CHARACTERIZATION OF THE COLOMBIAN WEB

ALEJANDRO MOLINA RAMÍREZ

UNIVERSIDAD EAFIT
DEPARTAMENTO DE INGENIERÍA DE SISTEMAS
MEDELLÍN
2009

CHARACTERIZATION OF THE COLOMBIAN WEB

ALEJANDRO MOLINA RAMÍREZ

Trabajo de grado para optar por el
título de Ingeniero de Sistemas

Asesor
JUAN GUILLERMO LALINDE PULIDO
Doctor en Telecomunicaciones

UNIVERSIDAD EAFIT
DEPARTAMENTO DE INGENIERÍA DE SISTEMAS
MEDELLÍN
2009

Nota de aceptación

____

____

____

____

Presidente del Jurado

____

Jurado

____

Jurado

Medellín ________________________________

AGRADECIMIENTOS

Al Doctor Juan Guillermo Lalinde Pulido, asesor del proyecto por su invaluable apoyo y guía durante la elaboración de este proyecto.

A todas aquellas personas que de una u otra forma colaboraron con la elaboración de este proyecto.

Finalmente a mi familia, profesores, amigos y compañeros por sus aportes y apoyo durante toda la carrera.

Table of Contents

Illustration Index

Drawings Index

1. Introduction

In this section we present the characteristics of the Web and of the studied sample, also the methodology used to collect documents.

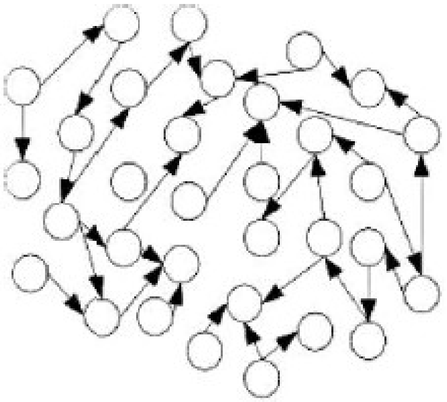
1.1. How is the Web?

The Web is more than a simple set of documents on different server, because there exists information relationships between the documents by means of the links established among them. This has many advantages, for both users when they search for information and for programs that crawl the web, searching for content to collect and index (as web search engines). Because of this, it is suggested that the Web can be modeled as a directed graph, in which every page is a node, and the links among pages are the arcs.

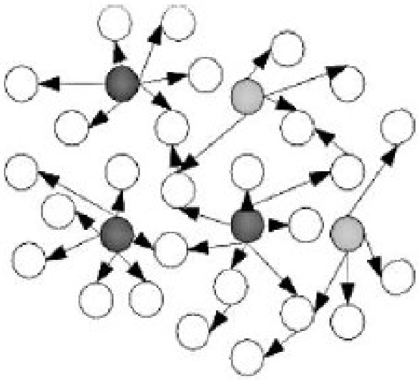
In general pages tend to link to other similar pages[1], this way, it is possible to recognize pages that are better than others, that is, pages that receive a higher number of references than normal. The web has a structure that can be denominated as free of scale network. Such networks, contrary to random networks are classified by an uneven distribution of links and because such a distribution follows the power-law[2].

$$P_r\left(\Gamma\left(p\right)=k\right)\alpha k^{-\theta}$$

Hihgly linked nodes act as centers that connect most of the other nodes to the network as shown on Drawings 1 and 2 where both networks have 32 nodes and 32 links, but follow different distributions.



Drawing 1: Random Network



Drawing 2: Free of Scale Network

This means that the distribution of links is rather skewed; a few pages receive many links while the majority receive very few or none at all. In this study it is shown that such distribution can be applied to many aspects of the Web, for which it can be said that they follow a “Zipf law”, in reference to Kingsley Zipf who proposed the distribution to model the frequency of appearance of words in texts[3].

According to this model, the probability of finding an element of a certain size x is proportional to

x^{α} where $\alpha > 1.$

When this distribution is plotted in a graph with a logarithmic scale, a straight line is found as is the case with many of the graphics in this study.

1.2. Studying the Web of a country

The free of scale networks are auto-similar that is, a small sample has the characteristics of the complete network (that is, the characteristics trascend the scale on which the network is viewed). it is shown in this study that this is the case of the Colombian Web, that presents characteristics very similar to the global network and networks of other countries, in spite of having just a small fraction of the total number of collectable pages in the global Web, estimated in 2005 to be around 11.5 billion pages[4].

The national Web can be defined as the set of pages related to a country.

Technically it is hard to distinguish whether a page is associated to the country of study, specially for the Colombian case and because it was not possible to get the complete list of domains from the (.co) domain registrar.

There are also studies done on other national domains as:

- Africa (9 countries)[5]
- Argentina (only universities)[6]
- Austria[7]
- Brazil[21] [8]
- China[9]
- Spain[22]
- Greece[26]
- Hungary[10]
- South Korea[19]
- Peru[11]
- Portugal[12]
- United Kingdom, New Zealand and Australia(only universities)[13]
- Thailand[14]

1.3. Collecting pages

The data collection was downloaded on February 2009, using the data gathering software WIRE[15]. The computer used for downloading had a 1GHZ processor and 512Mb of RAM, running Kubuntu 8.04.

The gatherer starts by downloading a set of initial websites (seeds), those are the initial known domains collected before. From the downloaded pages new links are extracted and then filtered, discriminating between Colombian domains (.co) and other domains. In total more than 4.5 million pages were downloaded, the data downloaded uses 37 GB of space on disk.

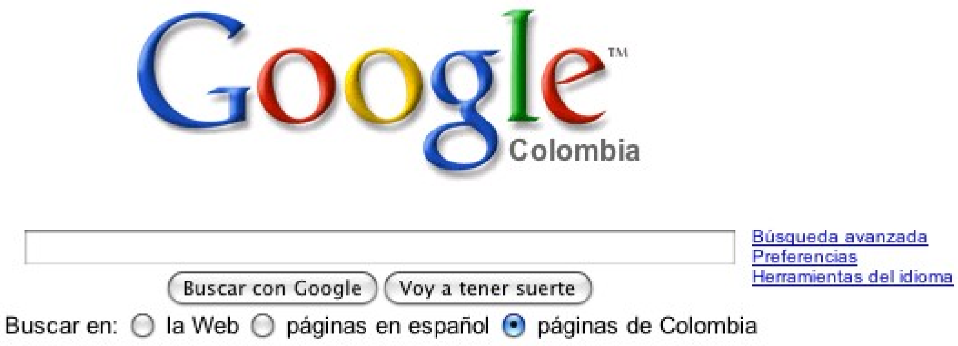
1.4. Seeds

In order to start crawling the web, an initial set of websites is required as entry point or gateway into the web, this initial set is very important because there are websites inside of the Web called islands, that can only be analyzed if the website address is known beforehand, this sites are special in that no link from any other site points to them, therefore they are practically invisible when crawling the web.

1.4.1. Obtaining the seeds

For this analysis of the Colombian Web, the registrar of the local domain (. co) could not provide us with the list of all the known domains, therefore this report is not an exhaustive analysis of the local Web in that many websites could have been left unanalyzed, furthermore several local websites are registered not using the local (.co) domain but instead use the global (.com) domain.

In order to obtain a somewhat meaningful list of initial seeds for the Colombian Web we used google, specifying that we wanted to find results only on Colombian websites as shown on Illustration 1.



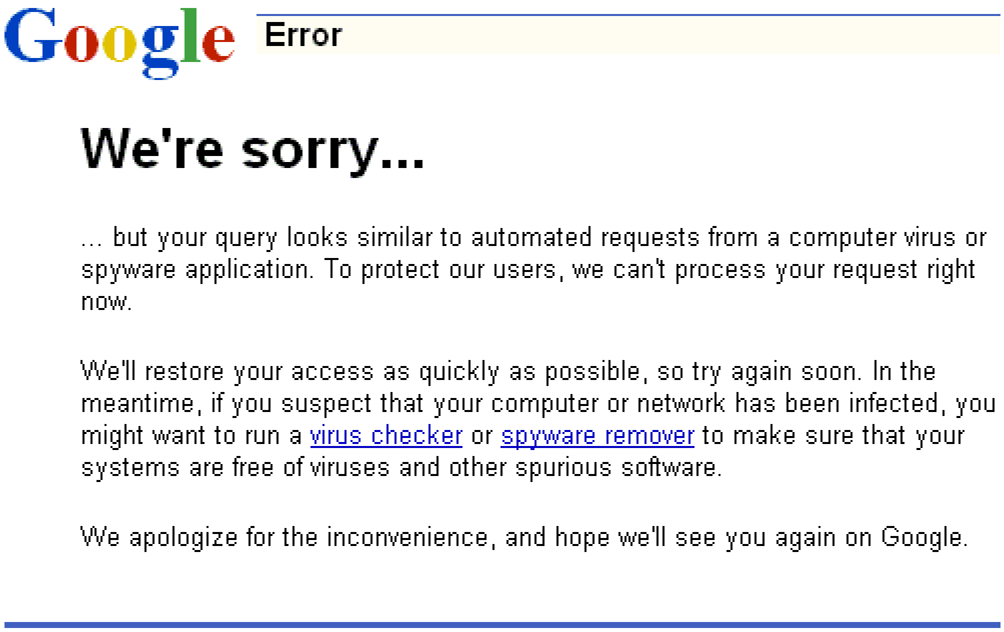
Buscar en: ○ la Web ○ páginas en español ● páginas de Colombia

Illustration 1: Search Only in Colombian Pages

Then several queries were searched following patterns as:

- site:.com a
- site:.com b
- site:.co a

Doing this for all the domain suffixes to be studied (.co, .com, .net, .org) and then crawling all the result pages returned by google searching for URLs[16].



Google Error

We're sorry...

... but your query looks similar to automated requests from a computer virus or spyware application. To protect our users, we can't process your request right now.

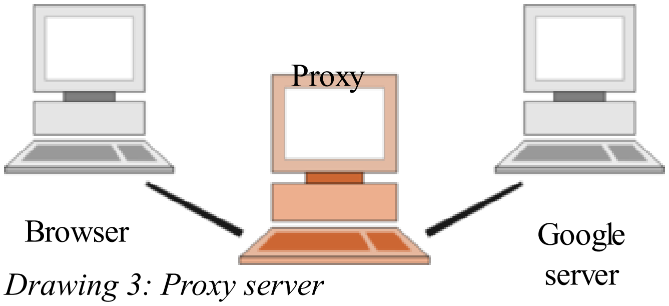
We'll restore your access as quickly as possible, so try again soon. In the meantime, if you suspect that your computer or network has been infected, you might want to run a virus checker or spyware remover to make sure that your systems are free of viruses and other spurious software.

We apologize for the inconvenience, and hope we'll see you again on Google.

Illustration 2: Google Error automated requests

Given that google protects itself from robot based behaviors as seen on Illustration 2, it was important for the process of the gathering to be verified by hand at every step.

In order to collect the set of websites, we created a proxy server and used it as a packet analyzer to log every URL seen on every page that was navigated.



Drawing 3: Proxy server

Given that google uses compression, the proxy had to uncompress all the responses then run regular expressions to extract the URLs, finally keeping a list of URLs that would be written to a file when the proxy was signaled to be shutdown.

After crawling google manually through a web browser and letting the proxy gather all the links, the resulting list was again filtered with a small script that would discard any link that did not belong to the initial defined set of domain suffixes (.co, .com, .net, .org), this was done because the resulting list of domains included sites that were known not to be from Colombia, and that ended in domain suffixes of other countries like Chile (.cl), this filtering was applied being conscient of the consequences it would have regarding sites that were indeed from Colombia, but had a domain suffix from another country like India (.in) as was the case for the website “vive.in”.

This proved that even when google does a rather good job at returning the list of websites from Colombia, it was not 100% accurate and included in the results sites that were not from the Colombian Web, including sites like:

- adobe.com
- youtube.com
- w3c.org

Also, the gathered sites returned by google were already presorted by a PageRank[29] Algorithm, making them already the most important ones.

1.5. Difficulties of the characterization of the Web.

The web is a non-centralized collection, in which different authors can contribute content on their own without a control mechanism that decides what is published or not. This is the main advantage of the Web from the point of view of the users, but it is also the main cause of difficulties when search and characterization is needed.

