acceso a la información. [5, 14]

Continuando con la historia de la Internet en nuestro país, cabe describir además el contexto internacional al cuál pertenece Colombia y con el cual debe interactuar haciendo alianzas estratégicas que permitan alcanzar los objetivos más rápidamente.

Así entonces, existe un proyecto llamado "*América Latina Interconectada con Europa - ALICE*" que empezó a ser desarrollado en junio de 2003 y que será financiado hasta marzo de 2008. [13]

Este proyecto es administrado por *DANTE (Delivery of Advanced Net-Work Technology to Europe)* y la *Comisión Europea* fue su principal fundador. Actualmente ALICE posee 4 socios europeos y 19 Latinoamericanos y está conectado con Europa a través de un enlace entre Brasil y Madrid (España) llamado *GEANT2*. [17]

La significación de la sigla ALICE ilustra muy bien el objetivo general del

mismo, pero uno de sus objetivos específicos es la "creación de un *backbone* de una red latinoamericana de telecomunicaciones". Dicho objetivo se hizo realidad por medio del nacimiento de "*RedClara*" cuya creación duró dos años y medio. Pertenecen a esta red catorce países de Latinoamérica y se está pensando en crear una segunda versión de la misma. [13, 17]

Gracias a esos proyectos, en Colombia se derivaron otros muy importantes con los mismos fines. Uno de estos fue la conformación de la *Red Nacional Académica de Tecnología Avanzada - RENATA*, en enero de 2006.

El objetivo de esta red nacional es "conectar a la sociedad científica y académica de Colombia con el mundo"; es una reunión de esfuerzos de personajes importantes en el mundo de las comunicaciones, y han hecho parte de este proyecto entre otros colaboradores, el Ministerio de Comunicaciones y Educación, representantes de la Comunidad Europea, miembros de Colciencias y representantes de los principales Departamentos involucrados en el proyecto. [13]

Las más de 50 instituciones de educación superior y centros de investigación conectados a esta nueva red son los principales beneficiados porque afortunadamente no tienen que experimentar la misma curva de aprendizaje que atravesaron los países pioneros en el uso de las TIC's, por lo tanto se darán pasos mucho más grandes hacia el desarrollo, alcanzándolo más aceleradamente gracias a experiencias previas vividas por otros países, en comparación con proyectos que tuvieron que empezar prácticamente de cero y sin poderse referenciar a experiencias similares de otras entidades interesadas en

estos temas.

El nodo principal de RENATA se "establece en la sede Morato, de Colombia Telecomunicaciones (en Bogotá), cuya red es llamada *RUMBO (Red Universitaria Metropolitana de Bogotá)* y sus cinco puntas las constituyen los nodos principales de las Redes Académicas Regionales (*RAREs*) de las ciudades de Cali *(RUAV - Red Universitaria de Alta Velocidad)*, Barranquilla (*RUMBA - Red Universitaria Metropolitana de Barranquilla*), Medellín (*RUANA - Red Universitaria Antioqueña*), Bucaramanga (*UNIRED - Red de Universidades*) y Popayán (*RUP - Red Universitaria de Popayán*). En estas puntas se interconecta cada uno de los operadores locales, que son quienes manejan las redes metropolitanas de las universidades".[14]

La capacidad que soporta esta red es de veinte RARE's y su conexión con el mundo de las redes avanzadas se realiza a través de *Colombia Telecomunicaciones*, institución a cargo del transporte de datos entre los integrantes de la red en Colombia y entrega este tráfico en el *PoP* de *RedCLARA* en Panamá.

Fue en marzo de 2006 cuando Colombia y Nicaragua fueron conectadas a 10 Mbps a el *PoP* (*Point-Of-Presence* o punto de acceso desde un lugar a el resto de la Internet) de RedCLARA en Panamá y de Tijuana (México) respectivamente.

Finalmente, se destaca que el 2 de mayo de 2007, los Ministerios de Comunicaciones y Educación, Colciencias y las principales redes universitarias

del país, crearon la *Corporación RENATA*, que tiene por objeto desarrollar la infraestructura de una red de alta velocidad , así como articular y facilitar acciones para la ejecución en Colombia de proyectos colaborativos de educación, innovación e investigación científica.[8]

El acta de constitución de la Corporación RENATA fue firmada por el Ministerio de Comunicaciones, Ministerio de Educación, Colciencias, y las cinco redes universitarias de todo el país.[8]

Haciendo un recuento de ese período y de los logros realizados con respecto a este tema de la llegada de la Internet a Colombia, se observa a la Universidad EAFIT como uno de los participantes definitivos en dicho proceso y aún figura como gran inversionista de recursos para el mejoramiento de la tecnología a nivel nacional y sobre todo aquella que tiene que ver con comunicaciones a través de la Internet.

Con los anteriores apartes de la historia y actualidad de la Internet en Colombia se observa claramente la proyección que se tiene y la gran consolidación de las condiciones del medio para cada día permitir la creación de más páginas Web, sitios, entre otros, y ser miembro activo y definitivo de esta Gran Red. Sabiendo esto, se puede decir que el dinamismo y crecimiento exponencial de la Web colombiana afectará directamente la información entregada en este proyecto de grado ya que la base de datos con las páginas y enlaces obtenida inicialmente sería de forma análoga una representación de la topología actual de la Web en Colombia y se evidencia la necesidad de la ejecución periódica de las aplicaciones de software desarrolladas en este

proyecto, para poseer dicha representación lo más actualizada posible.

1.1.3. Componentes de la Web

Para tener aún más claridad en lo que es la Internet, o de qué se compone la Web, se describe a continuación sus componentes más notables:

- **Dirección IP**: Ubicación de una computadora dentro de una red. Es un domicilio numérico que tiene dos representaciones. La primera es la más utilizada actualmente que es la versión IPv4 y la cual consta de cuatro números de hasta 4 cifras separados por puntos. [17]
  También está la nueva versión de direcciones IP llamada *IPv6* la cual fue creada en 1996 debido al aumento no esperado de la cantidad de dispositivos que se necesitan conectar a la Internet. Un ejemplo de estas nuevas direcciones IP es: 2001:0db8:85a3:08d3:1319:8a2e:0370:7334 y aún no se ha masificado su utilización ya que se calcula que todavía las direcciones IPv4 soportan durante unos cuatro años más la demanda de usuarios y el cambio a este nuevo estándar acarrea muchos costos que no han sido asumidos en su totalidad. [15]

- **Dominio**: Conjunto de computadoras y dispositivos de una red que conforman una unidad y comparten las mismas reglas y procedimientos (También comparten una característica común, como la de estar en un mismo país, organización o departamento). En el caso de la Internet, se dice que pertenecen a un mismo dominio aquellas computadoras y

dispositivos que tengan en común una parte de la dirección IP. En el caso de Colombia, es cualquier nombre de la forma $x.y$ donde $y = .co$. [17]

- **Dominio de Primer Nivel**: En la Internet existen varios tipos de terminaciones de dominios o *dominios de primer nivel*. Estos son los .com, .org, .es, etc. Los dominios de primer nivel indican el ámbito al que pertenecen, hay principalmente dos grupos, genéricos (.net, .com, .edu, entre otros) y territoriales (.co, .es, .ar, entre otros). [29]

- **Subdominio**: Los subdominios permiten organizar mejor el dominio por ejemplo por secciones, departamentos, servidores de aplicaciones, entre otros. Internamente son subdirectorios del servidor virtual. Se puede decir que son dominios dentro de otro dominio, ya que subdivide dicho dominio a conveniencia de la organización. [29]

- **Página**: Archivo accesible en red global identificado por un URL (*Universal Resource Locator*: Localizador Universal de Recursos) exclusivo. A pesar de la incongruencia con el nombre, se suele utilizar el término "página Web" para referirse a un grupo de páginas que conforman una unidad, es decir, un sitio Web. [17]

- **Sitio**: Área especifica de la red global donde se encuentra una página o conjunto de páginas Web que conforman una unidad debido a que comparten un mismo tema e intención. La propiedad y administración de cada sitio Web corresponde a un individuo, empresa u organización. Por lo general, aunque no necesariamente, las páginas de un determinado sitio Web suelen almacenarse en un solo servidor. Cada sitio

contiene por lo menos una página inicial o principal, la cual es el primer documento que ve el usuario al acceder al sitio; además puede contener otros documentos y archivos. De una forma más resumida se puede decir que es un servidor Web lógico identificado por un subdominio, por ejemplo: *dis.eafit.edu.co* que es un sitio perteneciente al dominio *eafit.edu.co*. [17]

- **Enlace**: También conocido como *"hiperenlace", "hyperlink" o "link"*. Es la base del *hipertexto*, es decir, la posibilidad de pasar de un punto de una página Web a otro lugar de ella misma *(ancla)*, o de otra diferente. [17]

- **Buscador**: Herramienta que permite encontrar contenidos en la Red, buscando a través de palabras clave. Se categorizan en buscadores por palabra (como *Lycos*, *Google* o *Altavista*) y de directorios o índices (como *Yahoo!*). [18]

  Cabe destacar que el enlace de un buscador con un sitio, no se reconocerá como válido para este estudio, ya que éste no es un enlace real sino que es producto de una búsqueda profunda en la Web realizada por un Robot o Agente Inteligente que son aquellos empleados en los buscadores por palabra.

  Dentro de esta tipología de buscadores se incluyen todos aquellos recursos de búsqueda que emplean "robots" o máquinas para recorrer e indexar automáticamente páginas a lo largo de la Red. Todas las páginas recorridas por los "robots" son sometidas a criterios de filtrado y análisis automático en un intento de eliminar aquellas cuyo objetivo sea

la "manipulación" de los resultados del buscador. Así mismo, la presentación de resultados por parte de este tipo de buscadores está basado en la aplicación de algoritmos internos de medición de relevancia de las páginas incluidas con respecto a los términos de búsqueda empleados por los navegantes así como su importancia según criterios internos. Todo el proceso anterior no es automático. Generalmente se suelen disponer varias bases de datos con distintos niveles de actualización las cuales se van sustituyendo progresivamente evitando actualizaciones "masivas" de todos los datos indexados.[16]

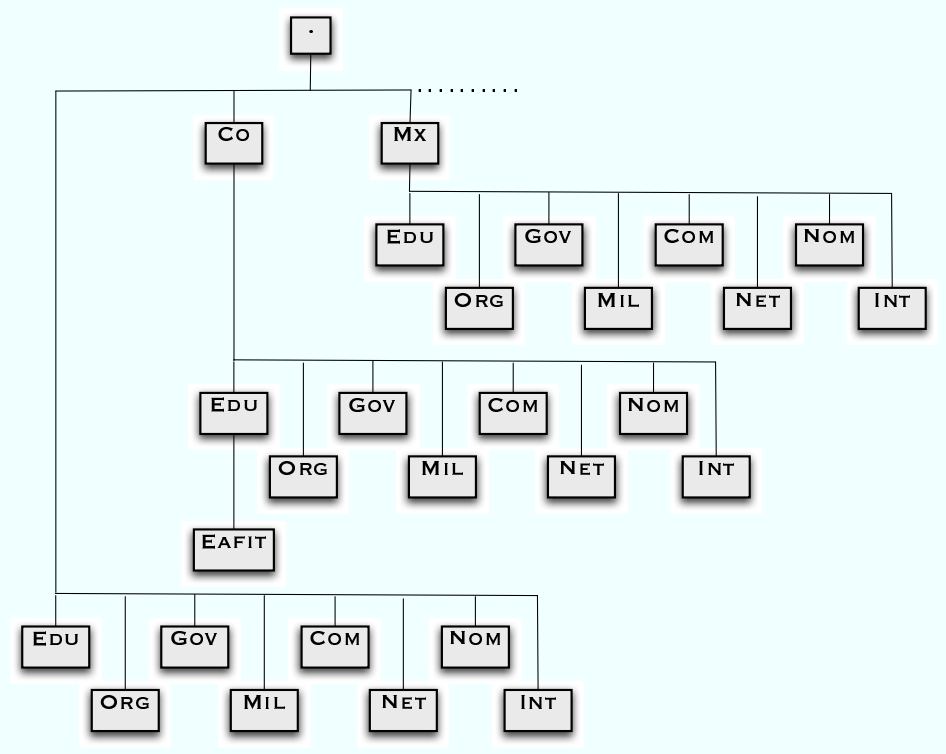

Figura 1.1: Representación del Sistema de Nombre de Dominio (*DNS*)

- **DNS**: (*Domain Name System*) es un conjunto de protocolos y servicios (base de datos distribuida) que permite a los usuarios utilizar nombres en vez de tener que recordar direcciones IP numéricas. Ésta es ciertamente la función más conocida de los protocolos *DNS*: la asignación de nombres a direcciones IP. [26]

1.2. BREVE EXPLICACIÓN DE LA TEORÍA DE GRAFOS

Los grafos son artefactos matemáticos que permiten expresar de una forma visual muy sencilla y efectiva las relaciones que se dan entre elementos de muy diversa índole. Un grafo simple está formado por dos conjuntos:

- Un conjunto V de puntos llamados vértices o nodos.
- Un conjunto de pares de vértices que se llaman aristas o arcos y que indican qué nodos están relacionados.

De una manera más informal podemos decir que un grafo es un conjunto de nodos con enlaces entre ellos, denominados aristas o arcos. En un grafo simple sólo hay un arco entre dos nodos . Si hay más de un arco hablamos de un multigrafo. Si los arcos se pueden recorrer en una en una dirección concreta pero no en la contraria lo llamamos grafo dirigido o dígrafo y los

arcos son entonces aristas. Si los arcos salen y llegan al mismo punto formando un bucle el grafo resultante se llama pseudografo. [11]

1.2.1. Grafos Dirigidos

En los problemas originados en ciencias de la computación, matemáticas, ingeniería y muchas otras disciplinas, a menudo es necesario representar relaciones arbitrarias entre objetos de datos. En este caso particular se representará la Web Colombiana.

Un grafo dirigido **G** consiste en un conjunto de vértices V y un conjunto de arcos A. Los vértices se denominan también nodos o puntos; los arcos pueden llamarse arcos dirigidos o líneas dirigidas. Un arco es un par ordenado de vértices (v,w); donde v es la cola y w la cabeza. El arco (v,w) se expresa a menudo como $v \rightarrow w$ y se representa como lo sugiere la figura 1.2.

