Characteristics of the Web of Spain

Ricardo Baeza-Yates
ICREA –
Department of Technology
Universitat Pompeu Fabra

Carlos Castillo
Cátedra Telefónica –
Department of Technology
Universitat Pompeu Fabra

Vicente López
Cátedra Telefónica –
Department of Technology
Universitat Pompeu Fabra

Abstract

The Web is a massive and interlinked collection of documents, built using a decentralized design to encourage the participation of many authors who publish information through a huge number of Web sites. Its characteristics are the result of the interaction between many organizations and individuals, and those interactions generate a large amount of diversity. This diversity means that several different topics are represented on the Web, and at the same time that the overall quality of pages and Web sites is very variable. The Web is very dynamic and is growing at a very fast pace, and even when some of its properties have been studied, there are several characteristics of it that are still not fully known.

This article reports the results of an in-depth study over a large collection of Web pages. On September and October 2004 we downloaded more than 16 million Web pages from about 300,000 Web sites from the Web of Spain. We show the characteristics of this collection at three different granularity levels: Web pages, sites and domains. For each level, we analyze contents, technologies and links and present statistics and models. We found that some of the characteristics of this collection resemble the ones of the Web at large, while others are specific to the Web of Spain, or have not been studied in the past.

Contents

1 Introduction

In this section we introduce the motivation for this work, the methodology used, and general characteristics of the studied Web collection.

1.1 Motivation

The World Wide Web has become in about 12 years the largest cultural endeavour of all times, equivalent in importance to the first Library of Alexandria. The main difference between both libraries is not that one was made of scrolls and ink, and the other one is made of hard drives, cables and digital signals. The main difference is that while in the Library books were copied by hand, most of the information on the Web has been reviewed only once, by its author, at the time of writing.

Digital technology allows fast reproduction of the work, with no human effort. The cost of disseminating content is lower due to new technologies, and has been decreasing substantially from oral tradition to writing, and then from printing and the press to electronic communications. This has generated much more information than we can easily handle.

On the dawn of the World Wide Web, finding information was done mainly by scanning through lists of links collected and sorted by humans according to some criteria. Automated Web search engines were not needed when Web pages were counted only by thousands, and most directories of the Web included a prominent button to "add a new Web page". Web site administrators were encouraged to submit their sites. Today, URLs of new pages are no longer a scarce resource, as there are thousands of millions of Web pages [Gulli and Signorini, 2005].

The open nature of the Web, which encourages many authors to publish contents, means that the results are unlike traditional, controlled, text collections. The Web is "massive, much less coherent, changes more rapidly, and is spread over geographically distributed computers" [Arasu et al., 2001].

The World Wide Web is the result of the interactions of many individuals and organizations, and from these interactions complex characteristics may arise. "While entirely of human design, the emerging network appears to have more in common with a cell or an ecological system than with a Swiss watch." [Barabási, 2001]. Some of the characteristics of the Web, most notably power-law distributions, can be partially explained by current models, while other characteristics have not been studied in detail in large collections.

1.2 How is the Web?

One of the greatest advantages of the Web is its capacity for relating information through links. These relationships gives users great flexibility when searching for information, and can be modeled by considering the Web as a directed graph (a digraph). In this graph, each node is a Web page and each edge represents a hyperlink between two pages.

These links are certainly not at random. Pages that are linked together are more likely to be on the same subject that pages taken at random [Davison, 2000, Menczer, 2004]. Besides, the best pages tend to attract more references [Caldarelli et al., 2002]. The Web as a graph has an structure that can be classified as a *scale-free network*. Scale-free networks, as opposed to random networks, are characterized by a skewed distribution of links, and they have been the subject of a series of studies (see for instance [Barabási, 2002] for an introduction). In scale-free networks, the distribution of the number of links of a page p follows a power-law:

$$Pr(\text{page } p \text{ has } k \text{ links}) \propto k^{-\theta} \tag{1}$$

For the Web, it has been observed that the the number of pages with k in-links is proportional to $k^{-2.1}$ [Broder et al., 2000]. Scale-free networks have a few highly-connected links that act as "hubs" connecting many other nodes to the network. An illustration of the difference between a scale-free network and a random network is shown in Figure 1.