The next anomalies constitute violations to standards or special situations that make it difficult to characterize web pages.

URL parameters and URL Rewriting: there are pages that have longer addresses than what they really should be. This is due to parameter passing in the address as if it were part of the access route, which contradicts the URL[16] standard, because parameters should appear after the "?" symbol, ie:

- Incorrect: http://website/directory/search/word/X/max/10
- Correct: http://website/directory/search?palabra=X&&max=10

This technique is known as URL Rewriting and its use has been extended with the arrival of Content Management Systems (CMS). Among its consequences are: 1) it can not be distinguished whether a page is static or dynamic and 2) several pages are gathered that have the same semantic meaning, given that many of this addresses accept many different parameters to deliver one same page (the identifier, the title, the section inside of the site, the date, etc.). This way, websites appear to have a much larger size than they really do, with more pages per site than average.

Content replication: It is common on the web, that many geographically distributed copies exist of the same documents. Normally what is replicated are complete large collections, and this is done to improve efficiency.

The consequences of this replicated content are websites with a large quantity of text, in the Colombian Web, the replicated content is about 7.50% or 333,820 pages. A manual inspection of the collection shows that there is more duplicated content not detected as such, because the web pages include design which changes, even though the content is still the same. Many other websites duplicate content among them intentionally, and not for efficiency purposes.

Spam in general: spam on the web refers to actions designed to mislead search engines and give some pages a higher ranking than deserved in the result of a query through a web search engine[17]. This actions include changes in the text, metadata or links to pages if the visitor is a harvester robot.

1.6. Structure of this Report

The different possible levels of analysis for the Web are: the smallest, at the level of words or text blocks or images, then pages, sub-sites (coherent units of multiple pages), sites, domains, up to the level of the whole Web of a country and the global Web. In the same way is this report structured, presenting observations of the Colombian Web at various levels: at the level of pages and documents on Section 2, at the level of websites on Section 3 and at the level of Domains on the Section 4. Section 5 presents conclusions, the glossary includes terms used in this document.

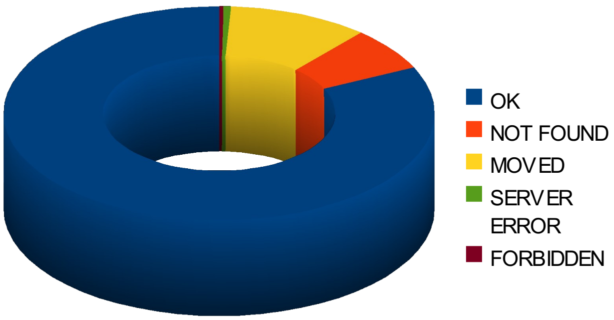
2. Characteristics of the Web pages

In this section the analysis of individual pages is presented, not considering its grouping to its website or domain. First the number of correctly downloaded pages is shown. Then meta data is analyzed, as the URL, title, size, content of the documents and links among them.

2.1. Downloaded pages vs invalid links

The harvester of pages works by extracting addresses of the websites that have been downloaded, and its frequent that among those addresses, are links to pages that no longer exist or that were simply miswritten. Every time the harvester connects to a web server, it receives a code that indicates the state of the page indicating whether the page exists or not, or if there is any other reason why the requested content could not be delivered. Drawing 4 shows the distribution of pages and their status codes. There are many codes, and they are grouped here as:

- *OK*: includes all successful requests: *OK(200)* and *PARTIAL CONTENT (206)*.
- *NOT FOUND*: the server could not find the requested document: *NOT FOUND(404)*.
- *MOVED*: includes all the request for which the server redirects the harvester to another web page: *MOVED(301)*, *FOUND(302)*, and *TEMPORARY REDIRECT(307)*.
- *SERVER ERROR*: includes all the failures on the server side: *INTERNAL SERVER ERROR(500)*, *BAD GATEWAY(502)*, *UNAVAILABLE(503)*, and *NO CONTENT(204)*.
- *FORBIDDEN*: includes all the requests that are not allowed, mainly because those are password protected pages: *UNAUTHORIZED(401)*, *FORBIDDEN(403)*, and *NOT ACCEPTABLE(406)*.



Drawing 4: HTTP Status Code Distribution

In experiments carried out at the CWR[18] successful OK requests were reported to have a probability of occurrence between 75% and 85%.

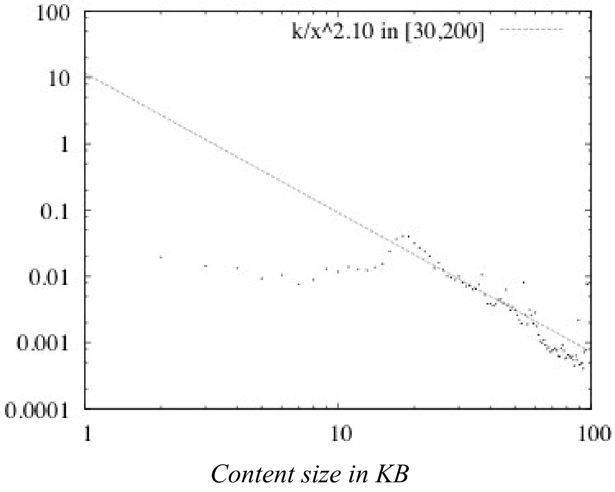
In the Colombian Web, the average of successful OK requests is 72.44% slightly below the lower boundary reported by the CWR[18].

Also the Not Found requests average 5.77% slightly higher than the 4.6% reported by the CWR[18].

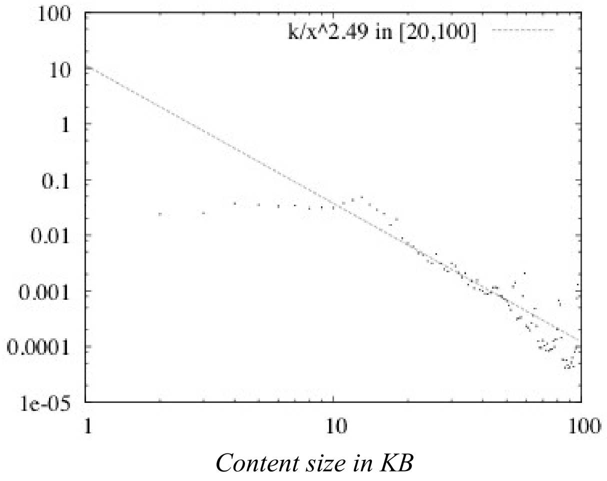
2.2. Text on the pages

From every downloaded page only the first 100kb were stored, this limit was enough for most pages.

Here we graphically show the size of the content of the pages, first only the content of the document, then the complete text (including html tags and code).



Drawing 5: Content Distribution (text with HTML)



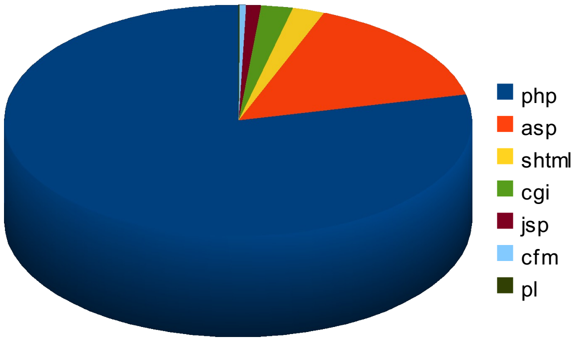
Drawing 6: Content Distribution (text without HTML)

The size of the contents of the documents follows the Zipf law with parameter 2.49, a lower value compared to the one found in Chile[18] and South Korea[19].

2.3. Dynamic pages

More than 1.3 Million pages (30.8%) of the downloaded pages were dynamic, that is, pages that were generated the moment they were requested and that did not exist previously. This is normal when it is required to query a database in the process of responding to requests.

It must be said that many dynamic pages exist that are not detected as such, this is one of the reasons why the percentage is low. It is estimated that the current tendency of having websites whose content is managed online (by using CMS´s) independently from design and structure of the documents, will continue to grow, because it is easier and more practical to have the content of a site in a database rather than HTML files, that are hard to modify either to add, change or remove information. It must also be considered that there are static pages, that have HTML and HTM file extensions, that are generated in batch constantly and automatically by the servers hosting them.

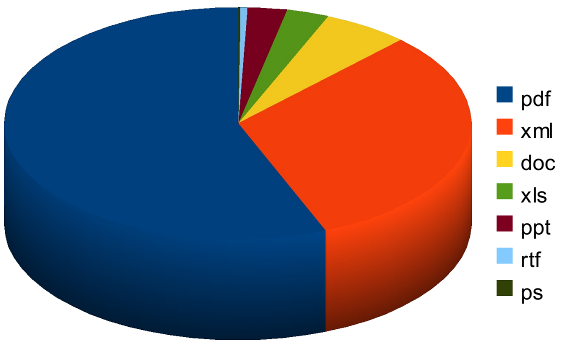


Drawing 7: Distribution of links to dynamic pages

In Drawing 7, the distribution of dynamic pages is shown, according to the application used to generate them. The most used application is PHP[20], an open source technology that dominates the Colombian web with 79% of usage. Its use is slightly higher than in Brazil[21] with 73%, Chile[18] with 75% and vastly superior than in Spain[22] with 46.24%. The ASP[23] technology, a proprietary and of restricted platform follows with 15.65%. In other countries or continents ASP dominates the market, like in South Korea[19] with 75% and Africa[24] with 63%.

2.4. Documents that are not HTML

We found approximately 1 million links to documents in formats different than HTML. The most popular formats are PDF (Acrobat) and XML (considered as SVG, RSS, RDF, XML, etc). Compared to proprietary formats DOC, XLS and PPT, the Open Document Format (the open source alternative) is almost non existent. In drawing 5 the distribution of this documents is shown.

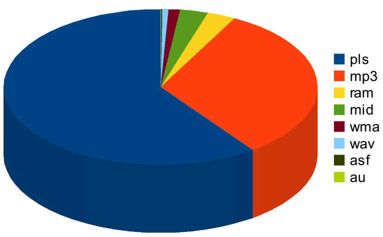


Drawing 8: Distribution of links to documents, excluding HTML pages.

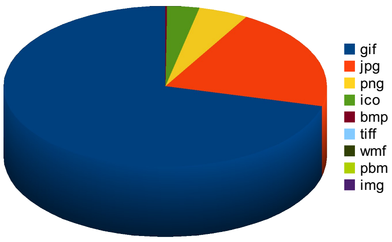
The PDF format is also the most used in other countries, like Austria[25], Brazil[21], South Korea[19], Greece[26], Chile[18] and Portugal[27]. In Spain[22], it is the second most used format with 41.43%.

2.5. Audio, Video and Images

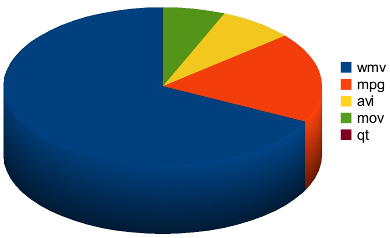
There are many links to multimedia files, more than 65000 links to audio files, less than 10000 links to video files and more than 35 million links to images. The distribution of formats is shown in Drawings 9, 10 and 11.



Drawing 10: Distribution of links to audio files



Drawing 11: Distribution of links to images



Drawing 9: Distribution of links to video files

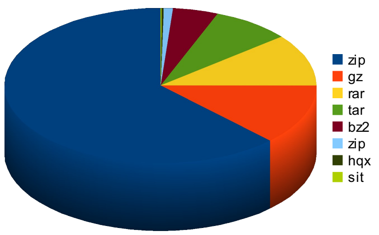
Regarding Audio files, the MP3 format is the most common after the PLS (playlist) format, this might be due the popularity of MP3 players, while other closed formats are not as common or are even disappearing as is the case in Chile[18]. Regarding video files, the closed format WMV (windows media video) with 67% is the clear winner over the rest, the formats MPG and AVI are not as popular, and the FLV format is practically non existent at least in terms of links. It must also be said that the total number of video links is far below the one found in other countries, this could be due to the availability of video streaming services (like youtube), that allow website owners to embed videos hosted somewhere else.