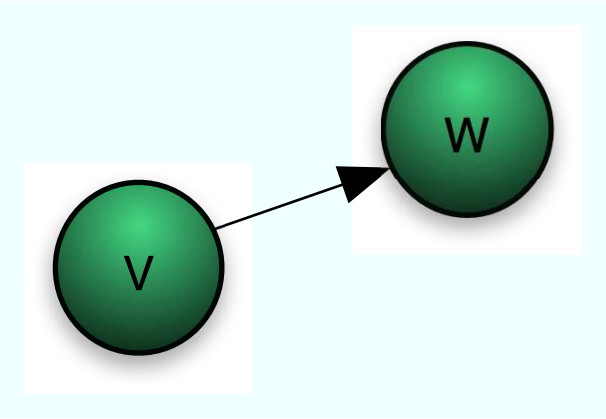

Figura 1.2: Grafo Dirigido Simple

Obsérvese que la "punta de la flecha" está en el vértice llamado *cabeza*, y la "cola de la flecha", en el vértice llamado *cola*. Se dice que el arco $v \rightarrow w$ va de v a w y que w es adyacente a v.

Los vértices de un grafo dirigido pueden usarse para representar objetos, y los arcos, relaciones entre los objetos. Tomando como ejemplo este proyecto de grado, los vértices serían páginas o sitios y los arcos serían los enlaces o vínculos entre dichas páginas o sitios. Como otro ejemplo ilustrativo, se podría tomar ciudades como vértices y los vuelos aéreos entre una ciudad y otra como arcos.

Un camino en un grafo dirigido es una secuencia de vértices v_1, v_2,. . . , v_n, tal que $v_1 \rightarrow v_2$, $v_2 \rightarrow v_3$,. . . , $v_{n-1} \rightarrow v_n$ son arcos. Este camino va del vértice v_1 al vértice v_n, pasa por los vértices v_2, v_3,. . . , v_{n-1}, terminando en el vértice v_n. La longitud de un camino es el número de arcos en ese camino, en este caso *n-1*. Como caso especial, un vértice sencillo v, por sí mismo denota un camino de longitud cero de v a v. Este sería el caso particular de las páginas llamadas "islas" o "islands", las cuales no están conectadas con ninguna otra, pero que pertenece al dominio colombiano. [1, 20]

Un camino es simple si todos sus vértices, excepto tal vez el primero y el último, son distintos. Un ciclo simple es un camino simple de longitud por lo menos uno, que empieza y termina en el mismo vértice.

En muchas aplicaciones es útil asociar información a los vértices y arcos de un grafo dirigido. Para este propósito es posible usar un grafo dirigido

etiquetado, en el cual cada arco, vértice o ambos pueden tener una etiqueta asociada. Una etiqueta puede ser un nombre *(URL)*, un costo o un valor de cualquier tipo de datos dado.

Para representar un grafo dirigido se pueden emplear varias estructuras de datos, como matrices de adyacencia, listas de adyacencia, o se podría definir un *TDA* (Tipo de Dato Abstracto); la selección apropiada depende de las operaciones que se aplicarán a los vértices y a los arcos del grafo. [1]

Para el caso de este trabajo de grado, los enlaces entre los vértices del grafo o páginas se almacenarán en una tabla de la base de datos, donde se hallarán los identificadores de la página origen y de la página destino, luego de analizar cada uno de los vínculos que tenga cada página visitada, de tal forma que sea útil al momento de implementar la visualización del grafo con sus respectivos enlaces o arcos.

1.2.2. Recorridos en grafos dirigidos

Para resolver con eficiencia muchos problemas relacionados con grafos dirigidos, es necesario visitar los vértices y los arcos de manera sistemática. La búsqueda en profundidad[1], que es una generalización del recorrido en orden previo de un árbol, es una técnica importante para lograrlo, y sirvió de estructura para construir un algoritmo eficiente, por ejemplo al momento de clasificar los *hosts* hallados en este estudio, de acuerdo a la *Teoría de la Conectividad de la Web*, sobre la cual se profundizará más adelante.

Siguiendo con el tema del algoritmo de búsqueda en profundidad, supóngase que se tiene un grafo dirigido **G** en el cual todos los vértices están en un principio marcados como no visitados. La búsqueda en profundidad trabaja seleccionando un vértice v de **G** como vértice de partida; v se marca como visitado. Después se recorre cada vértice no visitado adyacente a **v**, aplicándole también la búsqueda en profundidad de manera recursiva. Una vez que se han visitado todos los vértices que se pueden alcanzar desde v, la búsqueda de v está completa. Si algunos vértices quedan sin visitar, se selecciona alguno de ellos como nuevo vértice de partida, y se repite este proceso hasta que todos los vértices de **G** se hayan visitado.

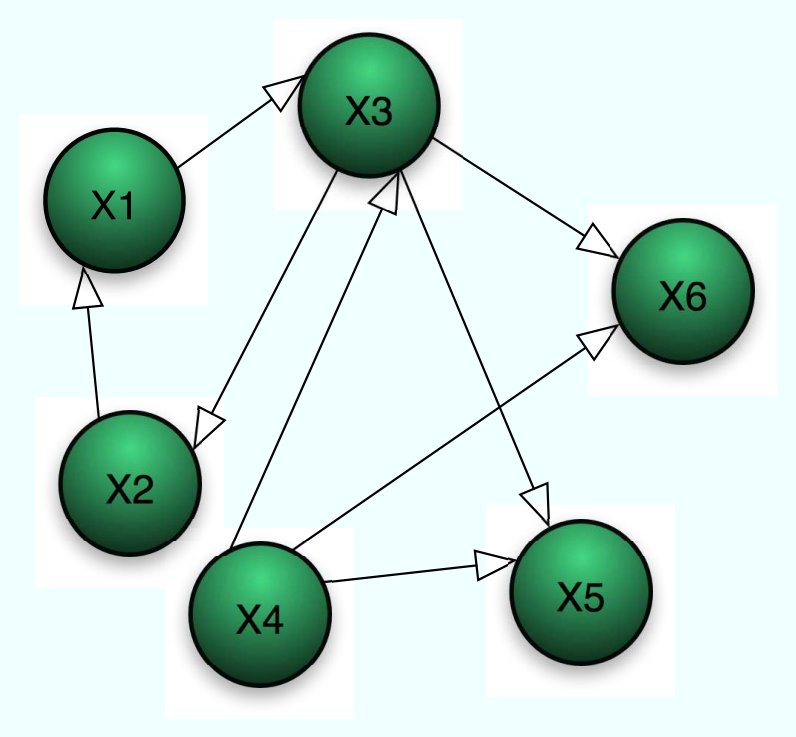

Figura 1.3: Representación Clásica de un Grafo Dirigido

Esta técnica se conoce como búsqueda en profundidad porque continúa buscando hacia adelante (más profunda) mientras sea posible. Por ejemplo, supongase que x es el arco visitado más recientemente. La búsqueda en profundidad selecciona algún arco no explorado $x \rightarrow y$ que parta de x. Si se ha visitado y, el procedimiento intenta continuar por otro arco que no se haya explorado y que parta de x. Si y no se ha visitado, entonces el procedimiento marca y como visitado e inicia una nueva búsqueda a partir de y. Después de completar la búsqueda de todos los caminos que parten de y, la búsqueda regresa a x, el vértice desde el cuál se visitó y por primera vez. Se continúa el proceso de selección de arcos sin explorar que parten de x hasta que todos los arcos adyacentes a x hayan sido explorados.

1.2.3. Clasificación topológica

La clasificación topológica es un proceso de asignación de un orden lineal a los vértices de un grafo dirigido acíclico tal que si existe un arco del vértice i al vértice j, i aparece antes que j en el ordenamiento lineal.

Componentes Fuertes

Existe una propiedad de los grafos dirigidos que es un concepto clave en este estudio. Un grafo dirigido es llamado fuertemente conexo (de ahora en adelante se llamarán *CFC*) si para cada par de vértices u y v existe un camino

de u hacia v y un camino de v hacia u, es decir, que de un nodo cualquiera se puede llegar a otro nodo cualquiera por medio de aristas dirigidas. Por ejemplo en la figura 1.3 el conjunto de vértices compuesto por X_1, X_2, X_3 forman un conjunto de *componentes fuertemente conexos*.

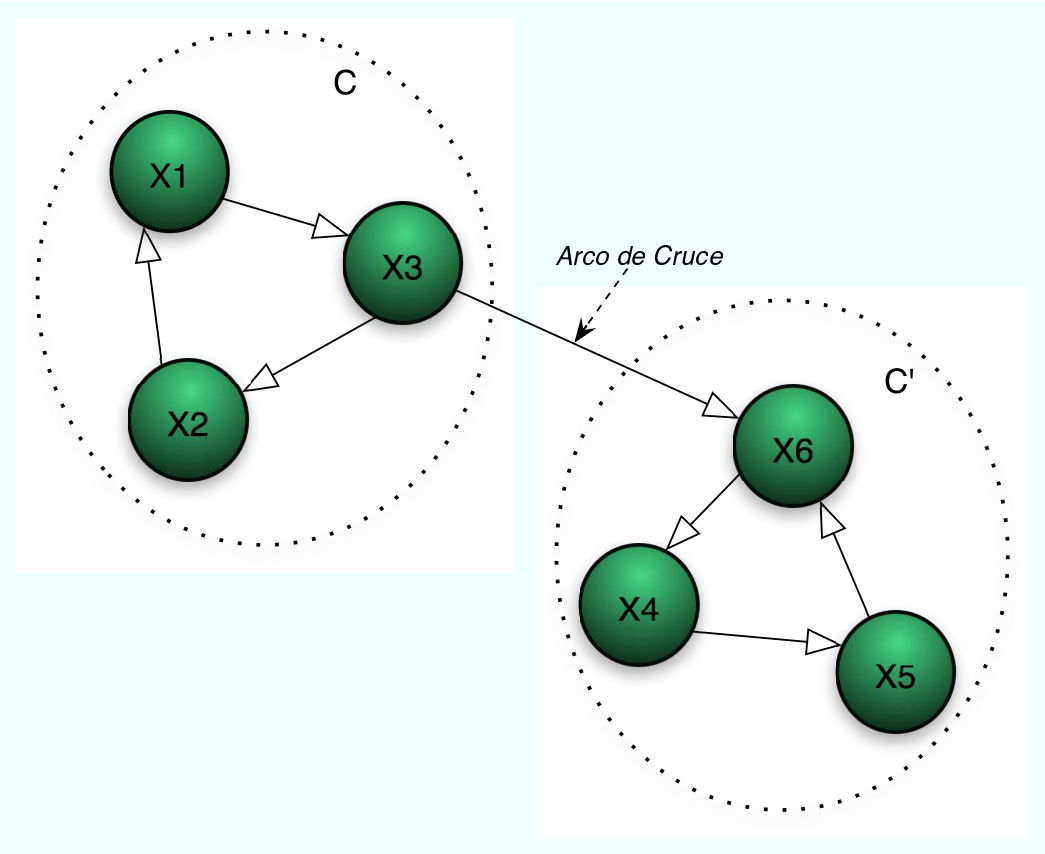

Figura 1.4: Ejemplo de GDA de Componentes Fuertemente Conexos

Sea **G = (V,A)** un grafo dirigido; se puede dividir **V** en clases de equivalencia V_i, $1 <= i <= r$ tales que los vértices v y w son equivalentes si y solo si, existe un camino de v a w y otro de w a v. Sea A_i, $1 <= i <= r$, el conjunto de arcos con cabeza y cola en V_i. Los grafos $G_i = (V_i, A_i)$ se

denominan CFC (o sólo componentes fuertes) de **G**. Un grafo dirigido con sólo un CFC, se dice que está fuertemente conexo.

Todo vértice de un grafo dirigido **G** está en algún CFC, pero ciertos arcos pueden no estarlo. Tales arcos, llamados arcos de cruce de componentes van de un vértice de componente a un vértice de otro. Se pueden representar las interconexiones entre los componentes construyendo un grafo reducido de **G**, cuyos vértices son los CFC de **G**. Así como lo muestra la figura 1.4.

Tomando dos componentes de este tipo de grafo como **C** y **C'**, si existe un arco en **G** que vaya de algún vértice del componente **C** a algún otro del componente **C'**, se dice que hay un arco de cruce del vértice **C** al vértice **C'**. El grafo reducido siempre es un *Grafo Dirigido Acíclico o GDA*, porque si existiera algún ciclo serían en realidad un solo componente fuerte, lo cual significaría que no se calcularon en forma adecuada los CFC. [1]

Existen diferentes tipos de algoritmos basados en la búsqueda en profundidad que sirven para encontrar todos los CFC de un grafo dirigido. Las cualidades de este estudio hacen posible utilizar una variante de dichos algoritmos con el fin de reducir el tiempo de búsqueda y la utilización de recursos informáticos. Se toma cualquier vértice v no analizado, se marca como analizado y se realiza una búsqueda en profundidad hacia adelante guardando todos los vértices que tengan un camino de v a ellos, haciendo la misma búsqueda de forma recursiva para cada uno de éstos y pasando el resultado a su anterior, al finalizar dicha búsqueda, v y todos los vértices a los cuales apunta ya cuentan con la lista de sus siguientes, por lo que no es necesa-

rio realizar la búsqueda cuando se vaya a analizar alguno de ellos. Luego se realiza una búsqueda en profundidad hacia atrás en v, guardando todos los vértices que tengan un camino hacia v, realizando la misma búsqueda de forma recursiva para cada uno de éstos y pasando el resultado a su siguiente, al finalizar esta búsqueda, v y todos los vértices que lo apunta ya cuentan con la lista de sus anteriores, lo que también reduce a la mitad el análisis de cada uno de dichos vértices. Luego de encontrar el conjunto de todos los vértices a los cuales se puede llegar desde v, y el conjunto de los vértices de los cuales se puede llegar a v, se obtiene un CFC de la intersección de dichos conjuntos. Este procedimiento se realiza hasta que todos los vértices sean analizados. Además, si al momento de hacer las búsquedas en profundidad hacia adelante o hacia atrás se encuentra un vértice terminal, es decir, que no cuenta con arcos de entrada o con arcos de salida, se marca como analizado para no perder tiempo en un análisis posterior.