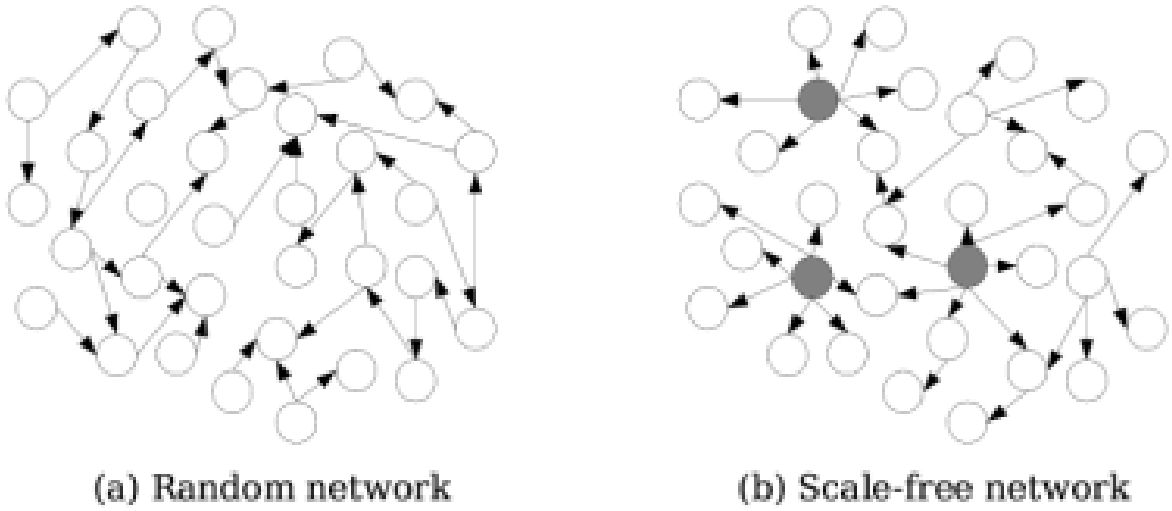

Figure 1: Examples of a random network and a scale-free network. Each graph has 32 nodes and 32 links.

When representing graphically the number of in-links and frequency in a logarithmic scale, a straight line appears; we find this distribution on the Web in almost every aspect, and we can see this in several graphics of this study. It has been said that "no paper on statistics of web pages is complete without a graph showing a power-law distribution" [Fetterly et al., 2005]. It is the same distribution found by economist Vilfredo Pareto in 1896 for wealth in population: 80% of the wealth is owned by 20% of the people. It is also the same distribution found by George Kingsley Zipf in 1932 for the frequency of words in texts, and that later turned out to be applicable to many other domains [Zipf, 1949].

1.3 Studying the Web of a Country

The Web graph is self similar [Dill et al., 2002], as a small part of the graph shares most of the properties of the full graph. This is the case for most scale-free networks (but not all of them [Barábasi et al., 2001]). Our collection of pages from the Web of Spain[1], shows characteristics that are very similar to those of the global Web, which is remarkable if we consider that the latter has at least 11×10^9 pages [Gulli and Signorini, 2005]: three orders of magnitude larger that our collection.

A national Web is the set of pages related to a country. Checking if a Web page belongs to a country is a difficult technical problem, so we use certain heuristics. Given that the organization that controls the assignment of addresses and symbolic names on the Internet[2] reserves certain

[1] We use "Web of Spain" instead of "Spanish Web", as the latter could be mistaken by the pages written in the Spanish language, which is not the case.

[2] IANA: Internet Assigned Numbers Authority. Home page available at http://www.iana.org/.

suffixes to each country, for instance, `.fr` for France and `.es` for Spain, a first approach is to say that the Web of Spain is the set of pages whose domain includes the suffix `.es` (note that the complete list of domains under a country-code top-level domain is typically not public).

In the case of Spain the country-code is not the best method for defining the Web of this country, as there are thousands of Web sites that do not use the `.es` domain, mostly for two reasons. First, a `.es` domain has a higher cost than a `.com` domain; approximately €100 against an average of €15 (per year) of a `.com`; second, to get a domain name under `.es` it is necessary to prove that the applicant owns a trade mark, or represents a company, with the same name as the domain being registered.

The heuristic we use for defining the Web of Spain is that a Web site is in Spain if its IP address is assigned to a network physically in Spain, or if the Web site's suffix is `.es`. This allows us to get much more pages than by looking only at the domain suffix; as shown in Table 6 only 16% of the domains with pages in Spain are under `.es`.