Regarding images, the GIF format is the most popular with 70% of the links. This might be due to the ability of presenting animations, also lossless compression (but allowing only a limited color pallet), and is usually used in community sites for smilies (images that represent a situation, feeling or emotion, like “:)”). JPG files are used mostly to interchange photos or as header images on the sites, with 20% of the links pointing at them. Unfortunately PNG files are not as common as the other formats with less than 5% of the share, in spite of being developed as a replacement to the GIF format.

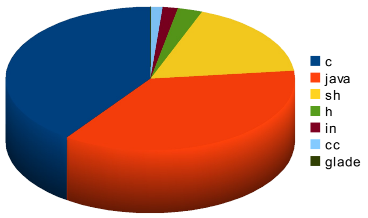
This might be due to a bigger file size than the GIF format and the lack of support in the most popular browser, Internet Explorer of Microsoft in its early versions.

2.6. Software, Source code and Compressed files

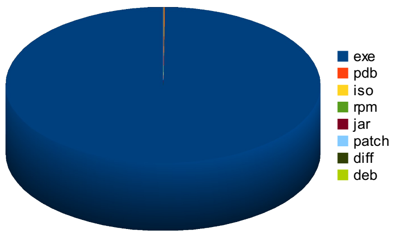
We found more than 680000 links to program files, almost 500000 more links compared to Chile[18] with 180000 links, slightly less than 60000 links to compressed files or 150000 links less compared to Chile[18] with 210000 links, and slightly less than 10000 links to source code files, a third of the links reported for Chile[18] with 35000 links, the distribution of the links is shown in drawings 12, 14 and 47.



Drawing 12: Distribution of links to compressed files



Drawing 13: Distribution of links to source code



Drawing 14: Distribution of links to software

Regarding links to software packages, windows (EXE) has a clear majority with 99.8%, compared to the rest of the links to software packages for other systems for a combined 0.02%.

The distribution of links to source code files, shows C as the dominant language with 40%, closely followed by Java with 36%, and Shell files with 17%, this means that half of the source files is meant or capable of running on Unix environments.

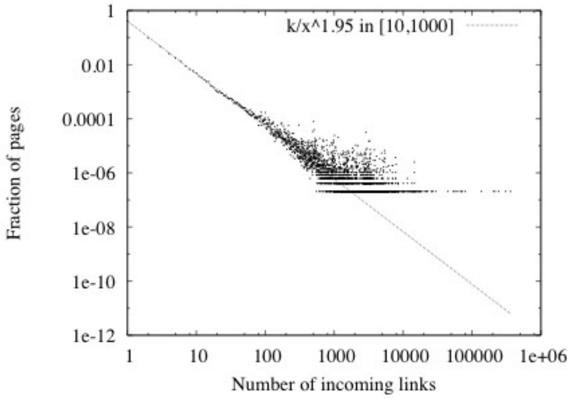
The distribution of compressed files shows that the majority of links point to ZIP files with 62.4%

followed by GZ with 12.65% and then RAR with 10.85% this phenomena shows similarities to what happens with software, given that most of the software is aimed at the windows platform, it is not strange to see that almost 73% of the links points to formats mostly used in that platform (ZIP and RAR), while the percentage of formats used in unix based platforms (GZ, TAR and BZ2) is around 25%, it must also be noted that the later formats are mostly used to distribute source code packages and with a combined number of links of less than 15000, shows a clear dominance of the windows platform.

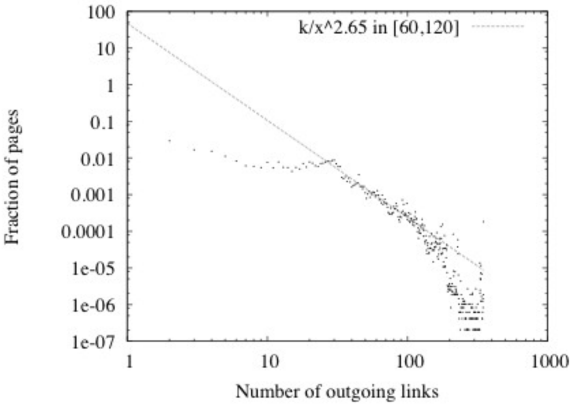
2.7. Links between Web pages

The number of links that a web page receives is called its “internal grade”, a name that comes from analyzing the web as a graph, in the same way, the number of outgoing links are called its “external grade”. The distribution of both grades are shown in drawings 15, 16, 17 and 18.

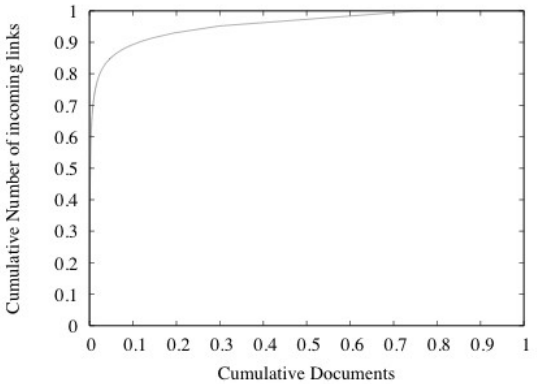
The internal grade of a page is a measure of its popularity in the web, while the external grade indicates the type of page being visited. A commercial page or from a particular brand will try to keep the number of outgoing links low, in order to keep the users in their site. Also, having a page with many links is easy, but receiving links from other pages is rather difficult. Close to 70% of all the documents sum up all the internal grade, while only around 40% of the documents sum app all the external grade.



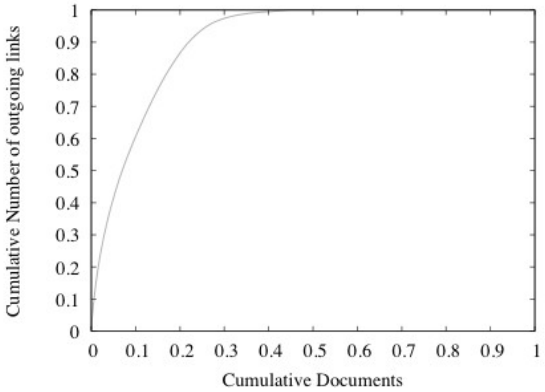
Drawing 15: Internal Grade



Drawing 16: External Grade

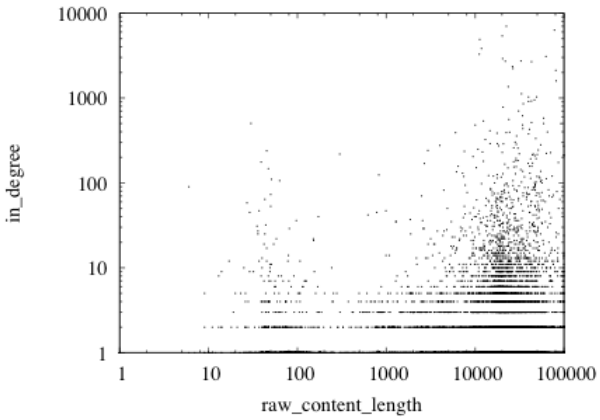


Drawing 17: Internal Cumulative Grade

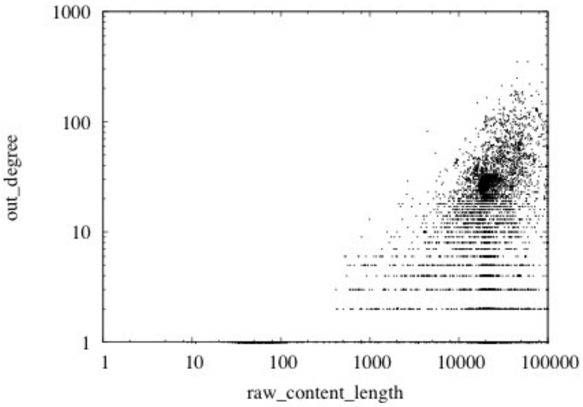


Drawing 18: External Cumulative Grade

Adjusting a Zipf distribution to the data, for the internal grade a parameter of 1.95 is found while for the external grade a parameter of 2.65 is found. Comparing this values to usual parameters[28] of 2.1 and 2.7, the Colombian web is on the average regarding its external grade, but its internal grade is below the average, closer to values found in Africa[24] with 1.9 or Chile[18] with 1.95.



Drawing 20: Size vs Internal Grade



Drawing 19: Size vs External Grade

In Drawings 19 and 20, the relation between the size of the pages and the grades is shown.

There exist a correlation between the size of the page and the external grade, because a small page can only have a few outgoing links specially if it is tiny, there is no clear relationship between size and internal grade, but it is possible to see that the smaller pages receive less links from the outside.

2.8. Ordering using link analysis algorithms

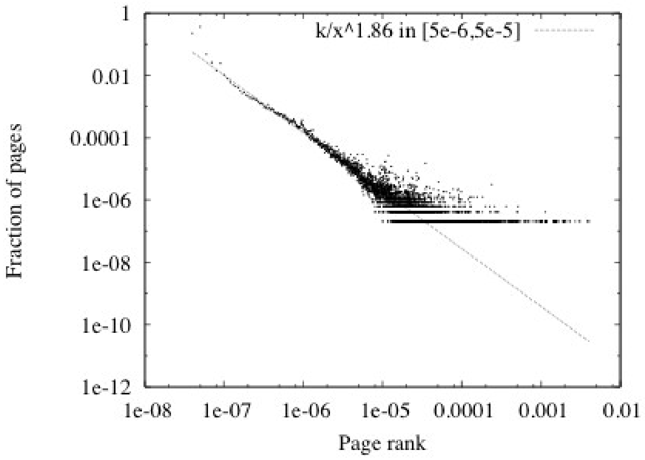
There are several link algorithms that try to infer how important is every page on the web, using the information of the links that each page receives. Comparing the distribution of *Pagerank*[29] with a variation of the algorithm *HITS*[30], in which the complete Web is used as the set to be analyzed. The later can be seen as a static version of HITS.

The Pagerank algorithm calculates for each page a score that reflects the quantity of links it receives from other pages also with a high link count. In a sense, it is a measure of quantity and quality of the received links.

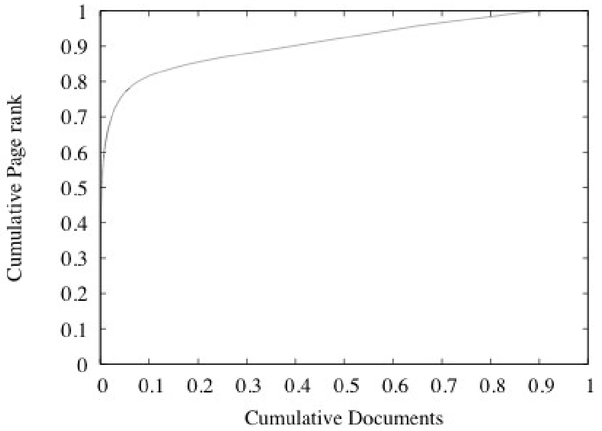
The HITS algorithm calculates two scores for each page: *Hub* and *Authority*. The Hub score shows how good a page is, in terms of how good are the links that the page has to other pages. The Authority score shows good a page is, in terms of how good are the links that it receives.

Because of the way the Pagerank algorithm is calculated, in which random values are introduced in the calculation (it considers that with a small probability, a page can be reached by chance), even pages with few incoming links have a Pagerank score higher than zero. Analyzing the scores, 85% of the pages accumulates 100% of the Pagerank.

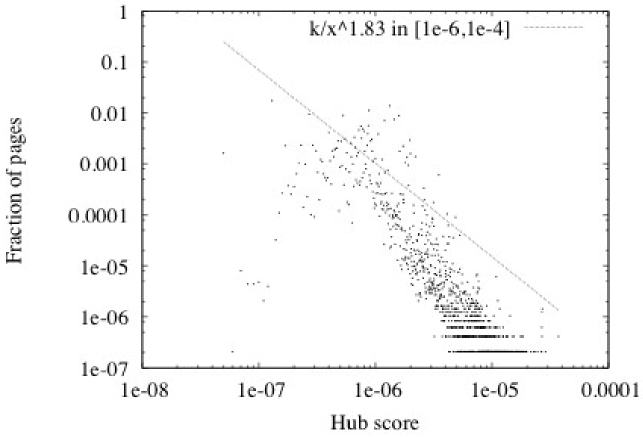
The scores can be seen in Drawings 22, 21, 23, 24, 25, 26.