Recopilando las definiciones más importantes de la Internet se puede llegar a una que ayude al lector a comprender el objetivo de esta investigación: La Internet es una malla inmensa de sitios y páginas que se encuentran relacionadas por vínculos o links que permiten al "navegante" ir de un lado a otro sin una estructura visible. Dicha malla se llamará de aquí en adelante grafo, las páginas son los nodos que comprenden ese grafo, y los vínculos o links serán las aristas o arcos que permiten una comunicación dirigida entre un par de nodos cualquiera. Esta adecuación a la teoría de grafos permite al lector tener una imagen más técnica y clara de la Internet.

1.3. LA TEORÍA DE LA CONECTIVIDAD DE LA WEB

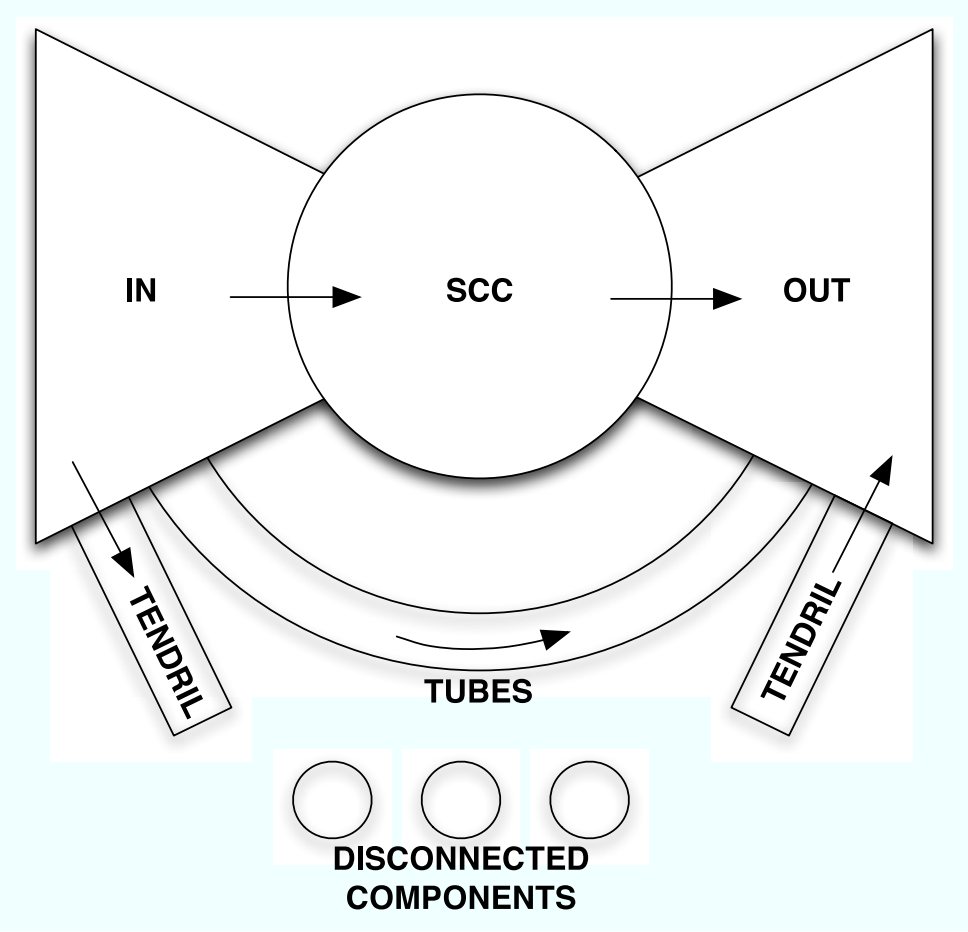

Figura 1.5: Resultado del estudio de la Web como un grafo dirigido [2]

En 1999, los centros estadounidenses de investigación de IBM y Compaq unieron fuerzas para realizar un estudio a gran escala de la estructura de la Internet, utilizando una aplicación desarrollada por la compañía Altavista

para recorrer la Web página por página llamada *Web Crawl*. Esta aplicación arrojó datos muy importantes sobre la conectividad de la Web y les permitió a la terna de compañías involucradas observar diferentes características que posee la Web desde el punto de vista de la conectividad. [2]

En dicha investigación se tomó el concepto de CFC aplicado a las páginas y enlaces, con el fin de agrupar los diferentes sitios en una categoría principal y otras derivadas de ella, pero que representa una forma más general de ver la Web y permite buscar una figura a nivel macro que se acople a dichas categorías. La clasificación resultante es la siguiente:

- **SCC**: (*Strongly Connected Component*) Este es el grupo de sitios que conforman el CFC de la Web.
- **IN**: Grupo de sitios que pueden acceder a *SCC* pero no pueden ser accedidos desde *SCC*.
- **OUT**: Grupo de sitios que son accedidos desde *SCC* pero que no pueden acceder a *SCC*.
- Existen varios grupos de sitios que no están en *SCC*, es decir, que no hacen parte del CFC, pero que tienen cierta relación con *IN* o con *OUT*:
    - **TENDRILS**: Grupo de sitios que son accedidos desde *IN* pero no están en *SCC*, y de sitios que acceden a *OUT* pero no están en *SCC*.

- **TUBES**: Grupo de sitios que son accedidos desde *IN* y acceden a *OUT* pero no están en *SCC*.

- **DISCONNECTED COMPONENTS**: Grupo de sitios desconectados de todos los demás.

Esta representación gráfica de la topología de la Web sirvió como base para que varios países se pusieran en la tarea de hacer un estudio similar en su grupo de sitios denotado por el dominio de primer nivel territorial, entre estos Inglaterra, Brasil y Chile, jugando éste último un papel muy importante en la historia de las investigaciones de la Web ya que modificó la *Teoría de la Conectividad de IBM–COMPAQ–ALTAVISTA* de 1999.

1.3.1. Modificaciones a la teoría de Conectividad de la Web

En el año 2000, el *Centro de Investigaciones de la Web de la Universidad de Chile* empezó una serie de estudios anuales que hasta la fecha han mostrado la evolución de su Web, dejando bien claro que la Web es altamente cambiante y dando cifras claves que permiten observar esos cambios. Pero la importancia de estos estudios es que no están basados en la Teoría de la Conectividad de IBM–COMPAQ–ALTAVISTA de 1999, sino en una modificación hecha por el jefe del Centro de Investigaciones de la Web y Director del Departamento de Ciencias de la Computación de la Universidad de Chile

en ese entonces. Esta modificación permitió segmentar aún más la estructura a la que la Internet parece acoplarse a la perfección.

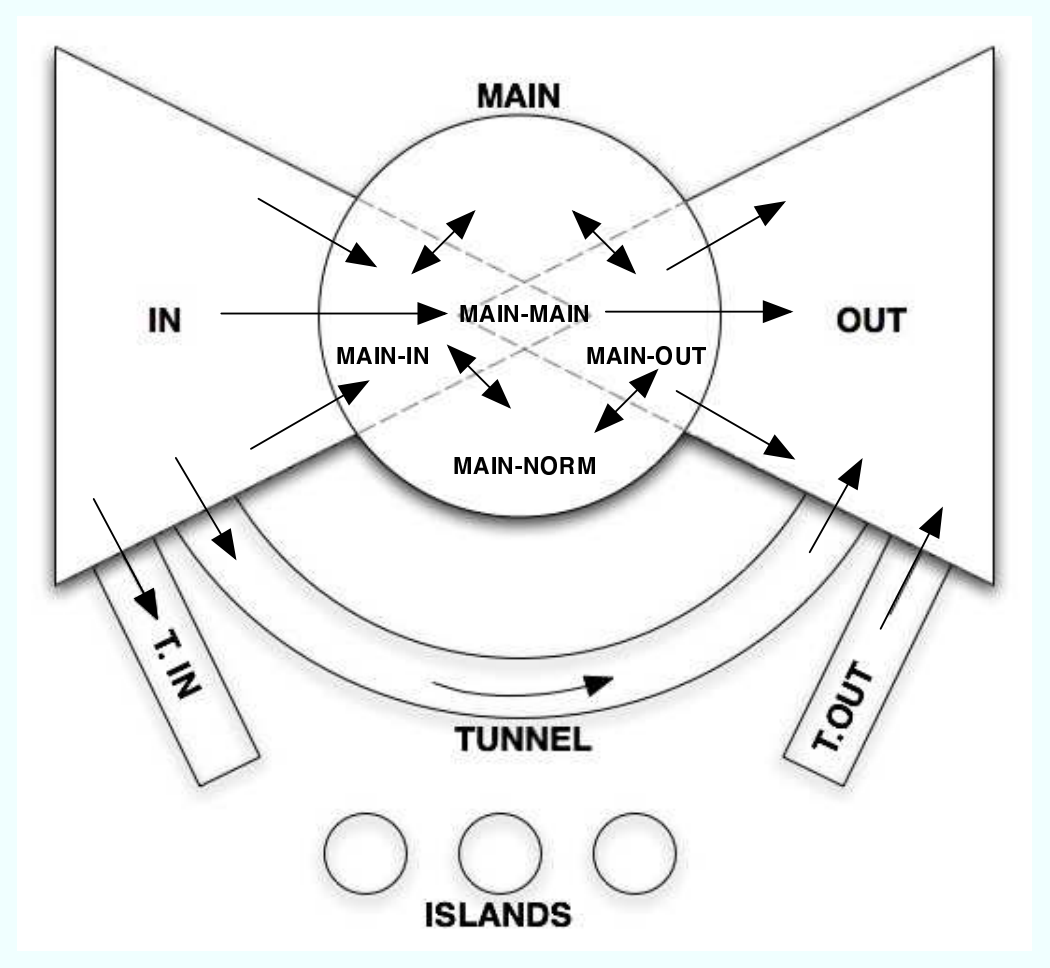

Figura 1.6: Subdivisiones adicionadas a la estructura original [20]

En el fondo la teoría de la conectividad no cambia demasiado, sólo se hacen una serie de subdivisiones que permiten entrar más en detalle en los estudios de la Web, dando mayor claridad en su evolución y en la manera de interpretar las cifras arrojadas por dichos estudios, estas subdivisiones son:

- **MAIN**: Este es el grupo de sitios que conforman el CFC de la Web, aquellos sitios que desde cualquiera de ellos se pueden llegar a los demás por medio de vínculos. Que a su vez se divide en:

  - **MAIN–MAIN**: Subgrupo de sitios que pueden ser accedidos directamente desde los componentes de *IN* y pueden acceder directamente a los componentes de *OUT*.
  - **MAIN–IN**: Subgrupo de sitios que pueden accedidos directamente desde los componentes de *IN*, pero no están en *MAIN–MAIN*.
  - **MAIN–OUT**: Subgrupo de sitios que pueden acceder directamente a los componentes de *OUT* pero no están en *MAIN–MAIN*.
  - **MAIN–NORM**: Subgrupo de sitios que no pertenecen a ninguno de los subcomponentes definidos anteriormente.

- **IN**: Grupo de sitios que pueden acceder a *MAIN* pero no pueden ser accedidos desde *MAIN*.

- **OUT**: Grupo de sitios que son accedidos desde *MAIN* pero que no pueden acceder a *MAIN*.

- Existen varios grupos de sitios que no están en *MAIN*, es decir, que no hacen parte del CFC, pero que tienen cierta relación con *IN* o con *OUT*:

  - **TENTACLE IN**: (*T.IN*) Grupo de sitios que son accedidos desde *IN* pero no están en *MAIN*.

  - **TENTACLE OUT**: ($T.OUT$) Grupo de sitios que acceden a OUT pero no están en *MAIN*.
  - **TUNNEL**: Grupo de sitios que son accedidos desde IN y acceden a OUT pero no están en *MAIN*.

- **ISLANDS**: Grupo de sitios desconectados de todos los demás.

Si se da el caso de encontrar una estructura compuesta por muchos posibles MAIN, es decir, CFC's de sitios, que reuniéndolos a todos no forman un CFC como tal, entonces se debe evaluar varios casos antes de determinar un *MAIN principal*:

- Se puede presentar un CFC que sea muy grande y que a su vez sea el más relacionado con los CFC que le circundan. Este sería el MAIN principal de este estudio.
- Si se encuentra un CFC muy grande, pero no es el más relacionado se debe recorrer cada uno de los CFC restantes que al menos sean apuntados por un CFC y apunten a otro CFC, calculando la diferencia entre la suma de los tamaños de los CFC que lo apuntan y la suma de los tamaños de los CFC a los cuales apunta, de tal manera que luego de analizar todos los CFC, se elige como MAIN principal el de la menor diferencia, buscando un equilibrio relativo.

Luego de realizar varias pruebas y cálculos con ejemplos hipotéticos, se llegó a la conclusión de que se elegirá como MAIN principal aquel que tenga

el mayor tamaño, siempre y cuando éste CFC sea apuntado y apunte a su vez al menos a un CFC en cada caso.

Capítulo 2

Metodologías y Tecnologías utilizadas para el desarrollo del sistema

2.1. HERRAMIENTAS DE DESARROLLO

La parte práctica de este proyecto de grado se realizó bajo estándares de desarrollo orientados al lenguaje de programación Java. Utilizando paradigmas de programación orientada a objetos que permiten la persistencia de los objetos contenidos en la base de datos.

2.1.1. Java como lenguaje de Programación

Se tomó la decisión de realizar el desarrollo práctico en dicho lenguaje ya que ofrece una base sólida para la manipulación de objetos, aplicación de patrones, conexiones a bases de datos que permiten multiples accesos en un mismo período de tiempo, tecnologías altamente estables que ofrece el acceso a recursos en línea, herramientas de manipulación de expresiones regulares y soporte para gráficos en tres dimensiones gracias a las bibliotecas de funciones de Java3D.

Además de las utilidades nativas del Java, se hizo uso de la tecnología *OSCaché*, la cuál ofrece una alternativa para guardar en memoria objetos Java recuperados de la base de datos y solicitados a manera de consulta, ayudando a reducir el tiempo de acceso a la base de datos y aumentando la rapidez en la ejecución de la aplicación.

El *OSCaché* ofrece entre otras ventajas, rápido acceso a memoria, persistencia en disco, excelente rendimiento y un sistema flexible de cacheo.[7]

Para la conexión a la base de datos se utilizó el pool de conexiones *Jakarta Commons*, el cual permite manejar una gran cantidad de conexiones simultáneas de forma rápida, facilitando la creación y mantenimiento de componentes reutilizables de Java.[19]

2.1.2. Eclipse

Eclipse es un entorno integrado de desarrollo multiplataforma, construida especialmente para el desarrollo en el lenguaje de programación Java por una comunidad de proveedores de herramientas y soluciones. Operando bajo el paradigma de "código abierto", con una licencia pública que provee derechos libres de redistribución. Esta plataforma provee a los desarrolladores una herramienta con bastante flexibilidad y control sobre su tecnología de software.