In the last years the Webs of different countries have been studied with different methodologies. The following is a list of other countries or regions that have carried these studies:

- Africa (9 countries) [Boldi et al., 2002]
- Austria [Rauber et al., 2002]
- Brazil [Veloso et al., 2000, Modesto et al., 2005]
- Chile [Baeza-Yates and Castillo, 2000, Baeza-Yates and Poblete, 2003, Baeza-Yates et al., 2003, Baeza-Yates and Castillo, 2005a]
- Greece [Efthimiadis and Castillo, 2004]
- Hungary [Benczúr et al., 2003]
- Portugal [Gomes and Silva, 2003]
- South Korea [Baeza-Yates and Lalanne, 2004]
- Thailand [Sanguanpong et al., 2000]
- United Kingdom, New Zealand and Australia (only university pages) [Thelwall and Wilkinson, 2003]

Some of these studies are summarized in [Baeza-Yates and Castillo, 2005b]. There are two previously published studies about the Web of Spain: an in-depth report on 27 specific Web sites from universities and public institutions [Alonso et al., 2003], and a preliminary study about a large sample of Web sites, approximately half of the ones we analyze in this document [Baeza-Yates, 2003].

1.4 Web Crawling

There are three main methods for obtaining a collection from the Web [Pitkow, 1999]:

- Modifying a Web browser to record user's actions. This arises privacy issues, and also limits the amount of data that can be obtained to the pages that are visited by the group of persons using the instrumented browser.

- Recording user sessions at a proxy level. This has the problem that only the sessions from one or a few organizations can be recorded, and therefore the subset of the Web that is obtained may have a limited topical scope.

- Using a Web crawler to download pages automatically starting from a set of pages. This has several limitations; see for instance [Thelwall, 2004, Chapter 2].

Our collection was obtained by running a Web crawler between September and October 2004, using a single PC with two Intel-4 processors running at 3 GHz and having 1.5 GiB[3] of RAM under Red-Hat Linux. For the information storage we used a RAID of disks with 1.6 TiB of total capacity, although the space used by the collection was less than 50 GiB.

We used Web crawling software developed by Akwan [da Silva et al., 1999]. The crawler starts by downloading a set of starting URLs, which in our case were obtained from the pages included in the old Buscopio search engine[4]. After downloading the pages, new URLs were extracted from the downloaded pages, and the process continued recursively while the pages were considered inside Spain, according to the definition outlined in the previous section.

We downloaded only HTML, plain text, Adobe PDF, Microsoft Word (`.doc`), and Adobe Postscript (`.ps`) files. To filter other types of files, we used the mime-type header returned by Web servers, a list of 130 extensions of known non-textual content (such as `.gif`, `.mp3`, etc.) and a list of 15 extensions related to the file types we were interested in.

While the amount of information available on the Web is finite, the number of Web pages is potentially infinite [Baeza-Yates and Castillo, 2004] due to the existence of dynamic Web pages. We restricted the crawler to download a maximum of 400 pages per site, except for the Web sites inside `.es`, that had a limit of 10,000 pages per site.

Once a page was downloaded, it was parsed to extract its text and links. A maximum of 300 KiB of text and 250 links per page were kept after parsing. The crawler followed the robots exclusion protocol [Koster, 1996], by downloading and obeying the `robots.txt` file, avoiding multiple simultaneous connections to the same host, and waiting at least 60 seconds between connections.

The crawler only tried to download each page once, and HTTP connections timed out after 60 seconds of inactivity. If a Web page was not available, other pages from the same site were retried until exhausting the list of URLs for that site.

Running a Web crawler is, in a certain way, like sending a vehicle for exploring the surface of mars: it would be ideal to know the terrain in detail before sending the vehicle, but the vehicle is needed to explore the terrain. The set of limits and other parameters we chose for the crawler, while arbitrary, were consistent with the ones used in other studies and are reasonable according to our experience on running large Web crawls in other countries.

During the time of the study we downloaded over 16 million Web pages, and processed them to extract links, text, and metadata. Table 1 summarizes the main characteristics of the collection. The definition of Web site and domain is given in Section 1.7.

1.5 Difficulties in Web Characterization

The Web is a decentralized collection, in which different authors may contribute contents on their own without a central authority controlling what is published and what is not. This is the main

[3] We use "GiB", "MiB", etc. for powers in base 2, while "GB" and "MB" mean powers in base 10.

[4] Available from 2001 to 2002 at `http://www.buscopio.net/`.

Table 1: Overview of the Web of Spain.

Web Pages	16.171.267	
Total Text	43 GiB	
Average text per page	2.855 B	
Web Sites	308.822	
Average pages per site	52,08	
Average text per site	149.521 B	146 KiB
Domains	118.248	
Average sites per per domain	2,61	
Average pages per domain	136,75	
Average text per domain	382.368 B	373 KiB

advantage of the Web, but also the main cause of difficulties for searching information and for characterization.