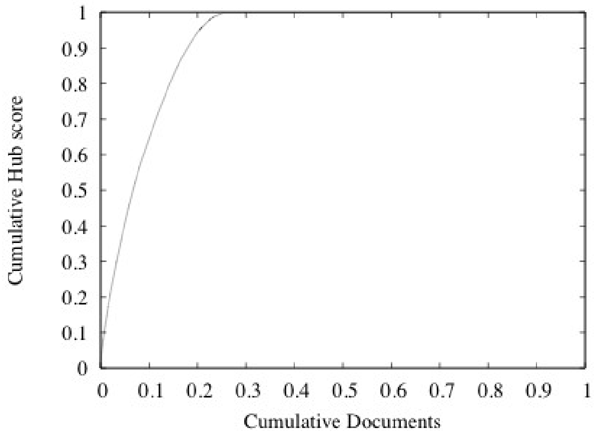
Drawing 22: PageRank



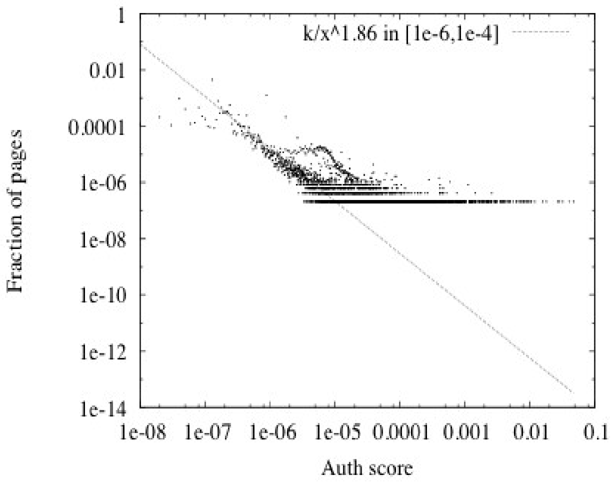
Drawing 21: Cumulative PageRank



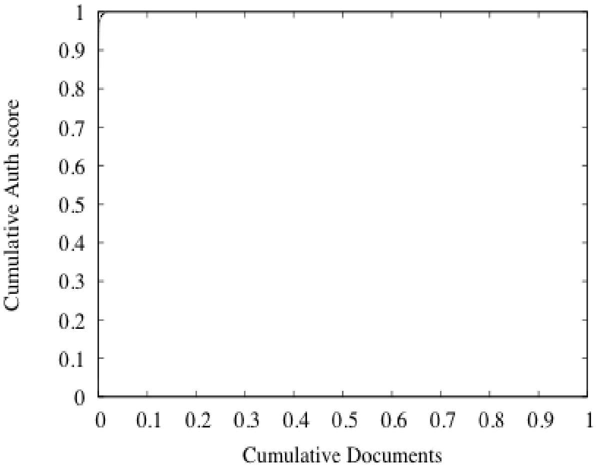
Drawing 23: Hub Score



Drawing 24: Cumulative Hub Score

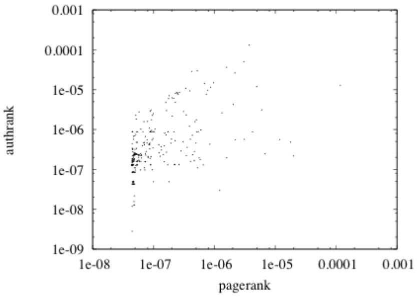


Drawing 25: Authority Score

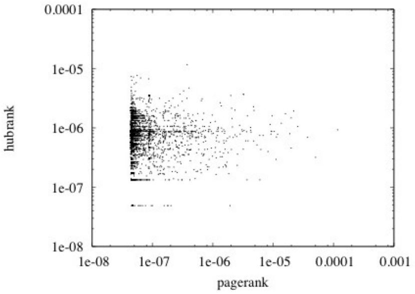


Drawing 26: Cumulative Authority Score

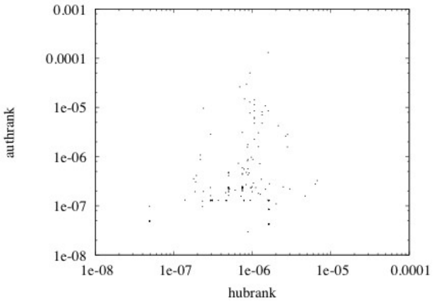
In contrast, a page needs quality links in order to have Hub and Authority scores different than zero, this way only 25% of the pages have a Hub score higher than zero and only 2% have an Authority score higher than zero.



Drawing 27: PageRank vs Authority score



Drawing 28: PageRank vs Hub score



Drawing 29: Hub score vs Authority score

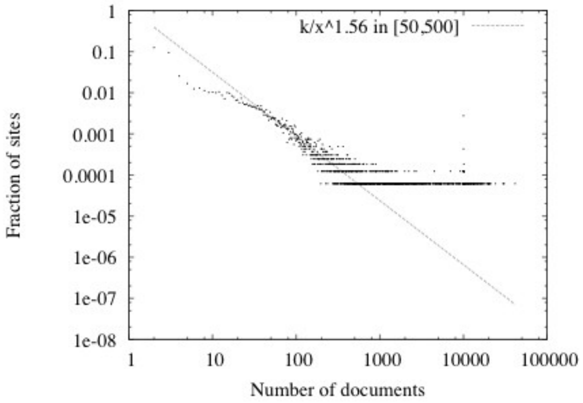
From a random set of 10.000 documents with scores higher than zero, we see no significant correlation between the link analysis algorithms, PageRank, Authority score and Hub score as shown in Drawings 27, 28, 29.

3. Characteristics of the Websites

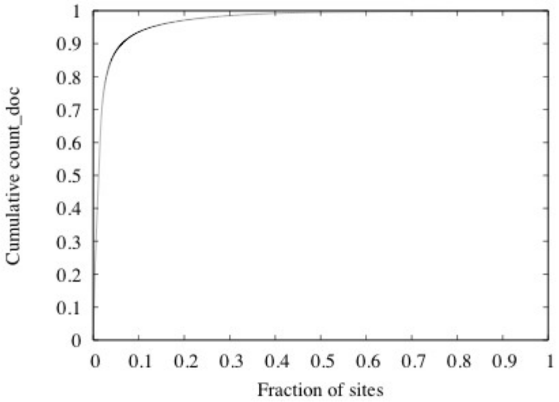
We define a website as a set of pages that share the server part of the URL[16]. Besides using the heuristic that http://www.site.co and http://site.co map to the same website[1].

3.1. Number of pages

There are on average 393 pages per site. The distribution of the number of pages per site is shown on Drawings 30 and 31.



Drawing 31: Distribution of documents per site



Drawing 30: Cumulative number of documents per site

The distribution is rather skewed, only 10% of the sites have 90% of the documents. There are many sites with few pages, which can be a sign of the low development of the Web. Comparing the data with the Zipf law, we get a parameter of 1.56 lower than the one found in Chile[18] with 1.74 or South Korea[19] with 2.5, higher than the one found in Spain[22] of 1.14 and similar to the one found in Brazil[21] with 1.6.

[1] Generally it is that way, there are even initiatives for stopping the www prefix usage on the web, some search engines allow webmasters to chose whether they want the website indexed with or without the prefix.

3.2. Sites with only one page

There are 4015 sites with only one webpage, that is 24.74% of the sites. A parameter not so high compared to the one found in Spain[22] of 60%. Among the possible reasons for this we have found:

- The browsing of the website is based on Javascript, and therefore it is necessary to interpret the code to navigate.
- The website is just a redirect to another website, either using the "Refresh" label or having one link to the other site.
- The website indeed has only one site.
- The page requires flash in order to visualize/navigate it. It is common among websites to have an introductory animation to the site, without really using flash to show content. This way many sites that are "normal", do not get indexed by search engines because of the lack of a "skip introduction" HTML link.
- The page contains only external links.
- The page contains internal links but they are malformed and the collector was not able to interpret them.
- The page uses Java Applets to handle the navigation.

3.3. Sites with many pages

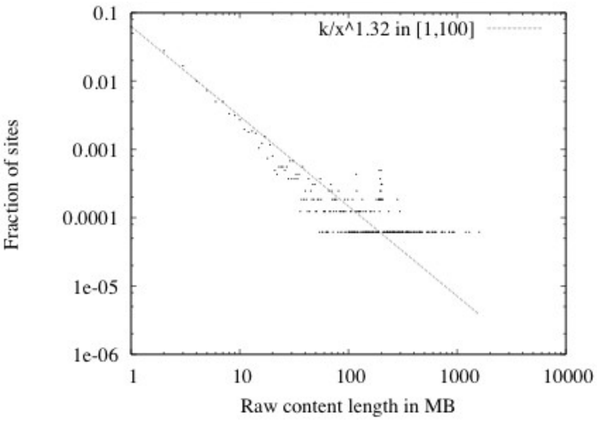
We also analyzed sites that have many pages. The top 30 sites with the most documents is shown in Table 1. Normally those sites are using CMS (Content Management System) that offer services like blogs, forums, image galleries. Current CMSs allow the usage of *URL Rewriting* to retrieve pages and even the usage of different parameters can lead to the same page. Besides that, there are also links to different internal parts of the document (like comments to a blog post, or different opinions in a forum), which create recursion in the pages. These systems do not have a static design (ie, a document may have links to other pages which get delivered with different dates) which makes it difficult to detect duplicated documents.

Website	pages	Comment
www.anuncol.com	41370	CMS, parameters in URL
www.industrialtaylor.com.co	30534	CMS, parameters in URL, malformed URL
www.freddyvera.com	29946	CMS, parameters in URL, malformed URL
biblos.javeriana.edu.co	29818	CMS, parameters in URL
foros.hispavista.com.co	24144	CMS, parameters in URL
www.paginasamarillas.com	21220	CMS, parameters in URL
www.i-local.org	20469	CMS, malformed URL
colombianpaintball.com	20402	CMS, parameters in URL
www.mitiempoextra.com	20330	CMS, parameters in URL
www.clinicalasvegas.com	20294	CMS, parameters in URL, malformed URL
lanota.com	20193	CMS, parameters in URL
www.yoquieroir.com	20135	CMS, parameters in URL
www.veoyalquilo.com	19976	CMS, parameters in URL
www.loteriadeltolima.com	19974	CMS, parameters in URL, malformed URL
www.pngbd.com	19944	CMS, parameters in URL, malformed URL
www.ingeominas.gov.co	19900	CMS, parameters in URL, malformed URL
www.tiendadecomputadores.com	19264	CMS, parameters in URL
www.hinchadaverde.com	19074	CMS, parameters in URL, malformed URL
www.bodytech.com.co	18248	CMS, parameters in URL, malformed URL
www.dinero.com	17537	CMS, parameters in URL
www.babillacine.com	17487	CMS, parameters in URL
www.bandolitis.com	17192	CMS, parameters in URL
www.cvxcol.org	16656	CMS, parameters in URL
www.empresarioccibague.com.co	16493	CMS, parameters in URL, malformed URL
www.colegioamericano.edu.co	16295	CMS, parameters in URL, malformed URL
www.colegiounidadpedagogica.edu.co	16197	CMS, parameters in URL
www.comfamiliar.com	16128	CMS, parameters in URL
www.unbosque.edu.co	16114	CMS, parameters in URL
www.fundacionartedevivir.org	16076	CMS, parameters in URL
redeparede.com.co	16011	CMS, parameters in URL

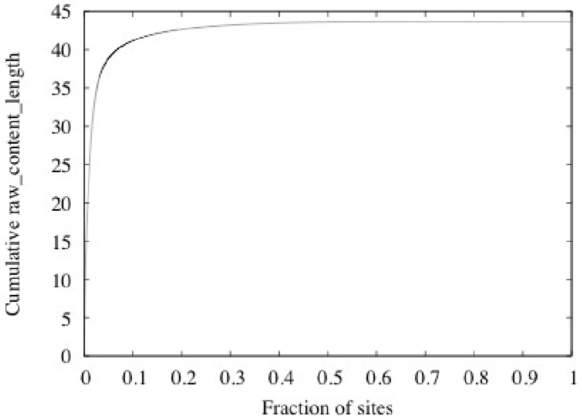
Table 1: Top 30 websites by pages

3.3. Size of the pages of a complete Website

In this section we analyze only the text of the collected pages, that is, in order to find the size of a website only the size of the HTML documents is taken into account, not the size of the images or any other documents or multimedia files.