2.1.3. Estandares de programación J2EE

J2EE son las siglas de *Java 2 Enterprise Edition* que es la edición empresarial del paquete Java creada y distribuida por *Sun Microsystems*. Comprenden un conjunto de especificaciones y funcionalidades orientadas al desarrollo de aplicaciones empresariales. [27]

Algunas de las funcionalidades aplicadas en este proyecto son:

- Acceso a base de datos (**JDBC**): *JDBC* es el acrónimo de *Java Database Connectivity*, una *Interface de Programación de Aplicaciones* ***API*** que permite la ejecución de operaciones sobre bases de datos desde el lenguaje de programación Java, independientemente del sistema de operación donde se ejecute o de la base de datos a la cual se accede,

utilizando el dialecto SQL (*Structured Query Language*) del modelo de base de datos que se emplee.

- Utilización de directorios distribuidos (**JNDI**): *JNDI* son las siglas de *Java Naming and Directory Interface*. Es una API para servicios de directorio. Permite a los clientes descubrir y ubicar datos y objetos utilizando un nombre y, como toda API Java que interactúa con sistemas host, es independiente de la implementación de fondo.[25]

- Uso de Beans: Un Bean es un componente de software que tiene la particularidad de ser reutilizable y así evitar la tediosa tarea de programar los distintos componentes uno a uno. Se puede decir que existen con la finalidad de ahorrar tiempo de desarrollo y es el caso de la mayoría de componentes que manejan los editores visuales más comunes.

- Se divide cada una de las aplicaciones en capas de acuerdo con el patrón (**MVC**): *Modelo Vista Controlador*. El cual es un patrón de arquitectura de software que separa los datos de una aplicación, la interfaz de usuario, y la lógica de control en tres componentes distintos. El patrón MVC se ve frecuentemente en aplicaciones Web, donde la vista es el conjunto de formularios y páginas, el control es el objeto encargado de escuchar y atender las peticiones de los usuarios, proporcionando datos dinámicos a las páginas y el modelo contiene las clases representativas de la aplicación.

2.2. SISTEMA DE GESTIÓN DE LA BASE DE DATOS

En este proyecto de grado se empleó como herramienta para la administración de la base de datos, el motor MySql en su versión 4.1. Se utilizaron, además, como herramientas para interactuar con la base de datos, en el Sistema Operativo Windows, la consola nativa de MySQL y el programa "MySql Control Center 0.94-beta" y, en el sistema operativo Mac OS X, la herramienta "Aqua Data Studio". Estas herramientas permiten ejecutar scripts de SQL, hacer importaciones y exportaciones de datos, creación de tablas, entre otras funcionalidades que fueron necesarias en todo el proceso.

2.2.1. ¿Por qué MySql?

Una de las principales razones para la elección de este sistema de gestión de bases de datos fue su naturaleza de "Software Libre", que permite su uso y modificación de forma gratuita [1]. Esto da vía libre para que no haya problemas de licenciamiento para aquellas personas que deseen profundizar aún más en este proyecto.

[1]Estas modificaciones son referentes a su código fuente, sin embargo tienen ciertas limitaciones ya que no pueden ser utilizadas o incluídas en software con fines comerciales. (MySQL usa el *GPL* (*GNU* Licencia Publica General) para definir qué se puede y qué no se puede hacer con el software en diferentes situaciones)

Otra razón igual o más importante es que MySql, como gestor de *bases de datos relacionales* , es rápido, confiable, robusto, soporta igualmente altos volúmenes de datos, tiene un gran conjunto de funciones desarrolladas para su usuario y además su conectividad, velocidad y seguridad lo hace altamente conveniente para acceder a bases de datos en la Internet. [3]

2.2.2. Características Particulares de MySql

Este SGBD (*Sistema de Gestión de Bases de Datos*) tiene compatibilidad con los Sistemas Operativos, aplicaciones y lenguajes de programación utilizados en el proyecto, que son Mac OS X y Microsoft Windows como sistemas operativos, además de Java y Java 3D como lenguajes de programación de las aplicaciones construídas que emplearon y ejecutaron sentencias SQL para el acceso y escritura en la base de datos.

Pensando en una posible división del procesamiento de los datos para este proyecto particular, se sabe que MySql puede soportar varios procesadores, porque manejando muchas conexiones al mismo tiempo, esta herramienta adopta las propiedades multihilos, de tal forma que cada conexión tenga su propio hilo, haciendo que ninguno de ellos deba esperar por la terminación de otro (a menos que un hilo esté modificando una tabla y otro hilo quiera accederla). Esto es muy útil cuando hay grandes procesos que pueden producir un bloqueo en otros que resulten más cortos o sencillos y que necesiten una prioridad al momento de su ejecución.

Otra de las características definitivas de MySql por la que fue utilizado en este proyecto es la funcionalidad que ofrece para el manejo de la *integridad referencial* de la base de datos. Esta particularidad, si es bien utilizada, permite mantener consistencia en los datos y con total tranquilidad se pueden crear, actualizar o borrar registros de la misma.

Podemos decir de manera simple que integridad referencial significa que cuando un registro en una tabla haga referencia a un registro en otra tabla, éste último registro debe existir. [23]

Dentro de las funciones indispensables en bases de datos de tipo transaccional está principalmente el manejo de *Claves Foráneas* o *Foreign Keys*, el manejo de *Índices* y también algo muy importante: el borrado y actualización en cascada, teniendo previamente en cuenta los tipos de tablas que soporta la versión de Mysql, ya que no en todas soporta esta última funcionalidad.

2.2.3. Funcionalidades empleadas en el proyecto

Las funciones ya comentadas que ofrece la versión 4.1 de MySql y que fueron empleadas e implementadas en la creación y manipulación de la base de datos, como fueron los tipos de tablas transaccionales, índices, claves foráneas, entre otras, se describen a continuación:

- **Tipos de Tablas**

Las tablas de la base de datos fueron creadas del tipo InnoDB [2]. La importancia de haber utilizado esta funcionalidad se evidencia claramente al querer por ejemplo borrar o actualizar registros de la base de datos. En este caso se mantiene la integridad referencial de la misma, gracias a el borrado y actualización en cascada que permite que al borrar o actualizar un registro en una tabla, borre o actualice automáticamente esos mismos registros o campos que estén en otras tablas relacionadas con la tabla inicial.

- **Claves Foráneas (*Foreign Keys*)**

  Para lograr el efecto cascada ya mencionado, es necesario que las tablas estén relacionadas a través de campos clave, que son definidos explícitamente en sus restricciones o *constraints* y se llaman claves foráneas. Éstas referencian la clave primaria de una tabla "padre" y son declaradas en el script de creación de cada una de las tablas. De manera resumida, una clave foránea es simplemente un campo en una tabla que se corresponde con la clave primaria de otra tabla.

  En el proyecto fue necesario definir muchas de estas restricciones, para controlar la correspondencia entre campos de diferentes tablas. Por ejemplo para relacionar las tablas de páginas (*tblpagina*) y saber a que host pertenece, existe el atributo *idhost* en la tabla *tblpagina* que es clave foránea, ya que ese mismo campo es clave primaria en la tabla de hosts (*tblhost*). De esta manera, en una misma consulta, se puede

[2] MySQL 4.1 soporta cinco tipos de tablas: MyISAM, ISAM, HEAP, BDB (Base de datos Berkeley), e InnoDB. BDB e InnoDB son ambas tipos de tablas transaccionales, es decir, que soportan el manejo del commit y rollback desde la aplicación.

traer información de una página o url y además adicionarle información del host al que pertenece dicha página, a pesar de que la información se encuentra en tablas separadas. A parte de éste, existen numerosos ejemplos más como el de los dominios y subdominios que pertenecen a un host, o también para saber los enlaces entre las páginas, fue necesario crear una tabla de enlaces (*tblenlace*) que se relaciona con la tabla *tblpagina* y muestra entre qué páginas hay un enlace directo a través de links en la Web colombiana, entre otros ejemplos.

- **Utilización de índices**

  Para que las funcionalidades de las claves foráneas operen correctamente, se debe definir por cada una de ellas un índice o *index* en el momento de crear la tabla, para lograr una mayor rapidez al relacionar tablas que tengan campos en común y en los cuales sea necesario implementar una integridad referencial. Las claves foráneas como otra de las características que ofrece esta versión de MySql, combinado con el manejo de índices, son importantes también a la hora de insertar registros y de realizar consultas que extraigan información de más de una tabla en un solo registro. Esto se realiza por medio de sentencias SQL que especifiquen la forma de relacionar o enlazar los registros de tablas con sus campos claves o atributos "compartidos", de tal manera que al momento de insertar o consultar registros se hagan previamente ciertas validaciones para que la información retornada sea confiable, claro está, contando con una buena definición de dichas sentencias que implica un conocimiento profundo de las tablas y sus relaciones en la base de datos, de acuerdo a el *Modelo Entidad-Relación* de la misma.

2.3. COMPONENTES PRINCIPALES DE LA APLICACIÓN

La aplicación está conformada por cuatro grandes procesos, cada uno de los cuales tiene acceso a la base de datos llamada Webcol. Dichos procesos son:

- Consultor
- Analizador
- Clasificador
- Visualizador

Reuniendo los dos primeros procesos, se forma el llamado *Crawler* que en resumen lo que realiza es obtener la porción de las páginas que componen la Web Colombiana, a través de la descomposición de el código HTML de las páginas de resultados del buscador empleado para el estudio y luego de esto visitar una a una las URL's halladas en la consulta inicial, guardando además los nuevos links que se generen de la descomposición del código fuente de las pgínas visitadas.

De una manera más detallada, se explica a continuación en qué consiste tanto el Consultor como el Analizador desarrollados en este proyecto:

2.3.1. Consultor

Este proceso se encarga de acceder a uno o más buscadores Web (tipo Google, Yahoo, Altavista, entre otros), consultando los sitios de dominio de primer nivel territorial colombiano (.co), y sus combinaciones con los demás dominios de primer nivel (.edu.co, .gov.co, .mil.co, .net.co, .com.co, .org.co, .gob.co, .int.co, .nom.co). Además, genera consultas automáticas haciendo una combinación entre los dominios de las páginas encontradas (catalogadas como "palabras clave") y los dominios de primer nivel antes mencionados.

Lo anterior se realiza con el principal objetivo de tomar los resultados del buscador y, por medio de técnicas de expresiones regulares, buscar direcciones Web de páginas o sitios de la Web colombiana, las cuales se guardan en una base de datos diseñada especialmente para almacenar dicha información y que luego el Analizador sea quien visite las páginas registradas por el proceso de consulta.

Su funcionamiento es sencillo, simula ser un navegador o *browser* para poder acceder a los buscadores con total libertad y consulta utilizando los parámetros del *request* de la página buscadora, luego toma el código HTML con los resultados y aplica técnicas de expresiones regulares para obtener los vínculos a las direcciones de la Web colombiana. Es necesario poseer previamente un criterio de definición de vínculos válidos, ya que cada buscador entrega los resultados mezclados con vínculos inservibles o irrelevantes como publicidad o direcciones hacia otros servicios del mismo buscador que deben

ser omitidos en este estudio.

Luego de obtener la lista de direcciones válidas del buscador, utiliza nuevamente técnicas de expresiones regulares para obtener el subdominio, dominio y extensión de cada una de las direcciones de la lista, además encola una nueva búsqueda relacionada con el dominio.

2.3.2. Analizador

Este proceso se encarga de tomar cada página no visitada contenida en la base de datos para obtener su código HTML y en éste buscar direcciones Web, tanto relativas como absolutas, para registrar en la base de datos las páginas con las que tiene relación por medio de vínculos o direcciones Web, además almacena su tamaño y fechas de creación y modificación de cada una de ellas.

Se puede dar el caso que el analizador encuentre la relación de una página registrada por el consultor, con otra que no se encuentre en la base de datos, y en este caso, el mismo analizador es el encargado de almacenar la nueva URL.

El principal objetivo de este análisis es obtener todos los enlaces que tienen las páginas con todas las demás, ya que este es el aspecto de mayor importancia para la posterior clasificación topológica de la Web colombiana.

2.3.3. Clasificador

Este proceso es el encargado de observar las características de conectividad de cada sitio y clasificarlo según la teoría de la conectividad de la Web [2, 20], por medio de consultas realizadas sobre la base de datos y la utilización de un algoritmo diseñado para tal fin.

Su comportamiento está totalmente ligado a la base de datos, ya que en gran parte sólo ejecuta sentencias SQL sobre los registros relacionados con los sitios o 'host', cada una de ellas buscando registros que cumplan con las características que poseen cada uno de los tipos de componentes de la Teoría de Conectividad de la Web tratada en el marco teórico de este proyecto de grado.

El único proceso complejo que se debió hacer es el diseño de una adaptación del algoritmo de búsqueda en profundidad para determinar el conjunto de los sitios fuertemente conexos o MAIN. Gracias a un análisis de las características de cada uno de los tipos de componentes de la Web se obtuvo un orden de clasificación que reduce el dominio de búsqueda antes de obtener el conjunto MAIN:

- Aplicación del algoritmo descrito anteriormente para detectar todos los posibles MAIN del grafo, es decir, todos sus CFC. Además, para encontrar los hosts terminales que pueden ser IN o TENTACLE OUT en caso que se hayan encontrado en la búsqueda en profundidad hacia

atrás, o los que pueden ser OUT o TENTACLE IN, en caso tal de que se hayan encontrado en la búsqueda en profundidad hacia adelante.

- Se toma cada CFC y se compara su tamaño para elegir como MAIN aquel que contenga mayor número de hosts.

- Para clasificar el IN se realiza una serie de consultas sobre la base de datos detectando los hosts que apuntan al conjunto de hosts que conforman el MAIN, guardando en cada iteración la distancia o las visitas que se deben hacer para llegar del host hasta el MAIN.

- Para el caso de la clasificación de los OUT se realiza de forma similar, pero hallando los hosts que son apuntados por los componentes del MAIN.

- El conjunto de hosts que representan el TUNNEL, son aquellos que no han sido clasificados y que, como la teoría lo dice, existe un camino desde el IN hasta el OUT sin relacionarse con el MAIN. Para poder clasificarlos primero se deben encontrar todos los host siguientes a IN y todos los host anteriores a OUT. La intersección de los conjuntos obtenidos se clasifica finalmente como TUNNEL. Además, se deben clasificar como TUNNEL los sitios que son siguientes o anteriores a los mismos TUNNEL.