Drawing 33: Size of the website



Drawing 32: Cumulative size of the website

In Drawings 32 and 33 the distribution of the size of the sites is shown, again the distribution is very skewed.

The distribution is adjusted to the Zipf law with parameter 1.32 for a size of up to 10 MB.

Table 2. shows the top 30 sites with the most text. It can be seen that there is a high usage of CMSs and that most offer either products, services or information (as forums, indexes, etc.), there is also a big amount of replication because of the usage of dates in the URL.

Website	Text MB	Comment
www.freddyvera.com	1670	Forum
www.tiendadecomputadores.com	1356	Products Catalog
foros.hispavista.com.co	1258	Forum
www.unbosque.edu.co	982	University
tuguiadeviajes.blogspot.com	973	Services Catalog
www.mitiempoextra.com	940	
www.anuncol.com	891	Advertisement
www.industrialtaylor.com.co	882	Products Catalog
www.polemiza.com	803	Default empty site
www.clinicalasvegas.com	769	Clinic
www.colegioamericano.edu.co	756	High school
www.loteriadeltolima.com	754	lotto
www.babillacine.com	672	Movies
www.empresarioccibague.com.co	657	Chamber of Commerce
www.ccpalmira.org.co	651	Chamber of Commerce
atajos.lapapa.com.co	629	Products Catalog
www.elvallenato.com	621	Forum
www.colegiounidadpedagogica.edu.co	611	High school
zonasite.com	573	
biblos.javeriana.edu.co	572	University
www.ingeominas.gov.co	565	Government
guia.hispavista.com.co	549	Forum
lanota.com	500	
www.rescateksar.org	493	Rescue/Search Dogs
www.factoringmarket.com	487	Finance
cafeinternet.com.co	474	Forum
www.revistalabarra.com.co	470	Magazine

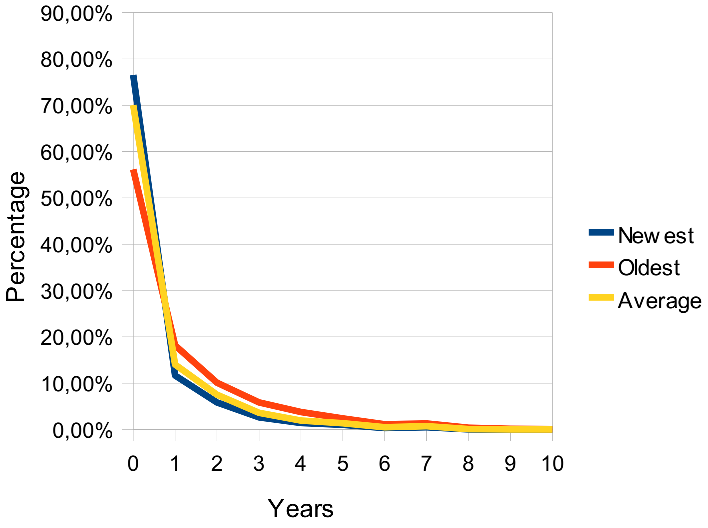
Table 2: Top 30 sites by size

3.4. Age

We measure the age of the websites, tracking the age of the oldest page, the age of the most recent one and the average. The age of the oldest page indicates a lower boundary of how old the website is, while the age of the newest page indicates when was the last time the website was updated.

From the data, we find that 76% of the sites were created in the last year and 88% were created in the last two years. This indicates that the Colombian web is growing at a very fast pace.

The results can be seen on Drawing 34.



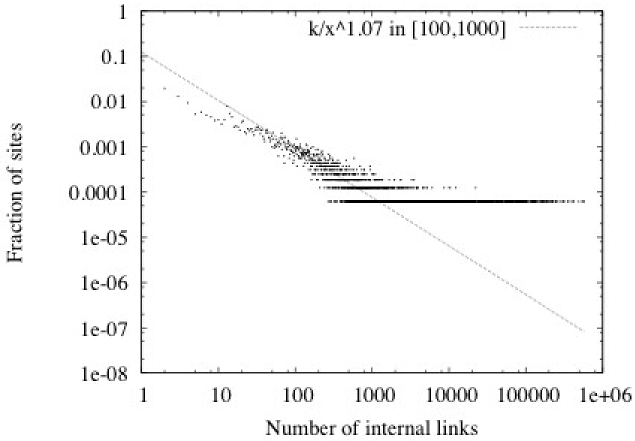
Drawing 34: Age of the websites

3.5. Internal Links

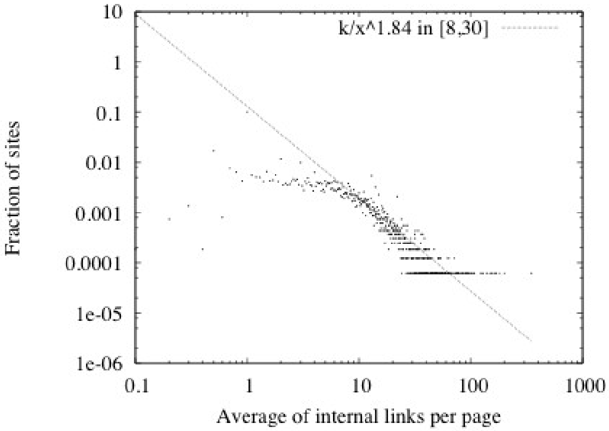
A link is considered as internal if it points to another page in the same website. An average site has 3164 internal links, and on average a page has 8 internal links. Besides this, there are some sites with a lot of internal links.

The distribution of the number of internal links per site is shown on Drawing 36.

This distribution is related to the distribution of pages per website, because a website with a low count of pages, can not have many internal links. However looking at the distribution of internal links, there does not seem to be an important correlation, as shown on Drawing 35. Measuring the distribution of internal links per page we find it follows the Zipf law on the central part with parameter 1.07.



Drawing 36: Distribution of the number of internal links



Drawing 35: Distribution of the number of internal links per page

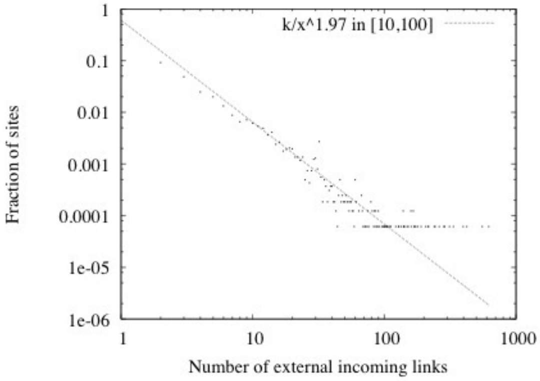
3.6. Links among Websites

Now we consider the links among websites, these are links between pages of different websites.

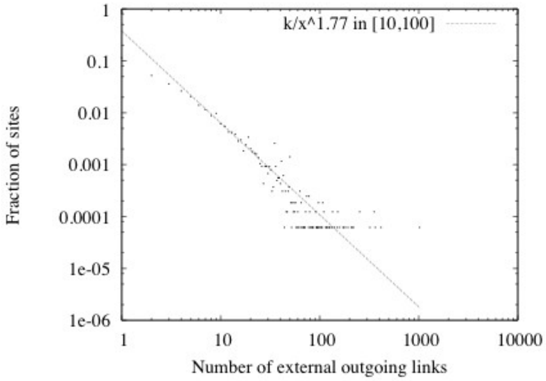
That is, if we have at least a link between say http://siteA.co/PageA.htm and http://siteB.co/PageB.htm, then we consider it a link between the two websites siteA.co and siteB.co (the internal links are not taken into account). this is also called the Hostrank or server graph[31].

There are 12,163 websites with more than one page, of those 3,392 have no incoming links from any other website in Colombia and 6,254 have no outgoing links to any other website in Colombia.

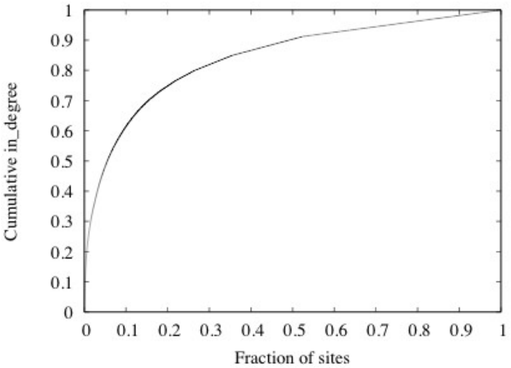
The distribution of the internal and external grade of the sites, also reveals a network free of scale, as shown on Drawings 37, 38, 39, 40.



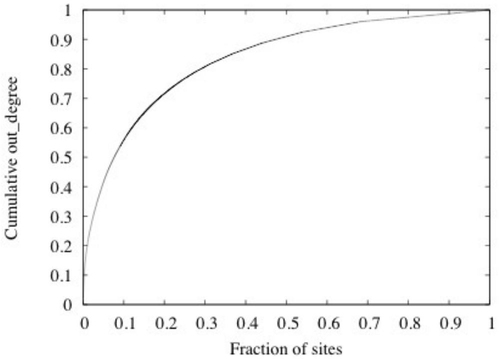
Drawing 37: Internal Grade



Drawing 38: External Grade



Drawing 39: *Internal Cumulative Grade*



Drawing 40: *External Cumulative Grade*

The parameters of adjustment to the Zipf law are 1.97 for the internal grade and 1.77 for the external grade, this can be compared to grades like Chile[18] (1.99, 1.91), Brazil[21] (1.9, 1.9), Greece[26] (2.0, 1.6) and Spain[22] (1.8, 1.3). it is estimated that the global[31] web has an internal grade of 2.34.

3.7. Most referenced Websites

The top 35 most referenced sites are shown on Table 3, all the websites that point towards a specific site are counted.

Because of the heuristic used on the collection of the initial seed of URLs, we see at the top websites that do not belong to the Colombian Web.

Site	**Links**
www.adobe.com *	620
www.youtube.com *	554
validator.w3.org *	423
www.macromedia.com *	397
jigsaw.w3.org *	334
www.colciencias.gov.co	312
www.contratos.gov.co	288
www.unal.edu.co	283
www.univalle.edu.co	261
www.icetex.gov.co	243
www.universia.net.co	239
www.uniandes.edu.co	217
www.mineducacion.gov.co	214
www.banrep.gov.co	206
www.presidencia.gov.co	190
www.udea.edu.co	179
horalegal.sic.gov.co	173
www.icfes.gov.co	170
javeriana.edu.co	168
www.geocities.com *	166
www.dnp.gov.co	166
www.colombiaaprende.edu.co	160
www.minproteccionsocial.gov.co	159
www.elespectador.com	159
www.minambiente.gov.co	152
www.mincomunicaciones.gov.co	149
www.colnodo.apc.org	142
www.dane.gov.co	141
www.lablaa.org	140
www.eafit.edu.co	139
biblioteca.univalle.edu.co	139
www.bogota.gov.co	134
www.mincultura.gov.co	129
www.sena.edu.co	128
www.semana.com	122

Table 3: Most Referenced Sites

** the site does not belong to the Colombian Web*

3.8. Sites with the most number of links

The 35 sites that have the most links are shown on Table 4, among them there does not seem to be an absolute majority of a particular type of site. There are directories, services, universities, community sites. Also we find the always common products and services catalogs.