- Los TENTACLE IN y OUT se clasifican haciendo un recorrido desde el IN hacia adelante y desde el OUT hacia atrás respectivamente, en dichos recorridos de los host no clasificados también se obtiene la distancia hasta los hosts del otro conjunto. Además, se deben clasificar

como TENTACLE IN los sitios que son anteriores a los mismos TENTACLE IN, y de igual forma, se deben clasificar como TENTACLE OUT los sitios que son siguientes a los mismos TENTACLE OUT.

- Las ISLANDS están compuestas por cada uno de los hosts que no han sido clasificados hasta el momento.
- Las subdivisiones de MAIN se realizan con consultas sencillas sobre la base de datos, pues dependen de la relación directa con los host que pertenecen a los conjuntos IN y OUT.

Luego de realizar esta clasificación es importante destacar que se encuentran CFC's cuyos componentes estén clasificados como ISLANDS, es conveniente realizar la clasificación tomando dicho CFC como un MAIN alterno o de una estructura desligada de la principal, además dicha información es de gran utilidad para el análisis del estado de madurez del la Web colombiana.

2.3.4. Visualizador

Este proceso es el encargado de mostrar los resultados del funcionamiento del sistema, permitiendo al usuario observar la Web como un todo, mostrando cercanía entre páginas bastante relacionadas. Todo construído en un universo virtual que permite al usuario moverse en diferentes direcciones y apreciar los enlaces como un puente de comunicación entre las páginas y sitios.

Con el fin de construir todo el espacio virtual, el visualizador busca cada

uno de los sitios en la base de datos y los ubica en rangos de coordenadas específicos dependiendo del tipo de componente y la distancia hasta el main, es decir, se toma un sitio de cada tipo de componente, lo ubica en el espacio virtual, consulta el sitio del mismo tipo que tenga la mayor cantidad de relaciones con él y lo ubica cerca al anterior. Al terminar de ubicar todos los sitios del mismo tipo de componente, pasa a otro tipo y modifica el rango de asignación de las coordenadas en el espacio virtual.

2.3.5. Base de datos WebCol

La información almacenada en las tablas de la base de datos que se tituló **WebCol** fue sometida a ciertos controles y correcciones, debido a que no existe un estándar único en el código fuente de las páginas de la Web colombiana y en ciertas ocasiones se encontraban estructuras que se salían de lo común o de los estándares tomados en cuenta (por ejemplo al momento de definir un link o hipervínculo dentro de una página o en los mismos parámetros de la dirección url, entre otras excepciones), lo cual hizo que la aplicación se comportara de manera diferente al momento de analizar sintácticamente las direcciones y las páginas a través del empleo de técnicas de expresiones regulares para obtener su extensión, vínculos, dominio, subdominio, etc. y por esta razón la información resultante no era correcta, y se almacenaban entonces datos "basura". Sin embargo, dichas excepciones sirvieron también como una oportunidad para retroalimentar la aplicación y redefinir las estructuras o la sintaxis base que tendría en cuenta la aplicación al momento

de navegar por la web colombiana e interactuar con el código fuente de las páginas que la compone.

Se realizó entonces una depuración contínua de la base de datos, auditando que se estuviera almacenando datos válidos y en caso de que no fuera así, se corregía la aplicación, luego se inicializaba nuevamente el estado de las URL's, y por último se continuaba con su ejecución, claro está, habiendo aplicado los nuevos cambios del analizador de tal manera que en algún momento futuro se volvieran a analizar estas páginas y finalmente se almacenara la información correcta.

En otras ocasiones fue necesario estrictamente borrar registros de la base de datos, porque había ciertos parámetros erróneos o sentencias inválidas dentro del código HTML de las páginas que de ninguna manera podían quedar válidos dentro de la base de datos, ya que poseían una sintaxis incorrecta o simplemente su información no se debía considerar válida dentro del conjunto de posibilidades con las cuales interactuaría la aplicación.

Esta base de datos creada para almacenar los registros extraídos por la aplicación desde la Web colombiana, consta de catorce tablas, de tipo *InnoDB*, con claves primarias, claves foráneas definidas en sus restricciones, índices y además con atributos que poseen tipos de datos soportados por MySql.

Estas tablas fueron creadas a través de código fuente en Java, que contenía el script de creación de las mismas. El modelo Entidad–Relación diseñado

para la base de datos y su diccionario de datos respectivo se puede observar en el manual del sistema, anexado dentro de la documentación del proyecto.

En un principio, el Consultor desarrollado, al ponerlo en estado de ejecución, combinaba automáticamente parámetros, generando todo un conjunto de búsquedas posibles dentro del dominio colombiano para llevarlas como al buscador empleado y ejecutarlas una a una para que éste retornara la mayor cantidad de resultados o de páginas posibles. Estas consultas pendientes por hacer se guardaban en una tabla temporal a la que el Consultor accedía y cada que ejecutaba una de ellas, actualizaba su estado para saber que ya la había realizado y continuar con las restantes. Había entonces una metodología de encolamiento implementada para garantizar que cada una de estas consultas se ejecutaran completamente.

Para esto, se partió de unas búsquedas predefinidas para cada uno de los dominios (*.edu, .gov, .net, .mil, etc.*) y según los resultados arrojados por estas consultas, se iban construyendo nuevas consultas. Por ejemplo en un principio en Google se ejecutó una de las sentencias predefinidas (*site:.edu.co*) y ésta retornó dentro de sus resultados el dominio *eafit*. Una de las búsquedas que el consultor automáticamente construye para una próxima consulta es "*site:.edu.co eafit*" y esta consulta a su vez retornó palabras claves adicionales que igualmente se combinaron con cada uno de los dominios para tratar de conseguir la mayor cantidad de páginas o resultados posibles, pero todo esto con un límite predefinido, ya que sino este proceso se convertiría en un ciclo infinito.

Igualmente, este constructor automático de consultas, fue sujeto a depuraciones constantes porque en un principio armaba consultas no válidas para el buscador y lógicamente no arrojaba ningún resultado en *Google*, lo cual hacía que la aplicación perdiera rendimiento, ocupando tiempo de procesamiento innecesariamente. Por esta razón, se monitoreaba periódicamente las consultas pendientes que almacenaba en dicha tabla temporal y cuando se encontraban consultas con algún tipo de anomalía, se redefinía el consultor con sus nuevas modificaciones, se borraban las consultas inválidas que habían en ese momento en la tabla temporal y se ponía en estado de ejecución nuevamente el consultor. Más detalles acerca de las cuatro aplicaciones desarrolladas en este proyecto se encuentran en el manual del sistema de este proyecto de grado.

2.4. EXPRESIONES REGULARES

Para interactuar con el código html de las páginas de resultados que retornaba el buscador, fue necesario emplear técnicas de expresiones regulares con el fin de extraer la información que es útil para este proyecto de grado, que en este caso serían cada una de las URL's resultantes de las búsquedas realizadas. Estas mismas técnicas se emplearon igualmente para trabajar con el código fuente – HTML o javascript – de cada una de las páginas resultantes al momento de visitarlas, ya que se hacía necesario extraer cierta información (por ejemplo su dominio, subdominio, extensión y otras URL's con las que esté enlazada) que de una u otra forma está enmarcada dentro de un lenguaje

finito o minigramática, que identifica a cada uno de los tipos de información que se pueden encontrar en el código fuente de las páginas o en su misma URL.

Teóricamente, una expresión regular (abreviada como regexp o regex), también llamada patrón, es una secuencia de caracteres que describe o define un conjunto de posibilidades que pueden adoptar ciertos símbolos o rangos de caracteres predefinidos a través de reglas de sintaxis, de tal forma que no se haga necesario listar cada una de esas posibilidades o formas que puede adquirir dentro de su conjunto. Esta técnica por ejemplo es muy utilizada por editores de texto y utilidades en general para buscar y manipular el contenido del cuerpo de los textos, basados en ciertos patrones. [28]

Las expresiones regulares utilizadas en este proyecto de grado, que hacen parte de una minigramática, representan una herramienta muy útil para determinar los componentes característicos de cada página, pero además existen un grupo de expresiones regulares más complejas llamadas *gramáticas completas*, que son las utilizadas para definir lenguajes de programación y para lograr que la máquina interprete sus instrucciones. Dentro de las utilidades de las expresiones regulares está por ejemplo la de distinguir las entradas válidas de las entradas erróneas dadas por los usuarios de una aplicación. [24]

2.4.1. Aplicaciones en el Consultor

Propiamente hablando del caso de este proyecto de grado, se tiene que la utilidad que brindan las expresiones regulares es al momento de manejar las URL's y el código fuente (HTML y javascript) de las páginas de resultados del buscador empleado, como de cada una de las páginas resultantes de la ejecución del consultor. Así entonces, fueron utilizadas sus funcionalidades tanto en el consultor como en el analizador.

En el consultor son empleadas para manipular la URL de *Google*, haciendo que la aplicación entendiera el estándar de su cadena e identificara en qué lugar de ella cambiar parámetros para realizar una búsqueda nueva, como por ejemplo cambiar los parámetros de paginación de los resultados de *Google*, o cambiando también las palabras clave para que el buscador arrojara nuevos resultados. Las expresiones regulares fueron utilizadas en el consultor también para analizar el código HTML de las páginas de resultados del buscador, para extraer de su código cada uno de sus vínculos o enlaces a otras páginas y almacenar estas nuevas URL's en una tabla de la base de datos, para su posterior visita. De esta manera se obtenían todas las páginas posibles de la Web colombiana, resultantes del análisis de cadenas, aplicando reglas de expresiones regulares para la identificación de la información válida para este proyecto de grado.

Por ejemplo si en *Google* se ejecuta la búsqueda con los parámetros **site:.edu.co eafit** debe traer todos los sitios que sean del dominio **.edu.co** y

que tengan la palabra clave "**eafit**". La URL que se obtiene al ejecutar esta consulta es la siguiente:

http://www.google.com.co/search?hl=es&q=site:.edu.co+eafit

Cuando el usuario se desplaza a la segunda página de resultados (agrupados de a 10 en este caso) la URL resultante es la siguiente:

http://www.google.com.co/search?**q=site:.edu.co+eafit**&**start=10**

Donde el parámetro "*start=10*" identifica la paginación de los resultados, es decir, que se está mostrando desde el resultado número 10 hacia delante. Esta cifra fue la que manipuló la aplicación, cambiando dicho parámetro para analizar cada una de las páginas de resultados.

El buscador arrojaba ciertos resultados pero omitía otros, y un ejemplo del mensaje que muestra en la última página de resultados es:
"Para mostrarle los resultados más pertinentes, omitimos ciertas entradas muy similares a los 647 que ya hemos mostrado. Si lo prefiere, puede repetir la búsqueda e incluir los resultados omitidos."
Si se da click en el vínculo subrayado, automáticamente el browser agrega otro parámetro al final de la URL, indicando el filtro y lo iguala a cero, denotando que el conjunto de resultados no tiene filtro de ningún tipo, por lo tanto el número de resultados aumenta en alguna proporción. La URL quedó finalmente de la siguiente forma:

http://www.google.com.co/search?q=site:.edu.co+eafit&start=10&**filter=0**

Éste último parámetro se añadió siempre en las URL's para Google, de tal forma que no excluyera ningún tipo de resultado al momento de ejecutar una consulta.

2.4.2. Aplicaciones en el Analizador

Luego de un análisis detallado de la URL manejada por el buscador, se procedió a estudiar también el código HTML de dichas páginas de resultados, para extraer todas las URL's retornadas luego de la ejecución de las búsquedas automáticas generadas por el mismo consultor.

La forma de identificar cuáles eran los vínculos o URL's resultantes, también estuvo a cargo de la aplicación de reglas de expresiones regulares sobre el código HTML de estas páginas de resultados. Una de las cadenas o patrones separadores de cada uno de los resultados es la expresión "**href=**" y seguido a esta cadena estaría el tipo de URL, ya sea *http://* ó *ftp://* Además, para que la URL fuera considerada como válida dentro del dominio colombiano, la expresión debía terminar en *.co*. Un ejemplo de estas URL's es la siguiente:

href="http://www.eafit.edu.co/principal.shtm"

De esta forma cada vez que la aplicación encontrara cadenas de caracteres

que concordaran con el patrón definido, extraería dicha cadena y la almacenaría como una URL válida en la base de datos, para que posteriormente el Analizador se encargara de visitar y analizar más profundamente su código fuente (en HTML o javascript), en busca de más enlaces o más URL's.

Estas URL's al igual que en ocasiones anteriores, fueron sometidas a un proceso de depuración, también empleando técnicas de expresiones regulares, en donde por ejemplo se eliminaron las comillas, se reemplazó la expresión *%3a* por el símbolo de dos puntos (:) entre otro tipo de depuraciones para que la URL fuera almacenada en la base de datos sin parámetros "basura" y quedara de la mejor forma posible para que luego el analizador recibiera esta dirección correctamente y la visitara en el momento en que le correspondiera, logrando que el browser no arrojara errores al visitarla.

Otra de las aplicaciones de las expresiones regulares nombrada anteriormente, es la de descomponer las URL's resultantes en sus diferentes partes como son dominio, subdominio y extensión, además de obtener palabras claves para próximas búsquedas. En la tabla 2.1 se puede ver un ejemplo de esta división.

URL	http://dis.eafit.edu.co/emontoya/agendajulio2001.htm
Dominio	eafit.edu.co
Subdominio	dis
Extensión	htm
Palabra Clave	eafit

Tabla 2.1: Ejemplo División en componentes en una url

Así entonces, esta URL se almacenaba en la base de datos y se le asignaba un host compuesto por el dominio y el subdominio, además de su extensión. Por otra parte, la palabra clave servía para combinarla al momento de generar más búsquedas posibles. Esto último a cargo del consultor, ya que extraer información adicional como tamaño, dirección IP, entre otros estuvo a cargo del Analizador

Conociendo la extensión de la URL, ésta se clasificaba, y en caso tal de que su terminación fuera la de un archivo normal (que en este caso llamamos *complementarios*), se almacenaba en una tabla adicional creada especialmente para este tipo de URL's. Aquí se almacena entonces cada uno de esos tipos de archivos, que se clasificaron dentro de los grupos: ficheros, audio, video, documentos y fotos. Esto con el fin de separar estos archivos complementarios de las páginas Web estrictamente hablando. Cabe anotar que sólo se almacenó la URL de estos archivos y no su contenido.

Esto también permite conocer un poco las proporciones de los tipos de documentos que se referencian en la Web colombiana, como son archivos de texto, presentaciones, archivos PDF, entre otros.

Finalmente, a modo de resúmen, el analizador empleó técnicas similares de expresiones regulares al momento de interactuar con el código HTML o javascript de las páginas Web. La información que se extrajo al ejecutar esta aplicación, fueron principalmente los links que tenía cada una de las páginas. Estos links tenían varias representaciones o formas de declararse. Por ejemplo en HTML la cadena ya mencionada "**href.=**", en javascript la cade-

na "**window.open**" o expresiones como "**action**" , "**location**", "**mailto**", "**localhost**", entre otras expresiones comunes en estos lenguajes orientados a la Web.

Las direcciones resultantes también se depuraron para que fueran almacenadas de la misma forma que hizo el consultor, de tal forma que el analizador la visite en el momento que le corresponda.

Por lo anterior se concluye que utilizando técnicas de expresiones regulares, se implementaron cuatro *analizadores sintácticos* para la identificación y clasificación de toda la información. Un parser para la URL de *Google*, otro para el código HTML de la página de resultados de *Google*, otro parser para las URL's de las páginas y un último para el código fuente de cada una de las páginas Web al momento de visitarlas con el Analizador.

2.5. ESTRATEGIAS APLICADAS EN EL DESARROLLO

Durante todo el desarrollo de la aplicación se utilizaron diferentes técnicas o estratégias que de una u otra forma reducían la complejidad algorítmica del código generado, y, por añadidura, el tiempo de corrida a costa de recursos computacionales como memoria RAM y espacio en disco duro.

2.5.1. Construcción del Crawler

Para poder recorrer la Web colombiana de una forma rápida y con unas bases bien definidas se decidió dividir la construcción del Crawler en dos partes importantes, la primera se encargaba de extraer la información ya depurada de las bases de datos de uno de los buscadores de contenido en Internet más populares de los últimos tiempos (*Google*), bautizada por su funcionalidad como **Consultor**. Esta aplicación tenía la gran misión de obtener un conjunto de páginas que servirían como punto de partida para un analisis posterior por medio de las URL's. Su ciclo de vida era bastante sencillo: Iniciaba con unas consultas base de los sitios con dominio de primer nivel territorial .co, traía una por una las páginas de resultados del buscador gracias a la detección de variables del request que permiten dicha paginación, tomaba cada página y por medio de técnicas de expresiones regulares anteriormente explicadas extraía cada una de las direcciones, vínculos o *links* que pertenecieran al subconjunto de la Web colombiana, luego se tomaba cada una de las direcciones y se guardaba en la base de datos para ser visitada posteriormente por otra aplicación que se encargaba del recorrido. Durante este proceso de buscar y guardar direcciones, el Consultor también se retroalimenta formando y encolando consultas nuevas con partes del dominio de cada dirección y construyendo sentencias de búsqueda con dichas partes del dominio de los sitios encontrados, aprovechando que el buscador ofrece la funcionalidad de variar los parámetros de búsqueda.

Luego de tener una mínima cantidad de páginas extraídas por el Con-

sultor, era hora de que otra aplicación iniciara su ciclo de vida, se trata del **Analizador**, que, como su nombre lo indica, toma cada una de las direcciones guardadas en la base de datos para visitarla y analizar su código HTML en busca de nuevas direcciones para guardarlas también en la base de datos y registrar el camino de la página origen a la página destino. Esta aplicación tiene la responsabilidad de registrar los caminos entre las páginas, se puede decir que "encontraba los arcos entre los nodos".

Capítulo 3

Análisis de Resultados

3.1. Resumen del Proceso

Para este *Estudio Demo* se realizó la ejecución del Consultor, que durante sus primeros dos meses contaba con cinco instancias que realizaban consultas en forma independiente al Buscador (Google en este caso). Luego de tener una muestra representativa de páginas del dominio colombiano guardadas en la base de datos, se inicia la ejecución del Analizador igualmente con cinco instancias que analizaron páginas diferentes dividiendo dicha ejecución en rangos de números, determinados por el atributo identificador de los registros en su tabla.

La ejecución contínua de estas dos aplicaciones, pasó a darle mayor impor-

tancia al Analizador, pues incrementaron sus instancias de ejecución mientras las del Consultor disminuían, llegando así a tener nueve instancias el Analizador y sólo una el Consultor, en sus últimos días de ejecución. Este método de separación de procesamiento de datos por instancias reduce el tiempo de búsqueda y análisis del Crawler, exigiendo al máximo tanto la conexión a Internet como la conexión a la base de datos.

Otra de las razones por las cuales se realizó de esta manera fue para dejar abierta la posibilidad de utilizar varios computadores o procesadores conectados en red, los cuales alimientarían la misma base de datos dándo como resultado un análisis más completo y en menor tiempo.

Luego de la ejecución del Consultor y Analizador y después de un análisis y corrección de los datos almacenados en la base de datos (para estar seguros de que la información contenida en la misma fuera válida), pasa a tomar un papel protagónico la ejecución de la aplicación llamada *Clasificador*, la cual con base en la teoría de la conectividad de la Web, expuesta en el primer capítulo de esta Monografía, clasifica los sitios de la muestra tomada en este estudio, ubicándolos dentro de alguno de los grupos de componentes, a través del análisis de sus enlaces y del nivel de conectividad que posean.

Con la anterior clasificación, la aplicación llamada *Visualizador* asigna por medio de algoritmos diseñados, una coordenada xyz en el espacio tridimensional a cada host de acuerdo a su correspondiente tipo de componente, y así finalmente proyectar en pantalla la estructura adoptada por la muestra.

3.2. Resultados Obtenidos

Cabe aclarar que esta Monografía es un estudio demo, porque no se analizó la totalidad de las páginas de la Web colombiana sino sólo una muestra de la misma. La razón principal es que luego de varios intentos fallidos por parte de los integrantes de este proyecto de grado, no se logró tener acceso autorizado con fines científicos a la base de datos que posee la Universidad de Los Andes, quien en la actualidad administra el dominio .co. De haber sido diferente habria sido posible un estudio verídico de la estructura actual de la totalidad de la Web colombiana, de la misma forma en que se se ha realizado ya en 9 países de África, al igual que en Argentina, Austria, Brasil, China, España, Grecia, Hungría, Corea del Sur, Perú, Portugal, Reino Unido, Nueva Zelanda, Australia, Tailandia y Chile [9].

Sin embargo, este no es el caso, por lo tanto se debió desarrollar la aplicación consultora, que extrae de Google la mayor cantidad de páginas del dominio colombiano, para luego analizarlas, clasificarlas y visualizarlas.

3.2.1. Principales Resultados del Consultor y Analizador

Luego de tener una base de datos con la información básica para cada página, se ejecuta el Analizador, el cuál adiciona más registros en la base de datos y actualiza otros ya existentes, con información adicional de las páginas,

hosts, dominios, etc. Con base en esta nueva base de datos, se realizaron ciertas consultas SQL que arrojaron resultados como los que se presentan a continuación:

De las 348749 URL's recopiladas por el Consultor, se concluyó a través de un *análisis sintáctico* de sus direcciones, que 13280 no pertenecían al dominio *com.co*, por lo tanto los hosts a los cuáles pertenecen esas páginas, no fueron tomados para ciertas clasificaciones ni para algunas tablas de resultados. Con esto se determina entonces que la muestra de páginas del dominio colombiano será de 335469, y el número de hosts se reduce de 13467 iniciales a 5902, ya que las 13280 páginas que se excluyeron, pertenecían a un conjunto de 7565 hosts que igualmente debieron ser omitidos para efectos de análisis de resultados. Dichos dominios excluídos se pueden tomar como dominios "salientes", o en algunos casos que se salen del estándar, se convierten en dominios inválidos.[1]

Con este filtro aplicado a la muestra trabajada, se puede observar de manera notoria, la diferencia entre los hosts excluídos con los no excluídos en cuanto al número de páginas que los componen y cómo la muestra de hosts "válidos" se redujo a menos de la mitad, pero la muestra del número de páginas se redujo sólo en un 3.8 %, exponiendo el hecho de que los hosts excluídos poseen en promedio menos número de páginas que las que tienen los hosts del dominio colombiano tomados para la muestra.

[1]Es bien conocido que existen algunos dominios que a pesar de que no terminan en .co, son colombianos, sin embargo éstos no se tuvieron en cuenta para la muestra porque se complicaría bastante detectarlos, debido a que el estudio ya no sería a nivel de URL's sino de direcciones IP.

En la tabla 3.1 se pueden ver los diez sitios con mayor número de páginas.

SITIO Y NÚMERO DE PÁGINAS	
eltiempo.terra.com.co	21159
becas.universia.net.co	8443
www.fac.mil.co	7009
www.portafolio.com.co	6520
www.ejercito.mil.co	5506
www.universia.net.co	4809
www.conavi.com.co	4582
www.presidencia.gov.co	4220
www.geo.net.co	3899
articulo.mercadolibre.com.co	3450

Tabla 3.1: Hosts con más páginas

Como ejemplo podemos tomar el host *www.eafit.edu.co*, el cual en la misma tabla 3.1 está en el puesto 75 de los resultados de la consulta sobre toda la muestra, con un total de 635 páginas que lo componen. Dicho host pertenece al dominio *eafit.edu.co*, que posee, según la tabla 3.2, además del ya mencionado, otros 20 sitios con una cantidad de de páginas asociadas a cada uno.

Cabe anotar que esta muestra de páginas del dominio colombiano fue tomada a mediados del año 2006, y por esta razón se puede asegurar que actualmente el número de páginas y hosts de toda la Web colombiana han cambiado, algunas habrán desaparecido, y nuevas habrán nacido. De igual

manera ocurre con las clasificaciones o el tipo de componente al cual pertenece cada uno de estos hosts, ya que pueden haber algunos que hayan aumentado su nivel de conectividad y estén actualmente enmarcados dentro de otro tipo de componente diferente al que se muestra en la base de datos de este estudio.

Haciendo un análisis a nivel de dominio, se presenta en la tabla 3.2 los principales dominios, de acuerdo al número de hosts por los cuales están compuestos, ordenados de mayor a menor.

DOMINIO Y NÚMERO DE HOSTS	
uniandes.edu.co	199
univalle.edu.co	178
udea.edu.co	118
unicauca.edu.co	71
blog.terra.com.co	54
terra.com.co	50
tripod.com.co	48
puj.edu.co	38
web.com.co	37
javeriana.edu.co	33
coomeva.com.co	32
unalmed.edu.co	29
poligran.edu.co	22
eia.edu.co	21
unisabana.edu.co	21
udistrital.edu.co	20
utp.edu.co	20
eafit.edu.co	20

Tabla 3.2: Principales Dominios según su cantidad de Hosts

En la tabla 3.2, podemos observar claramente la gran presencia o participación de las entidades educativas de nivel superior (dominio *edu.co*), evidenciando su papel protagónico dentro del desarrollo de la Web colombiana, según la muestra trabajada. Entre estas entidades se presenta como una de ellas, la Universidad EAFIT con su dominio *eafit.edu.co*, que ya se había mencionado anteriormente.

Otra de las clasificaciones que se obtienen gracias a la ejecución del Consultor (con su empleo de técnicas de expresiones regulares), son el tipo de páginas que contiene la muestra, es decir su extensión. El porcentaje de participación de los principales tipos de páginas se puede observar en la tabla 3.3. Allí se destaca una tendencia en la programación Web hacia el lenguaje de programación PHP, puede ser porque es uno de los lenguajes Web que llevan más tiempo o por la facilidad en su utilización.

A través de la aplicación Consultora, se puede ver en la tabla 3.4 información bastante valiosa sobre la cantidad de hosts que posee cada sector o dominio de primer nivel en la Web, como lo son el organizacional, comercial, educación, militar, etc.

Se observa entonces en la tabla 3.4, que predominan los sectores comercial y educativo, lo que tiene bastante sentido porque cada institución educativa en el territorio colombiano cuenta con su sitio Web, adicionalmente existen diferentes redes de carácter educativo, que aportan de manera considerable en términos de sitios y páginas, por ejemplo: RENATA y sus Redes Académicas Regionales RARES. También se puede observar dos nuevos dominios que no

EXTENSIÓN Y PORCENTAJE(%)	
HTML	30.35
PHP	28.13
HTM	16.59
ASP	12.75
JSP	4.79
SHTM	3.12
ASPX	1.87
PHP3	1.32
CGI	0.49
CFM	0.25
OTRAS EXTENSIONES	0.36

Tabla 3.3: Porcentajes de Tipos de Páginas o Extensiones

se había tenido en cuenta en este estudio, *.info.co* y *.arts.co*, que pertenecen al contexto jurídico y a las artes respectivamente.

Por otro lado, la aplicación Analizadora encontró dentro del código fuente de las páginas Web, todos aquellos archivos a los que éstas hicieran referencia. A estos archivos se les llama *complementarios* y tienen una extensión dependiendo de su naturaleza. En la tabla 3.5 se puede observar un ordenamiento que indica el porcentaje de participación que tiene cada una de estos tipos de archivos en la Web colombiana según la muestra de 55836 complementarios.[2]

[2]De los 55836 complementarios de la muestra, un poco más del 4 % tenían URL's cuya longitud sobrepasaba los 130 caracteres, por esta razón se guardaron automáticamente cada una de las direcciones en un archivo de texto, para consultas posteriores

DOMINIO Y PORCENTAJE(%)	
.com.co	32.24
.edu.co	32.10
.gov.co	16.51
.org.co	12.79
.net.co	5.05
.mil.co	0.90
.nom.co	0.25
.int.co	0.07
.info.co	0.02
.arts.co	0.02
OTROS DOMINIOS	0.05

Tabla 3.4: Porcentajes de los dominios de primer nivel, según el número de sitios que los componen

En la tabla 3.5 se evidencian las tendencias en el uso de archivos gráficos para la Web, los archivos GIF y JPG son los únicos tipos de archivos de imagenes que están al tope de la tabla. Además, sobresale la extesión de archivos de estilos y los archivos de animaciones Flash, como un testimonio de que en la Web colombiana existe la preocupación por el diseño gráfico.