Site	Links
www.encuentromedellin2007.com	575,579
www.revistalabarra.com.co	558,234
economia.uniandes.edu.co	540,125
www.ddhhcolombia.org.co	489,412
www.mitiempoextra.com	420,603
www.imageninvisible.org	414,836
www.ccpalmira.org.co	396,262
m3lab.encuentromedellin2007.com	358,861
cafeguaguau.com	356,568
gcn.mincultura.gov.co	315,103
www.asopadrescomfenalco.com	314,834
www.loteriadeltolima.com	309,113
www.deltaasesores.com	289,911
www.gerencie.com	276,322
comunidad.wilkinsonpc.com.co	256,721
www.colombialink.com	248,233
www.observatoriodejuventud.org	244,819
www.supernotariado.gov.co	242,606
www.cvxcol.org	227,642
www.colombiaaprende.edu.co	224,713
www.newmanschool.edu.co	224,637
www.sealedair.com.co	220,020
www.vanguardia.com	219,899
www.clinicalasvegas.com	214,364
comerciocaqueta.com	207,919
www.cirugiaplasticacolombia.com	202,620
www.funiber.org *	196,837
cafeinternet.com.co	193,513
www.estereofonica.com	190,457
jpnascar.com	188,067
colombiamania.com	187,107
www.fotografiacolombiana.com	182,145
www.pezplata.com	175,845
www.museos.unal.edu.co	175,269
www.dalailamacolombia.com	175,198

Table 4: Top 35 Sites by number of links

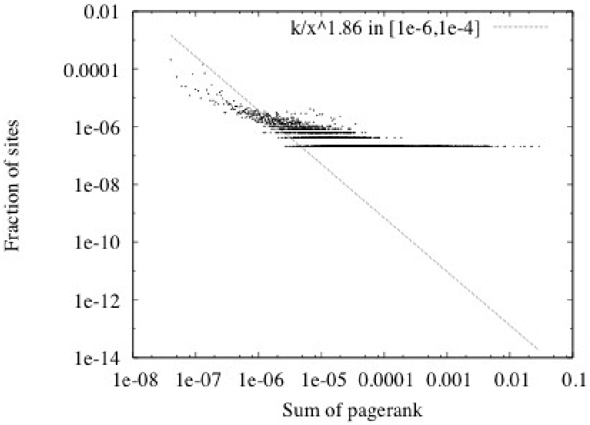
** does not belong to the Colombian Web*

3.9. Sum of the scores by links

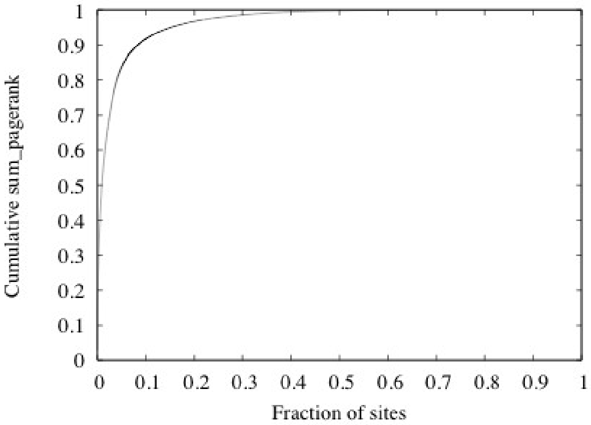
Studying the scores shown in Drawings 22, 21, 23, 24, 25, 26 and adding them by websites, we find a measure of the quality of the site. The results are shown on Drawing 41, 42, 43, 44, 46, 45.

An important note on the found data is that the best pages of the Colombian Web are distributed among many websites.

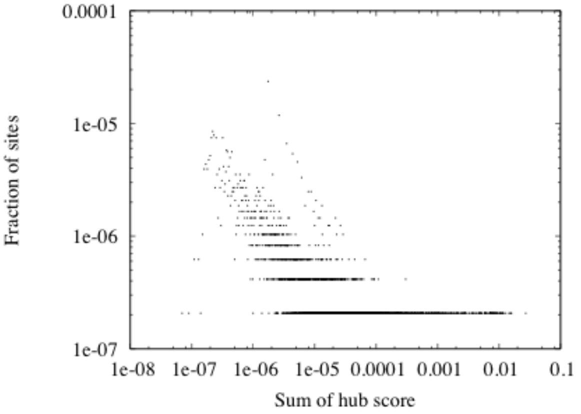
Besides that, the distribution of the PageRank follows the Zipf law, with a parameter of 1.86.



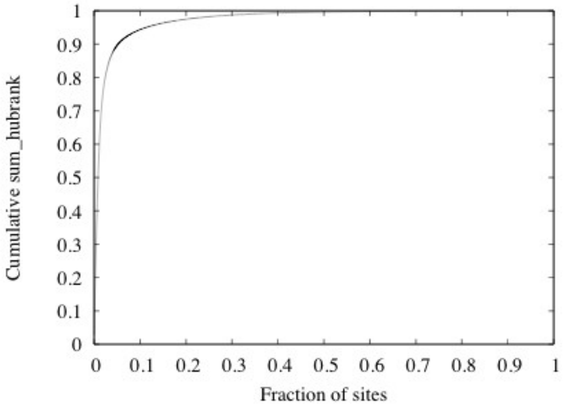
Drawing 41: Sum of PageRank



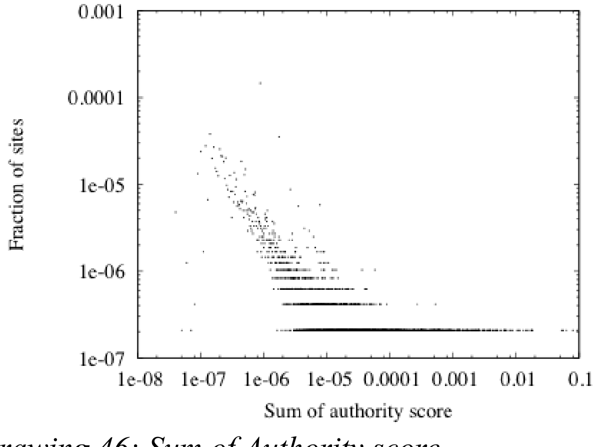
Drawing 42: Site cumulative of the sum of Pagerank



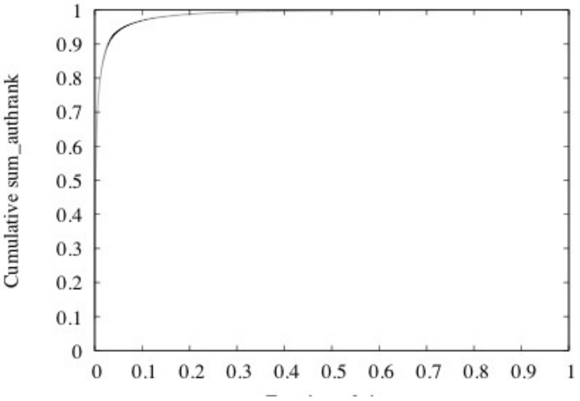
Drawing 43: Sum of Hub score



Drawing 44: Site cumulative of the sum of Hub score



Drawing 46: Sum of Authority score



Drawing 45: Site cumulative of the sum of Authority score

3.10. Strongly connected components

One of the basic components of graph theory is connectivity, it can be said that a part of a graph is connected if there is a path from any node to any other node inside that part of the graph. In a graph there can also be strongly connected components, that is, a connected part of the graph in which all the nodes that are connected, can be reached by strictly following the direction of the paths. Not all the Colombian Web is strongly connected.

Studying the distribution of the sizes of all the strongly connected components in a graph of the websites, we find a giant strongly connected component, as it was observed by Broder and others[32] this is a typical sing of a free of scale network.

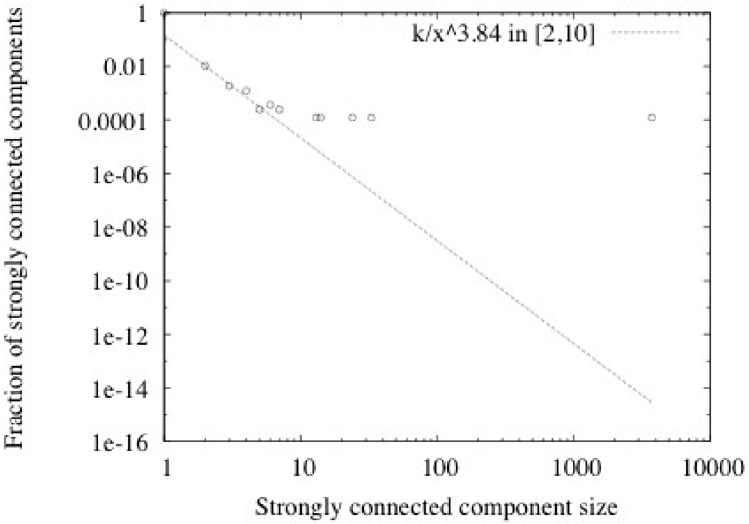
The distribution of the sizes of the strongly connected components is shown in Table 5.

A website is considered to have a component size of 1 if it has at least one incoming or one outgoing link. The strongly connected component corresponds to 46.57% of the nodes, around three times higher compared to Chile[18] with 14.03% or Spain[22] with 15.1% or South Korea[19] 15.1%. This difference mainly arises because of the subset of initial URLs used was already belonging to an at least highly connected component, and many Islands and lowly connected components were not seen.

Size of the SCC	Number of components
1	8040
2	84
3	15
4	10
5	2
6	3
7	2
13	1
14	1
24	1
33	1
3744	1

Table 5: Size of the Strongly Connected Components

When the sizes are represented graphically a Zipf law is observed with parameter of 3.84 similar to the one found in Spain[22] of 3.84, and also comparable with the ones found in Chile[18] of 3.4, South Korea[19] of 2.6, Greece[26] of 4.20 and 2.81 of the Global Web[31].



Drawing 47: Strongly Connected Component Size

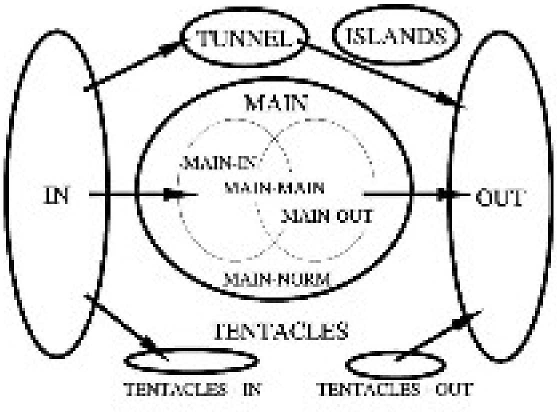
3.11. Structure of links among Websites

The strongly connected component seen on Table 5, can be used as starting point to distinguish several components of the Web. These were defined by Broder and others[32] as:

- *MAIN*, the sites on the strongly connected component.
- *OUT*, sites that are reachable from *MAIN*, but have no link towards *MAIN*.
- *IN*, sites that can reach *MAIN*, but have no links from *MAIN*.
- *ISLANDS*, sites not accessible either to or from *MAIN*.
- *TENTACLES*, sites only connected with *IN* or *OUT*, but in reverse direction to the links.
- *TUNNEL*, a component that links the *IN* or *OUT* components, but not going through *MAIN*.

In [33] the notation was extended, distinguishing in the *MAIN* part the following components:

- *MAIN-MAIN*, the sites that are reachable directly from *IN*, or that can reach *OUT* directly.
- *MAIN-IN*, sites that are reachable directly from *IN* but are not in *MAIN-MAIN*.
- *MAIN-OUT*, sites that can reach *OUT* directly, but are not in *MAIN-MAIN*.
- *MAIN-NORM*, sites not belonging to the previously mentioned categories.



Drawing 48: Macroscopic structure of the Web

The distribution of the websites in components is shown on table 6. The websites on the components IN and ISLANDS can only be found if their address is previously know, because they are not reachable following links. Also in this table the percentage of pages and the distribution of sites in components by its domain is also shown.