Comparando los resultados parciales de este estudio demo, con los obtenidos en un estudio real de la Web chilena realizado en el 2006 [9], se puede observar que la tendencia es muy similar, ya que hay un predominio notable de los archivos pdf dentro de las páginas en ambos estudios.

TIPO DE ARCHIVO Y PORCENTAJE(%)	
PDF	51.76
DOC	17.79
CSS	6.17
XLS	5.61
GIF	5.26
JPG	5.16
PPT	2.11
SWF	1.09
ZIP	0.83
RTF	0.81
OTROS TIPOS	3.41

Tabla 3.5: Porcentaje de Participación de los tipos de archivos o complementarios en la Web colombiana

Continuando con el tema de los tipos de complementarios encontrados en la muestra, se agruparon ciertas extensiones de archivos de acuerdo a su naturaleza, donde se tomaron las siguientes extensiones dentro de cada tipo:

- **DOCUMENTOS** (xls, doc, pdf, txt, etc)
- **FOTOS** (jpg, bmp, gif, jpeg, png, ico)
- **FICHEROS** (exe, zip, rar, sit, img, tar, msi, class, css)
- **VIDEOS** (mov, avi, mpeg, mpg, ogm, swf)

- **AUDIO** (wav, mp3, mid, mp4, rma)

La Tabla 3.6 presenta el orden de aparición de cada naturaleza y su porcentaje de participación en la muestra, revelando como mayoría los documentos e imagenes. Aunque la información relacionada con los archivos cambia rápidamente, porque la utilización de canales con mayor ancho de banda permiten el uso de archivos cada vez más grandes.

TIPO DE ARCHIVO Y PORCENTAJE(%)	
DOCUMENTOS	79.95
IMAGENES	11.93
FICHEROS	7.67
VIDEOS	1.12
AUDIO	0.14

Tabla 3.6: Porcentaje de Participación de las naturalezas de los archivos complementarios en la Web Colombiana

Otros datos generales obtenidos por medio de consultas SQL sobre la base de datos son los siguientes:

- Según la muestra trabajada en este estudio, los sitios de la Web colombiana están compuestos en promedio por 56 páginas.
- El promedio del tamaño que manejan las páginas Web estudiadas es de 713 Kilobytes.

3.2.2. Principales Resultados del Clasificador

La aplicación clasificadora que se desarrolló, agrupa cada uno de los sitios o hosts de la muestra estudiada, según la teoría de conectividad de la Web, explicada en el primer capítulo de esta monografía.

Este proceso de clasificación se realizó de una manera bastante cuidadosa, ya que cualquier decisión mal tomada al momento de implementar los algoritmos, repercutía directamente en la estructura resultante de la clasificación de esta muestra y por ende de la Web colombiana.

Luego de definir los algoritmos para dicha clasificación, se procedió a ejecutar esta aplicación, analizando uno a uno los enlaces entre las páginas y los hosts a los cuales pertenecían, para, finalmente, conformar una estructura de clasificación de los hosts, según los enlaces entrantes y salientes de cada uno, definiendo así su nivel de conectividad en la Web, identificado por el término llamado *Tipo de Componente* que fue explicado también en el primer capítulo de este documento.

La clasificación y el porcentaje de participación de cada uno de estos tipos de componentes, según el número de sitios asociado a cada uno (Con la muestra de 5902 hosts del dominio colombiano), se puede observar en la tabla 3.7.

Con la información suministrada por la tabla 3.7, podemos concluir que la Web Colombiana está compuesta en su mayoría por componentes aislados o

TIPO DE COMPONENTE Y PORCENTAJE(%)	
ISLANDS	46.85
OUT	28.26
IN	6.27
MAIN-MAIN	5.29
MAIN-IN	5.00
TENTACLE-IN	4.74
MAIN-NORM	1.29
TENTACLE-OUT	1.22
MAIN-OUT	0.85
TUNNEL	0.29

Tabla 3.7: Porcentaje de participación de los tipos de componentes según el número de hosts o sitios que los componen

de tipo "ISLAND" con casi el 50 % de participación, seguido por los "OUT" que tiene un 28 % del total de hosts de la muestra. Si se sigue observando los porcentajes y el orden de aparición de cada tipo de componente se ve una gran similitud con respecto a los resultados del estudio realizado en Chile.[9]. La diferencia, que se pone al descubierto entre las estructuras de los dos países radica en la posición en que figuran los componentes "TENTACLE-IN" y "TENTACLE-OUT", donde en la de Chile figura primero los T-OUT y en la muestra colombiana figuran primero los T-IN, y visceversa, aunque la diferencia de porcentajes de participación entre uno y otro no es significativa.

Una vez obtenida la clasificación de la muestra en tipos de componentes

de acuerdo a la tabla 3.7, se pueden entrar a analizar algunos de ellos, como es el caso particular de la tabla 3.8, en donde se analizan específicamente todos los componentes tipo *MAIN* (que son: MAIN-IN, MAIN-OUT, MAIN-MAIN y MAIN-NORM), de tal forma que se obtengan los dominios de primer nivel colombiano que hacen parte de él. Mostrando los sectores con sitios que más se encuentran enlazados entre sí.

DOMINIO Y PORCENTAJE(%)	
.edu.co	28.10
.gov.co	24.15
.com.co	22.92
.org.co	18.28
.net.co	4.50
.mil.co	1.64
.arts.co	0.14
.int.co	0.14
.nom.co	0.14

Tabla 3.8: Porcentajes de participación de los Dominios de Primer Nivel Territorial, en el "MAIN" de la muestra, según el número de sitios que los componen

En la tabla 3.8 se puede apreciar claramente cómo cambia el comportamiento o la clasificación de los dominios de primer nivel territorial, cuando se analiza determinado tipo de componente. Mientras que en la tabla 3.4 mostraba al dominio comercial (*.com.co*) en primer lugar, ahora a nivel del componente *MAIN* (El cuál comprende 733 hosts en la muestra), dicho sector

comercial pasa a un tercer lugar y el sector educativo junto con el dominio *gov.co*, se ubican cómo los principales protagonistas de los componentes del *MAIN* de la muestra, cada uno con porcentajes de participación significativos.

De la información suministrada en las tablas anteriores, se pueden derivar dos tablas más, indicando los hosts que más tienen conectividad, evaluando tanto la cantidad de enlaces entrantes como los salientes, los cuales se llamaron en este estudio hosts *anteriores* y hosts *siguientes* respectivamente.

El primer caso es para los quince hosts colombianos de la muestra, que son más referenciados o apuntados por otros hosts, y los resultados que indican el número de veces que son referenciados dichos hosts, se pueden apreciar en la tabla 3.9.

Para el segundo caso, se tiene la tabla 3.10 que da a conocer los quince hosts que poseen más enlaces salientes, o, lo que es lo mismo, los sitios que más hosts *siguientes* tienen, todos dentro del dominio colombiano.

En la tabla 3.9 de *Anteriores por Hosts*, la consulta sobre la base de datos arrojó en total 2768 registros de hosts que al menos poseen un enlace entrante, es decir, al menos un host referencia dichos sitios. Por otro lado, la consulta de *Siguientes de Hosts* sólo arrojó 1259 registros en total que al menos poseen un enlace saliente dentro del dominio colombiano. Con esta diferencia entre el número de hosts retornados por cada consulta se puede reconfirmar por otra vía(Como lo muestra la tabla 3.7), el hecho de que en

NÚMERO DE ANTERIORES POR HOST	
967	*eltiempo.terra.com.co*
843	*articulo.mercadolibre.com.co*
665	*articulo.deremate.com.co*
634	*www.eltiempo.com*
511	*www.fac.mil.co*
465	*www.mercadolibre.com.co*
439	*www.portafolio.com.co*
372	*www.ejercito.mil.co*
321	*www.universia.net.co*
234	*www.humboldt.org.co*
231	*pvirtual.uptc.edu.co*
222	*www.funlam.edu.co*
190	*www.javeriana.edu.co*
175	*www.caracol.com.co*
164	*www.terra.com.co*

Tabla 3.9: Número de hosts que apuntan al sitio relacionado(Anteriores de un Host)

la muestra trabajada, hay una mayoría de hosts ubicados dentro del tipo de componente *OUT* que de componentes tipo *IN*. Lo anterior se concluye debido a que la naturaleza de los "OUT" es tener anteriores y la naturaleza de los componentes "IN" es tener hosts siguientes. En resumen se puede detectar claramente que existe una mayor cantidad de enlaces entrantes que de enlaces salientes en los hosts colombianos de la muestra estudiada.

Otro análisis que se puede realizar, gracias a datos específicos recopilados por las aplicaciones desarrolladas, es el correspondiente a los hosts colombia-

NÚMERO DE SIGUIENTES POR HOST	
854	*articulo.mercadolibre.com.co*
691	*listado.mercadolibre.com.co*
685	*listado.deremate.com.co*
530	*www.hchr.org.co*
509	*eltiempo.terra.com.co*
388	*www.universia.net.co*
379	*www.invemar.org.co*
343	*portafolio.com.co*
331	*boletines.latinpyme.com.co*
313	*www.cab.int.co*
276	*www.mineducacion.gov.co*
271	*www.sirideec.org.co*
241	*virtual.uptc.edu.co*
234	*www.amigomed.edu.co*
234	*www.humboldt.org.co*

Tabla 3.10: Cantidad de hosts colombianos a los cuales apunta cada sitio relacionado(Siguientes de un Host)

nos que más referencian sitios fuera del dominio colombiano. Estos resultados se presentan en la Tabla 3.11 y las posiciones de los hosts en comparación con las obtenidas en la tabla 3.10, cambiaron notablemente, ya que algunos sitios dejaron de ser principales, otros diferentes pasaron a serlo y en general todos cambiaron de posición.

Se puede observar además cómo la Universidad EAFIT, específicamente su Departamento de Informática y Sistemas (*DIS*), figura dentro de esos quince hosts que más enlaces salientes fuera del dominio colombiano posee

NÚMERO DE SIGUIENTES POR HOST	
1247	*www.nuestracolombia.org.co*
692	*eltiempo.terra.com.co*
456	*matematicas.udea.edu.co*
375	*www.colombiaaprende.edu.co*
364	*dis.eafit.edu.co*
331	*www.acis.org.co*
306	*www.usergioarboleda.edu.co*
301	*uvirtual.ean.edu.co*
236	*www.javeriana.edu.co*
202	*www.epm.net.co*
189	*www.proexport.com.co*
184	*sat.edu.co*
174	*www.cta.org.co*
169	*www.imagine.com.co*
168	*lmejia.emcali.net.co*

Tabla 3.11: Cantidad de hosts *fuera del dominio colombiano* a los cuales apunta cada sitio relacionado

dentro de la muestra analizada, además de otras Instituciones de Educación Superior, que como se mencionaba anteriormente, juegan un papel definitivo en la misión investigativa y de desarrollo de la Web en Colombia que está encomendada a los organismos interesados en estos temas tecnológicos.

Por último, se sabe que los enlaces reales están dados a nivel de páginas, y debido a esto, es bastante importante mencionar un valor agregado que se adoptó en este estudio, que es el concepto de *Distancia entre Componentes Adyacentes*. Éste se utilizó en la base de datos como una clasificación o dato

adicional a cada host, que indica el nivel de cercanía entre dos sitios, donde cada uno pertenece a un tipo de componente diferente, que son adyacentes entre sí.

En toda la base de datos de la muestra no se encontraron sitios con una distancia mayor a 3, con respecto a sus componentes adyacentes. De hecho, se puede observar en la tabla 3.12, cómo están organizados los tipos de componentes de acuerdo a las distancias a las cuales se encuentran de sus tipos de componentes adyacentes.

SITIOS	*TIPO DE COMPONENTE*	*DISTANCIA*
1614	OUT	1
363	IN	1
279	TENTACLE–IN	1
72	TENTACLE–OUT	1
4	TUNNEL	1
53	OUT	2
7	IN	2
1	TENTACLE–IN	2
1	OUT	3

Tabla 3.12: Cantidad de sitios que hay en los determinados tipos de componentes, de acuerdo a la distancia a la cual se encuentran con respecto a sus tipos de componentes adyacentes

Los resultados de la tabla 3.12 reflejan cómo el tipo de componente *OUT* tiene presencia en las tres distancias encontradas y posee también la mayor

participación dentro de los componentes que están a la distancia menor con respecto a su grupo de componentes adyacentes.

Hay ciertos tipos de componentes a los cuales no se les calculó distancia alguna debido a su naturaleza, por lo tanto en las tablas observadas, la muestra se reduce considerablemente.

Este tema de las distancias entre tipos de componentes es posible representarlo gráficamente gracias a la aplicación que fue desarrollada en este proyecto para tal fin, llamada "Visualizador". Que se verá más detalladamente en el manual del sistema y en el manual del usuario diseñados especialmente para estos casos.

Adicional a los resultados ya mostrados en las tablas de este capítulo, se pueden añadir otras más en un futuro muy próximo, luego de una segunda corrida de las aplicaciones desarrolladas en este proyecto de grado y durante un mayor tiempo, para obtener una muestra más significativa, o lo que sería ideal, trabajar con la base de datos *completa* que está en poder del administrador del dominio colombiano *.co* que es la Universidad de los Andes. Así entonces, estas futuras *corridas*, podrían evidenciar cada vez con mayor certeza la estructura actual de la Web colombiana, al igual que su cambio y desarrollo, permitiendo llegar a otro tipo de conclusiones y seguir escribiendo una historia *verídica* de la evolución de la Internet en el país que se base en hechos y no sólo en teoría.

Conclusiones

- Las *islas* encontradas por el aplicativo serán determinadas sólo en la etapa de Consulta ya que cada una de ellas sólo puede ser enlazada gracias a un buscador, es decir, el analizador no puede llegar hasta una isla con un enlace desde el MAIN porque entonces no se trataría de una isla.

- Si un buscador es considerado como un sitio Web, la base de datos de enlaces que éste posee, lo convertiría en un sitio de un nivel superior que se conecta con todos los demás y esto afectaría directamente la topología resultante de este estudio, ya que no existirían *islas* porque toda página tendría al menos un enlace con el buscador. Por esta razón se determina que un enlace válido es aquel que es encontrado por el analizador, y, los enlaces del consultor no serían válidos ya que no se tratan de enlaces reales, o que estaban pensados al momento de crear la página, sino que son producto de un análisis de la Internet por robots de búsqueda o agentes inteligentes.

- Basándonos en la teoría de grafos y analizando el comportamiento de

las páginas Web con sus enlaces, se puede decir que la topología de la Internet está compuesta por todas las clases de grafos posibles, pero las páginas que componen un sitio tienden a comportarse como un multigrafo. Además, la relación entre los sitios, corresponde a un grafo dirigido o no dirigido dependiendo de los enlaces entre sí.

- La utilización de una base de datos como repositorio para el almacenamiento de información fue útil para reducir el tiempo de procesamiento que emplea el algoritmo de búsqueda de componentes fuertemente conexos o posibles MAIN, ya que la relación entre estructuras permite realizar recorridos hacia adelante o hacia atrás sin ninguna diferencia en cuanto a la complejidad algorítmica.

- Si se presentan varias *islas* de tamaño considerable, mostrando una topología de la Web colombiana fragmentada, puede ser porque este estudio no se realiza en la totalidad de páginas, o porque se presentan ciertos nichos de hosts principales que hace que la Web sea discriminada por algún factor característico.

- Dado el caso en que se encuentre un conjunto de vértices fuertemente conexos en el *MAIN* y dicho conjunto no sea a su vez fuertemente conexo, luego de varios cálculos con ejemplos reales, se definió como criterio principal que se escogería como *MAIN* principal, aquel CFC que tenga más páginas o más hosts asociados.

- El análisis de la Web colombiana, y en especial la realización de estudios de este tipo, siempre serán meramente una aproximación a la

realidad sólo hasta que NIC Colombia decida hacer pública la totalidad de la base de datos de dominios colombianos, proyecto que esta en negociaciones desde finales de 2003 pero que aún no se ha hecho efectivo por razones desconocidas.

- La Universidad EAFIT puede hacer uso del paquete de software ofrecido en este proyecto de grado, para ejecutar el Crawler durante un rango de tiempo mucho mayor y con tecnologías más avanzadas, de tal forma que el estudio adquiera una dimensión mayor y la muestra poblacional sea más grande, haciendo que los resultados se acerquen con mayor claridad a la realidad de la Web colombiana.

- La ejecución de las aplicaciones consultora y clasificadora se debió detener en un momento determinado, ya que su ejecución no llegaba a su fin debido a que siempre iban a encontrar más y más enlaces para ser analizados porque siempre habría al menos una instancia del Consutor en ejecución generando y guardando consultas temporales que podían arrojar resultados no existentes en la base de datos, por lo tanto las páginas sin visitar siempre incrementaban.

Glosario

- *Agente Inteligente*: Equivalentes en términos computacionales a un proceso del sistema operativo, que existen dentro de cierto contexto o ambiente, y que se pueden comunicar a través de un mecanismo de comunicación inter-proceso, usualmente un sistema de red utilizando protocolos de comunicación.

- *ARPA*: Acrónimo inglés de Advanced Research Projects Agency (Agencia de Proyectos de Investigación Avanzada). Organismo del Departamento de Defensa de Estados Unidos creado en 1958 como consecuencia tecnológica de la llamada Guerra Fría en contra de la Unión Soviética y de la cual surgieron una década después los fundamentos de ARPANET, red origen de lo que es hoy en día Internet.

- *ARPANET*: La red de computadoras ARPANET (Advanced Research Projects Agency Network) fue creada por encargo del Departamento de Defensa de los Estados Unidos como medio de comunicacion para los diferentes organismos estadounidenses. El primer nodo se creó en la Universidad de California y fue la espina dorsal de Internet hasta 1990,

tras finalizar la transición al protocolo TCP/IP en 1983.

- *BACKBONE*: Mecanismo de conectividad primario en un sistema distribuido. Todos los sistemas que tengan conexión al backbone (columna vertebral) pueden interconectarse entre sí, aunque también puedan hacerlo directamente o mediante redes alternativas.

- *BITNET*: Red de sitios educativos (investigación y universitarios) separada de Internet, pero el correo electrónico es libremente intercambiado entre BITNET e Internet. Los Listservs son la forma mas popular de los grupos de noticias originados en BITNET. Las computadoras de BITNET son usualmente mainframes corriendo el sistema operativo VMS (variante de UNIX).

- *CRAWLER*: programa diseñado para recorrer la web siguiendo los enlaces entre páginas. Esta es la forma habitual empleada por los principales buscadores para encontrar las páginas que posteriormente forman parte de sus bases de datos.

- *DANTE*: Organización sin ánimo de lucro creada en 1993 con el objetivo de mejorar las redes de comunicación de los organismos de investigación europeos.

- *FTP*: Es uno de los diversos protocolos de la red Internet, concretamente significa File Transfer Protocol (Protocolo de Transferencia de Archivos) y es el ideal para transferir datos por la red.

- *GEANT2*: Es la red de investigación y educación Pan-Europea de septima generación, sucesora de la red de investigación multi-Gigabit Pan-

Europea GÉANT. El proyecto dentro de el cual se financia la red comenzó oficialmente el 1 de septiembre de 2004, y funcionará por cuatro años.

- *GOPHER*: Servicio de Internet consistente en el acceso a la información a través de menús. La información se organiza de forma arborescente, de forma que sólo los nudos contienen menús de acceso a otros menús o a hojas, mientras que las hojas contienen simplemente información textual. De esta forma, es un predecesor de la Web, aunque sólo se permiten enlaces desde nudos-menús hasta otros nudos-menús o a hojas, y las hojas no tienen ningún tipo de hiperenlaces.

- *INTERRED*: una Interred es un sistema de comunicación compuesto por varias redes que se han enlazado juntas para proporcionar unas posibilidades de comunicación ocultando las tecnologías y los protocolos y métodos de interconexión de las redes individuales que la componen.

- *MILNET*: Una de las redes DDN (Defense Data Network) que constituyen Internet y que está dedicada a comunicaciones militares estadounidenses no clasificadas. Fue construida con la misma tecnología que ARPANET y continuó operando después de la desconexión de ésta.

- *NAP*: Punto de Acceso a la Red. Es una facilidad de intercambio público de red donde los proveedores de acceso a Internet (ISPs: Internet Service Providers) pueden conectarse entre sí. Los NAPs son un componente clave del backbone de Internet porque las conexiones dentro de ellos determinan cuánto tráfico puede rutearse. También son los puntos de mayor congestionamiento de Internet.

- *NEWS*: Forma habitual de denominar el sistema de listas de correo mantenido por la red USENET.

- *NSFNET*: Acrónimo inglés de National Science Foundation's Network. La NSFNET comenzó con una serie de redes dedicadas a la comunicación de la investigación y de la educación. Fue creada por el gobierno de los Estados Unidos (a través de la National Science Foundation), y fue reemplazada por ARPANET como backbone de Internet. Desde entonces ha sido reemplazada por las redes comerciales.

- *PARSER*: Un parser es así mismo un programa que reconoce si una o varias cadena de caracteres forman parte de un determinado lenguaje, es utilizado por ejemplo en compiladores.

- *POP*: Point of Presence. Punto de acceso a Internet. Ubicación de un punto de acceso a Internet. Un POP tiene necesariamente una única dirección de IP. POP también se refiere a la construcción o el lugar donde las líneas de telecomunicación de alto ancho de banda terminan, líneas subsidiarias (de cobre o fibra, T1, etc.) continúan luego desde el POP.

- *REQUEST*: En inglés, request significa pedir, solicitar. En efecto, la acción de escribir una dirección en la línea de URL de tu navegador, se traduce en solicitar un determinado fichero a un servidor, esta acción es denominada como hacer un request al servidor, es decir, permite el acceso a toda la información que pasa desde el navegador del cliente al servidor.

- *TCP/IP*: Conjunto básico de protocolos de comunicación de redes, popularizado por Internet, que permiten la transmisión de información en redes de computadoras. El nombre TCP/IP proviene de dos protocolos importantes de la familia, el Transmission Control Protocol (TCP) y el Internet Protocol (IP).

- *TELNET*: Es el nombre de un protocolo (y del programa informático que implementa el cliente) que sirve para acceder mediante una red a otra máquina, para manejarla como si estuviéramos sentados delante de ella. Para que la conexión funcione, como en todos los servicios de internet, la máquina a la que se accedía debe tener un programa especial que reciba y gestione las conexiones.

- *TIC*: Las TIC se conciben como el universo de dos conjuntos, representados por las tradicionales Tecnologías de la Comunicación (TC) - constituidas principalmente por la radio, la televisión y la telefonía convencional - y por las Tecnologías de la Información (TI) caracterizadas por la digitalización de las tecnologías de registros de contenidos (informática, de las comunicaciones, telemática y de las interfases).

- *Tipo de Dato Abstracto*: Un tipo de dato abstracto o TDA es un modelo matemático compuesto por una colección de operaciones definidas sobre un conjunto de datos para el modelo.

- *Token Ring*: Arquitectura de red desarrollada por IBM con topología lógica en anillo y técnica de acceso de paso de testigo. Cumple el estándar IEEE 802.5.

- *URL*: La dirección de una fuente de información. Está compuesto por cuatro partes distintas: el tipo de protocolo (http, ftp, gopher), el nombre de la máquina, la ruta del directorio y el nombre del archivo.

Bibliografía

[1] AHO. Alfred, HOPCROFT. John, and ULLMAN. Jeffrey. *Estructuras de Datos y Algoritmos*. Adisson Wesley Iberoamericana, México, 1998.

[2] BRODER. Andrei, KUMAR. Ravi, MAGHOUL. Farzin, RAGHAVAN. Prabhakar, RAJAGOPALAN. Sridhar, STATA. Raymie, TOMKINS. Andrew, and WIENER. Janet. Graph Structure in the Web. Technical report, AltaVista Company, IBM Almaden Research Center, Compaq Systems Research Center, California, Estados Unidos, 2000.

[3] Cristian Aramayo. Introducción a MySql. http://www.salnet.com.ar/inv_mysql/pag01_intro.htm, Última Visita: Junio 2006.

[4] Portal Cientfico Portal Gerona. Definición de Internet de forma clasificada. http://www.gerona.inf.cu/, Última Visita: Enero 2006.

[5] Colnodo. Proyecto "Monitor Politicas de Internet en America Latina y el Caribe". http://lac.derechos.apc.org/investigacion/tic_colombia.doc, Noviembre 2001.

[6] NAP Colombia. Network Access Point. http://www.nap.com.co, Última Visita: Abril 2006.

[7] OPEN SYMPHONY QUALITY COMPONENTS. OSCache. http://www.opensymphony.com/oscache, Última Visita: Mayo 2007.

[8] Agenda de Conectividad. Gobierno y universidades crearon hoy la corporación RENATA. http://www.agenda.gov.co/BulletinBoard/view_one.cfm?MenuID=3&ID=330, Última Visita: Mayo 2007.

[9] Centro de Investigación de la Web (CIW) y Yahoo! Research. Características de la web chilena 2006. http://www.ciw.cl/material/web_chilena_2006/index.html, Última Visita: Mayo 2007.

[10] Universidad de los Andes. Historia de la conexión de la Universidad de los Andes a Internet. http://interred.wordpress.com/2002/05/12/colombia-historia-de-la-conexion-de-uniandes-a-internet-2/, Última Visita: Marzo 2006.

[11] Revista digital de InfoVis.net. Teoría de grafos. http://www.infovis.net/printMag.php?num=137&lang=1, Última Visita: Marzo 2006.

[12] Artículos en Línea Albanet. Artículos sobre la historia del internet. http://www.albanet.com.mx/articulos/HISTORIA.htm, Última Visita: Marzo 2006.

[13] America Latina Interconectada Con Europa ALICE. V Conferencia Iberoamericana de Delegaciones Universitarias y Redes de Educación Superior. http://archive.dante.net/alice/ALICEbrochure.pdf, Junio 20 2006.

[14] FURLAN. Luis. Boletín Bimensual DeClara. Technical Report Año 2 Número 6, Cooperación Latino Americana de Redes Avanzadas, Marzo 2006. http://www.redclara.org/07/02_02/06.htm.

[15] Adriano MORAN. IPv6: Nuevo código de direcciones IP. http://www.consumer.es/web/es/tecnologia/internet/2006/07/26/153823.php, Última Visita: Mayo 2007.

[16] SEOLUCIÓN-Adesis Netlife. Tipología de buscadores: robots y directorios. http://www.seolucion.com/articulos/040308-robots-y-directorios.asp, Última Visita: Mayo 2007.

[17] Instituto Peruano de Marketing. Glosario del Navegante. http://www.ipm.com.pe/glosarionave.htm, Última Visita: Abril 2006.

[18] Proyecto Pixel. Glosario en Línea. http://www.proyectopixel.com/glosario.htm, Última Visita: Marzo 2006.

[19] The Apache Jakarta Project. Jakarta Commons. http://jakarta.apache.org/commons, Última Visita: Mayo 2007.

[20] BAEZA YATES. Ricardo and POBLETE. Barbara. Evolution of the chilean web structure composition. In *First Latin American Web Con-*

gress Book, page 11, Santiago de Chile, Chile, Noviembre 2003. The Institute of Electrical and Electronics Engineers, Inc.

[21] TANENBAUM. Andrew S. *Redes de Computadoras*. PEARSON, México, tercera edition, 1997.

[22] Glosarios Técnicos en Línea. Definición del término Internet. http://www.chenico.com/glosarioi.htm, Última Visita: Marzo 2006.

[23] TomaToma.ws. Integridad Referencial en MySql. http://www.tomatoma.ws/articulo.php?topic_id=355&forum_id=30, Última Visita: Junio 2006.

[24] Gregory M TRAVIS. *JDK 1.4 Tutorial – A practical guide to coding with the new features of Java*. Manning Publications Company, Estados Unidos, 2002.

[25] Java Users Group Argentina. Glosario de Java Users Group Argentina. http://cricava.com/java/glossary, Última Visita: Septiembre 2006.

[26] Biblioteca virtual Wikipedia. Definición del término Internet. http://es.wikipedia.org/wiki/Portada, Última Visita: Marzo 2006.

[27] Enciclopedia Wikipedia. J2ee. http://es.wikipedia.org/wiki/J2EE, Última Visita: Septiembre 2006.

[28] Enciclopedia Wikipedia. Regular Expression. http://en.wikipedia.org/wiki/Regular_expression, Última Visita: Julio 2006.

[29] Foros y tutoriales técnicos en línea. Definición de dominio de primer nivel. http://www.desarrolloweb.com/articulos/8.php?manual=13, Última Visita: Marzo 2006.