	Total of Sites	Only with Links	Pages	Internal Links	CO	EDU	COM	ORG
IN	11.94%	9.00%	21.62%	22.37%	24.38%	0.34%	58.75%	16.53%
ISLAND	19.88%	1.45%	8.14%	5.78%	12.82%	0.21%	74.90%	12.03%
OUT	30.91%	7.66%	6.15%	4.40%	82.61%	0.13%	12.66%	4.55%
TIN	4.96%	0.49%	0.73%	1.05%	71.64%	0.00%	25.21%	3.15%
TOUT	1.13%	0.53%	1.47%	0.75%	26.81%	0.00%	60.87%	12.32%
TUNNEL	0.39%	1.02%	0.06%	0.10%	72.92%	0.00%	25.00%	2.08%
MAIN_MAIN	6.40%	16.59%	22.96%	27.11%	73.26%	0.64%	18.77%	7.33%
MAIN_NORM	10.76%	27.92%	6.17%	7.25%	77.23%	0.61%	14.21%	7.94%
MAIN_OUT	10.45%	27.11%	29.12%	27.11%	72.93%	0.63%	14.08%	12.35%
MAIN_IN	3.17%	8.23%	3.58%	4.08%	66.32%	0.26%	23.06%	10.36%
MAIN	30.78%	79.85%	61.83%	65.55%	74.09%	0.47%	18.02%	7.43%

Table 6: Distribution of Sites by components and domains

4. Characteristics of the domains

The domain of a page is defined as the suffix of its name on the web, following the next rule:

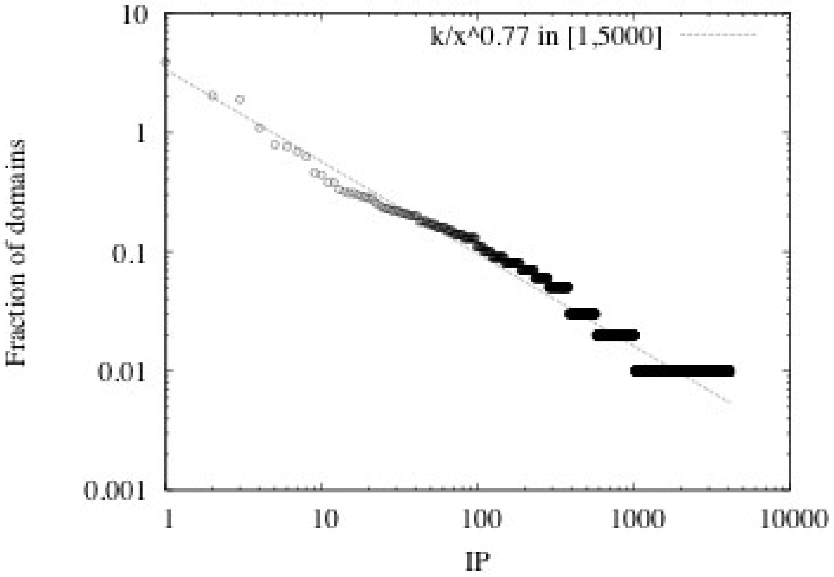
- if the address of a website is of the form www.A.co and www.B.A.co, then the domain is A.co

In total 11245 domains were found.

4.1. IP address and hosting provider

We did DNS lookups on the website addresses of each one of the studied domains, being able to contact 77.45% of them. The sites that could not be contacted are very likely non existent anymore.

We grouped the IP addresses by Domains, in order to count how many domains use the same IP. The Distribution of the number of domains by IP is shown on figure 49.



Drawing 49: Distribution of the number of domains by IP

In total there are around 4135 IP addresses for all the domains. This means that every address has on average 2.7 domains, the distribution does not follow a Zipf law because the adjustment parameter was of 0.77 lower than the minimum of 1.

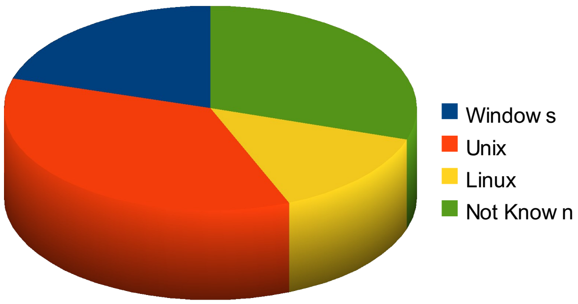
4.2. Web server software

For each IP address we find out what software is used for the web server and what operative system is being used. This was done using an HTTP HEAD requirement which asks only for the header of the initial page of the site. A typical answer has the form:

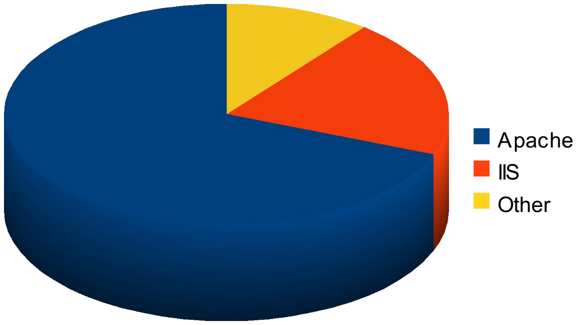
HTTP/1.1 200 OK

Server: Apache/1.3.33 (Debian GNU/Linux) PHP/4.3.10-9 mod_ssl/2.8 …

In some cases (as in the example), the information gathered is rather complete, including the name of the server (Apache), the version (1.3.33), and operative system (Linux) also including the installed extensions (PHP and ModSSL). The distribution of the operative systems is shown on figure 50.



Drawing 50: Operative System



Drawing 51: Web Server

The dominant web servers are Apache followed by Microsoft IIS (Internet Information Server), with Apache having more than two thirds of the market (with 69.05%) and IIS (with 20.33%) barely doubling the installed based of the other web servers (10.61%).

This distribution[34] follows quite precisely the global trend found on 2006 where Apache had a market share of 69% and IIS 21%, the current trend is lower for Apache with 45.95% and IIS with 29.27% as of 2009.

Regarding the Operative System, Unix/Linux have around 49.71% while Windows has around 20.33% of the share but there is another 29.96% of the sites hosted where the Operating Systems information is not delivered therefore it can not be clearly determined which one has a bigger market share. If the unknown sites follow the same distribution of the ones known then it can be said that Windows has a lower penetration rate compared to open source alternatives or commercial Unixes.

This is comparable to Chile[18] where Unix/Linux has 31% of the market and Windows 20% with an unknown range of 48% and also comparable to Spain[22] where Windows has 43% of the market and Unix/Linux has 41%.

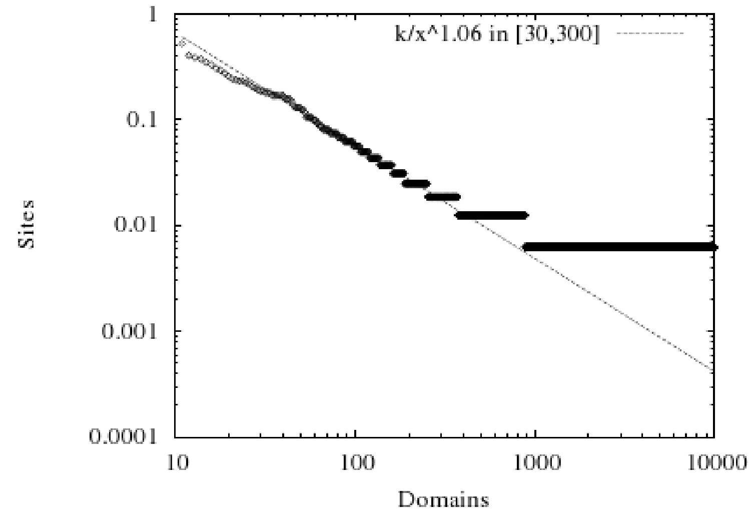
4.3. Number of sites per Domain

On average we find that there are 1.44 sites per domain. There are 10356 Domains with only 1 site, although there are several domains that have many more sites than the average. The distribution of the number of sites per domain is shown on Drawing 52.

The Top 30 domains with more sites are shown on Table 7. Many are domains of universities or government related.

Domain	Sites	Percentage of sites
univalle.edu.co	297	1.83%
uniandes.edu.co	263	1.62%
unal.edu.co	226	1.39%
udea.edu.co	167	1.03%
boyaca.gov.co	123	0.76%
cundinamarca.gov.co	119	0.73%
antioquia.gov.co	116	0.71%
terra.com.co	92	0.57%
unicauca.edu.co	91	0.56%
quebarato.org	86	0.53%
santander.gov.co	84	0.52%
comunidadcoomeva.com	65	0.40%
coomeva.com.co	62	0.38%
narino.gov.co	60	0.37%
puj.edu.co	56	0.35%
javeriana.edu.co	53	0.33%
eafit.edu.co	49	0.30%
tolima.gov.co	47	0.29%
bolivar.gov.co	44	0.27%
nortedesantander.gov.co	41	0.25%
cauca.gov.co	39	0.24%
evisos.com.co	38	0.23%
valle.gov.co	37	0.23%
uniminuto.edu	37	0.23%
huila.gov.co	36	0.22%
quebarato.com.co	35	0.22%
atlantico.gov.co	33	0.20%
unalmed.edu.co	32	0.20%
unisabana.edu.co	31	0.19%
mercadolibre.com.co	30	0.18%

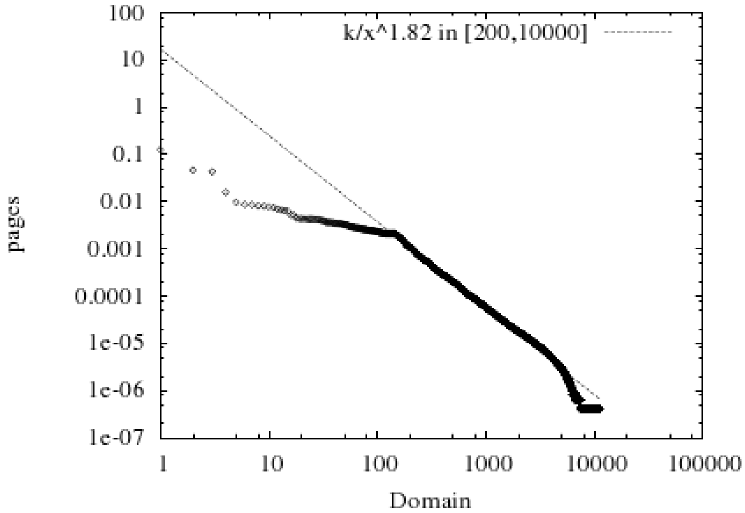
Table 7: Top 30 Domains by number of sites



Drawing 52: Sites per Domain

4.4. Number of pages per domain

On Average there are 429 pages per domain. All the domains have at least 2 pages and there are 3720 domains of this size or 33% of the total number of domains, comparable to the 21% found in Chile[18]. The distribution of the number of pages per domain is rather skewed and it is shown on Drawing 53, it follows a Zipf distribution on its central part with parameter 1.82, comparable to the one found in Chile[18] of 1.67 and Spain[22] of 1.18.



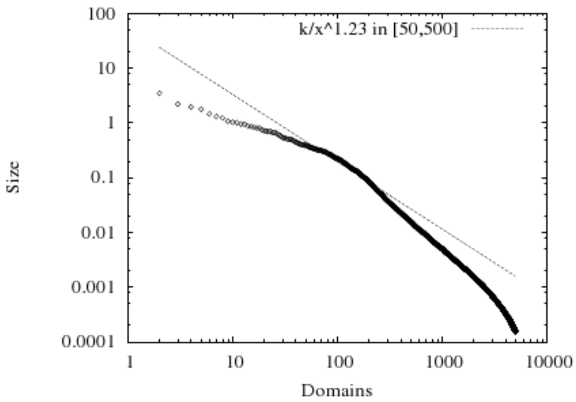
Drawing 53: Distribution of pages per Domain

4.5. Total size of the Domains

On average the domains have a size of 7 MB, this is due to the fact that many sites have a certain amount of repeated content because of the CMS's. The distribution of the domains and their sizes is shown on Drawing 54. The top 30 domains by size are shown on table . Following the same behavior observed in Chile[18], many of these domains are commercial and online auctions, there are also some universities that make it to the list.

Domain	Size in MB
quebarato.org	11600.34
quebarato.com.co	3099.38
mercadolibre.com.co	1960.49
hispavista.com.co	1725.27
freddyvera.com	1593.01
tiendadecomputadores.com	1293.65
blogspot.com	1163.9
unal.edu.co	1092.76
unbosque.edu.co	958.91
lapapa.com.co	912.17
mitiempoextra.com	896.79
anuncol.com	850.57
industrialtaylor.com.co	841.24
javeriana.edu.co	773.64
polemiza.com	765.98
clinicalasvegas.com	734.06
colegioamericano.edu.co	721.68
loteriadeltolima.com	719.21
adoos.com.co	657.24
babillacine.com	640.87
empresarioccibague.com.co	626.94
ccpalmira.org.co	620.97
encuentromedellin2007.com	620.11
elvallenato.com	593.11
ingeominas.gov.co	588.92
colegiounidadpedagogica.edu.co	582.89
zonasite.com	547.2
terra.com.co	532.59
udea.edu.co	506.58
lanota.com	477.31

Table 8: Top 30 Domains by Size in MB



Drawing 54: Distribution of the size of the Domains

4.6. Top level Domains

In the collection of websites in Colombia, there are many sites with the country domain (.co) but there are also many others with different top level domains (.com, .org, .net) on Table 9 the distribution of the suffixes is shown.

The number of .com domains is quite high compared to Chile[15], where the national domain (.cl) has around 99% of the distribution.

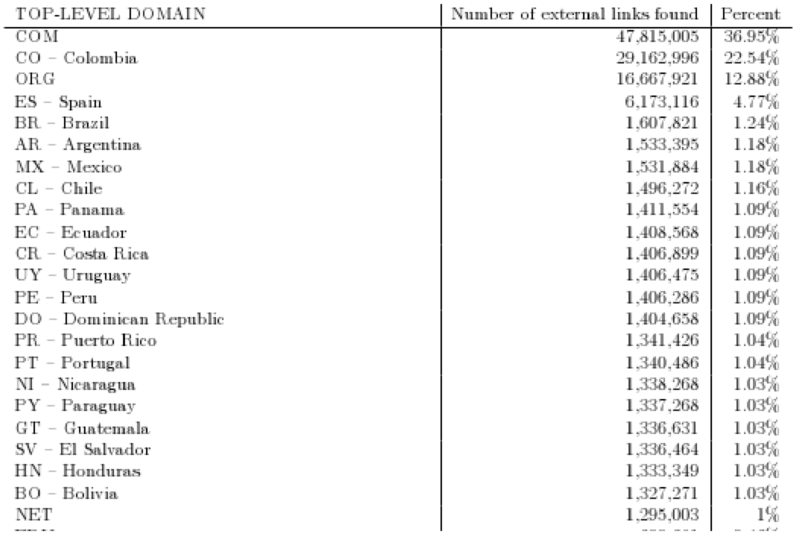
In the Colombian case, it must also be said that this distribution is not complete, given that the collector would only consider a site to be from Colombia if it had the .co domain, and all the other domains (.com, .org, .net) would be ignored leaving the ones presented here as the ones initially gathered from the seeds and its percentage did not grow during the recollection.

Domain	Percentage
com	29.58%
co	61.97%
net	0.03%
org	8.42%

Table 9: Distribution of Top level domains

4.7. External Top level Domains

Finally, information about external level Domains is presented on table 10, there are more than 128 million links to this domains and it can be seen that most of the links after the .com TLD have the common characteristic of being mostly to countries where Spanish is the main language.



TOP-LEVEL DOMAIN	Number of external links found	Percent
COM	47,815,005	36.95%
CO – Colombia	29,162,996	22.54%
ORG	16,667,921	12.88%
ES – Spain	6,173,116	4.77%
BR – Brazil	1,607,821	1.24%
AR – Argentina	1,533,395	1.18%
MX – Mexico	1,531,884	1.18%
CL – Chile	1,496,272	1.16%
PA – Panama	1,411,554	1.09%
EC – Ecuador	1,408,568	1.09%
CR – Costa Rica	1,406,899	1.09%
UY – Uruguay	1,406,475	1.09%
PE – Peru	1,406,286	1.09%
DO – Dominican Republic	1,404,658	1.09%
PR – Puerto Rico	1,341,426	1.04%
PT – Portugal	1,340,486	1.04%
NI – Nicaragua	1,338,268	1.03%
PY – Paraguay	1,337,268	1.03%
GT – Guatemala	1,336,631	1.03%
SV – El Salvador	1,336,464	1.03%
HN – Honduras	1,333,349	1.03%
BO – Bolivia	1,327,271	1.03%
NET	1,295,003	1%

Table 10: External Top level Domains

5. Conclusions

To analyze the Colombian web, we have taken a photo of it during the month of February of 2009. This is similar to taking a photo of cells on a certain period of time, what can be seen on the photo changes rapidly and might not even exist anymore, such as what happens with websites where some might disappear others appear or even experience growth or reduction.

One of the most notable characteristics of the Colombian web is the speed at which it is growing where around 76% of the sites were created in the last year, keeping the web young and offering new alternatives, services and ways of doing commerce, keeping in mind that compared to other national webs, it still remains small.

Unfortunately it is hard to study the Colombian web given the preference of the .com domain over the local .co suffix which makes it difficult to obtain the complete list of sites in order to make a more exhaustive study, but keeping in mind that the Web behaves as a free of scale network, the study of a subset of all the sites already provides significant information that could represent the state of the national web.

A study as the one presented here, has many applications. The most direct one is the development of better search engines and data structures for the web. An example of this is the appearance of CMS systems focused on the user, which brings the web faster to users and also gives them a better experience, but make it harder for the collectors and indexers to find information or even find which site is more important than others in order to provide better search results.

It can also be mentioned that many sites are still islands, not connected to other sites which makes them less important for search engines, but that could have valuable information.

In this study it is also possible to see the importance of multimedia distribution sites as youtube.com or the relevance of the pdf format as a global standard.

Finally it is also very interesting that the most important sites on the national web belong to government, universities or newspapers, bringing quality to the Web, and providing services that are made available to all the population which leads to more development for the country.

6. List of terms

The following list of terms includes common terms used on Internet in general and of the Web and that are used on this document:

AJAX Asynchronous Javascript and XML. It is a technology that allows the browser to continue interacting with the server after the page has been loaded. It is used so that pages do not need to be reloaded or refreshed in order to update information.

CMS Content Management System. It is a web application that takes control of the management and publication of the content of a site. Ie: blogs, forums, galleries and advanced personalized applications.

Domain The form of assigning names to computers on Internet follows a hierarchical structure. A group of computers whose names share a common suffix (like “.co” or “eafit.edu.co”) constitute a domain.

IP Address A sequence of four numbers (in the IP version 4 standard) that identify the location of every computer connected to Internet.

Internet International network that connects thousands of smaller networks. “Internet” in uppercase refers to the net that its currently in use, while “internet” in lowercase refers to the concept of connecting several networks.

Metadata Data about a Web page which is not its main content (or “data about the data”). Usually it includes an address, date, size, keywords, description, etc.

Hostname Name associated to an IP address (ie: “www.eafit.edu.co)

Page Every entity on the web that has an URL associated to it. In this document a more restrictive definition is used, which does not consider images, videos, music and other multimedia or compressed files as pages.

Static Page Every page that exists before being requested.

Dynamic Page Every page that is created the moment it is requested.

Service It is a program that can be executed using Internet. Ie: email, online chat, www.

Server A computer connected to Internet that provides a service.

Website Name of a computer that provides a Web page hosting service.

URL Standard used to refer to an address on the Web, ie: "http://www.site.co/page.html". Defined in 16.

World Wide Web Also simply called Web, is one of the services that can be provided by servers connected to Internet.

7. References

1 Brian D. Davison. Topical locality in the web. In SIGIR ’00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pages 272–279, New York, NY, USA, 2000. ACM Press.

2 Albert-László Barabási. Linked: The New Science of Networks. Perseus Books Group, May 2002.

3 George K. Zipf. Human Behavior and the Principle of Least Effort. Addison-Wesley (Reading MA), 1949.

4 A. Gulli and A. Signorini. The Indexable Web is more than 11.5 Billion pages. In WWW ’05: Special interest tracks and posters of the 14th international conference on World Wide Web, pages 902–903, New York, NY, USA, 2005. ACM Press.

5 P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Structural Properties of the African web. The Eleventh International WWW Conference, May, 2002.

6 G.H. Tolosa and F.R.A. Bordignon. Análisis de Enlaces en el EspacioWeb de las Universidades Argentinas. 2006.

7 A. Rauber, A. Aschenbrenner, O. Witvoet, R.M. Bruckner, and M. Kaiser. Uncovering Information Hidden in Web Archives. D-Lib Magazine, 8(12):1082–9873, 2002.

8 Eveline A. Veloso, Edleno de Moura, P. Golgher, A. da Silva, R. Almeida, A. Laender, Ribeiro B. Neto, and Nivio Ziviani. Um retrato da Web Brasileira. In Proceedings of Simposio Brasileiro de Computacao, Curitiba, Brasil, 2000.

9 Guowei Liu, Yong Yu, Jie Han, and Guirong Xue. China Web Graph Measurements and Evolution. In Web Technologies Research and Development (APWeb), pages 668–679, Shanghai, China, 2005. Springer Berlin / Heidelberg.

10 A.A. Benczur, K. Csalogany, D. Fogaras, E. Friedman, T. Sarlos, M. Uher, and E. Windhager. Searching a small national domain–a preliminary report. Poster Proceedings of Conference on World Wide Web, 2003.

11 Gabriel H. Tolosa, Fernando R. Bordignon, and Pablo J. Lavallén. Caracterización del espacio web de perú. 2006.

12 D. Gomes and M.J. Silva. A characterization of the Portuguese Web. 3rd ECDL Workshop on Web Archives, Trondheim, Norway, 21, 2003.

13 M. Thelwall and D. Wilkinson. Graph Structure in Three National Academic Webs: Power laws with anomalies. Journal of the American Society for Information Science and Technology, 54(8):706–712, 2003.

14 S. Sanguanpong, P.P. Nga, S. Keretho, Y. Poovarawan, and S. Warangrit. Measuring and Analysis of the Thai World Wide Web. Proceeding of the Asia Pacific Advance Network conference, pages 225–230, 2000.

15 Ricardo Baeza-Yates and Carlos Castillo. WIRE: Web Information Retrieval Environment, 2006. http://cwr.cl/projects/WIRE

16 T. Berners-Lee, L. Masinter, and M. McCahill. RFC1738: Uniform Resource Locators (URL). Internet RFCs, 1994.

17 J Cho, N. Shivakumar, and H. Garcia-Molina. Finding Replicated Web Collections. ACM SIGMOD, pages 355-366, 1999.

18 Ricardo Baeza-Yates and Carlos Castillo. Caracteristicas de la Web Chilena 2006. Technical report, Center for Web Research, University of Chile, 2007.

19 Ricardo Baeza-Yates and Felipe Lalanne. Characteristics of the Korean Web. Technical report, Korea-Chile IT Cooperation Center ITCC, 2004.

20 The PHP Group. PHP: Hypertext Preprocessor, 2009. http://www.php.net

21 Marco Modesto, Álvaro Pereira, Nivio Ziviani, Carlos Castillo, and Ricardo Baeza-Yates. Um novo retrato da web brasileira. In Proceedings of XXXII SEMISH, pages 2005-2017, São Leopoldo, Brazil, 2005.

22 Ricardo Baeza-Yates, Carlos Castillo, and Vicente López. Caracteristicas de la web de españa. El Profesional de la Informacion, 15(1), January 2006.

23 Microsoft ASP: Active Server Pages. 2006. http://msdn.microsoft.com/asp.net/.

24 P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Structural Properties of the African Web. The Eleventh International WWW Conference, May 2002.

25 A. Rauber, A. Aschenbrenner, O. Witvoet, R.M. Bruckner, and M. Kaiser. Uncovering Information Hidden in Web Archives. D-Lib Magazine, 8(12):1082–9873, 2002.

26 Efthimis Efthimiadis and Carlos Castillo. Charting the Greek Web. In Proceedings of the Conference of the American Society for Information Science and Technology (ASIST), Providence, Rhode Island, USA,

7. References

November 2004. American Society for Information Science and Technology.

27 D. Gomes and M.J. Silva. A Characterization of the Portuguese Web. 3rd ECDL Workshop on Web Archives, Trondheim, Norway, 21, 2003.

28 G. Pandurangan, P. Raghavan, and E. Upfal. Using PageRank to Characterize Web Structure. 8th Annual International Computing and Combinatorics Conference (COCOON), pages 330–339, 2002.

29 L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank Citation Ranking: Bringing Order to the web, 1998.

30 Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 46(5):604–632, 1999.

31 S. Dill, R. Kumar, K.S. McCurley, S. Ra jagopalan, D. Sivakumar, and A. Tomkins. Self-Similarity In the Web. ACM Transactions on Internet Technology, 2(3):205–223, 2002.

32 A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Ra jagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: experiments and models. Proceedings of the ninth WWW Conference, 2000.

33 Ricardo Baeza-Yates and Carlos Castillo. Relating Web Characteristics With Link Based Web Page Ranking. In Proceedings of String Processing and Information Retrieval SPIRE, pages 21–32, Laguna San Rafael, Chile, 2001. IEEE CS Press.

34 Netcraft. Netcraft, 2009. http://www.netcraft.com