In the studied collection, we detected the following anomalies that constitute either bad implementations of W3C standards by Web page authors, or special situations that make Web characterization more difficult.

Parameters in the URL As shown in Figure 4, we found a few very long URLs. By inspecting them, we detected that in most cases they are addresses in which the parameters to a program are passed inside the "path" part of URL addresses. This is syntactically correct but semantically contradicts the standard [Lee et al., 1994], as the parameters for calling programs should appear at the end of the URL after a question mark "`?`", for example:

- Parameters inside the path of the URL: `http://site/dir/search/word/X/max/10`.
- Parameters according to the standard: `http://site/dir/search?word=X&max=10`.

The consequence of this is that it is not possible to make a perfect separation between static and dynamic pages, and this may lead to crawl several times pages with semantically the same content.

Content replicas (*mirrors*) A common practice on the Web is to create several geographically distributed copies of the same contents, to ensure network efficiency as the users can download the copy that is "closer" to their location. Normally, these replicas are entire collections having a large volume. In [Cho et al., 1999], it was found that the replicated information was between 20% and 40% of the total Web contents, and that the most replicated collections on the Web were the software site Tucows, the Linux Documentation Project (LDP), the manual of the Apache Web Server and the specification of the Java API. More recent studies have found that about one third of the Web pages are exact duplicates [Fetterly et al., 2005], and Section 3.2 shows that in the Web of Spain today the large replicated collections are roughly the same than in the full Web in 1999.

The consequences of these replicas are that there are many sites having the same contents; besides, as these collections are normally very large, these sites appear as having an amount of content that is several times larger than the average.

Spam in general Spam on the Web refers to actions oriented to deceive search engines and to give to some pages a higher ranking than they deserve in search results. These actions include changes in the page contents, the metadata and/or links [Gyöngyi and Garcia-Molina, 2005]. Recognizing spam pages is an active research area, and it is estimated that over 8% of what is indexed by search engines is spam.

It can be argued that as spam is a part of the Web, spam pages should be included in a Web characterization study. However, spam pages use computational resources and bandwidth that could be used for downloading pages with content that is actually viewed by users, so we try to avoid them as explained in Section 1.6.

Domain name spamming (DNS wildcarding) Some link analysis ranking functions assign less importance to links between pages in the same Web site [Bharat and Henzinger, 1998]. Unfortunately, this has motivated spammers to use several different Web sites for the same contents. A usual technique for doing this is to configure DNS servers to assign hundreds or thousands of site names to the same IP address. This is called "DNS wildcarding" [Barr, 1996]. On the Web of Spain, we observed that 24 out of the 30 domains which appeared to have the largest amount of Web sites were configured to use DNS wildcarding.

1.6 Data Cleaning

We made a preliminary crawl in July 2004, but we stopped the process when we noticed that we were collecting a large number of spam pages, most of them under the `.com` top-level domain. We manually checked the domains with the larger number of sites and pages to generate a black list that is composed of 200 domain names that contain pages including thousands of nepotistic links, with little or no information content.

This help us reduce the bandwidth usage, but is far from being a perfect solution for eliminating spam. One of the most revealing signs of spam is the presence of unexpected regularities on certain variables [Fetterly et al., 2004], and in Figure 14 we can see that there are some large groups of pages sharing mostly the same connectivity, indicating that it is likely that they are generated automatically to improve the score they obtain in link-based rankings algorithms.

1.7 Organization of this Work

While the first Web characterization studies focused mostly on Web pages [Pitkow, 1999], pages can be grouped in larger units to understand better some aspects of the Web. This is the idea behind the Alternative Document Model [Thelwall, 2002], and also behind the visual language for depicting different levels of the Web presented in [Björneborn and Ingwersen, 2004].

The objective of our research is broad: **to study contents, technologies and links on the Web**. To achieve this objective, we analyze our data using different granularity levels, depicted in Figure 2. The smallest level is that of words or blocks of text or images, which we do not consider except for a brief discussion on the distribution of the vocabulary of texts. The next unit is the level of pages, which we separate into static pages (pages that exist before being requested) and dynamic pages (pages that are created on request, usually connected to a database).

We decided for this study to group pages by host names, as it is easier than identifying coherent units inside hosts. In this article, a Web site is identified by the `host` part of a URL; therefore, `http://ex.example.com/dir/page.html` belongs to the Web site "`ex.example.com`". As an heuristic to avoid duplicates, we merged "`www.example.com`" and "`example.com`" under the same Web site as it is common that they both have the same contents.

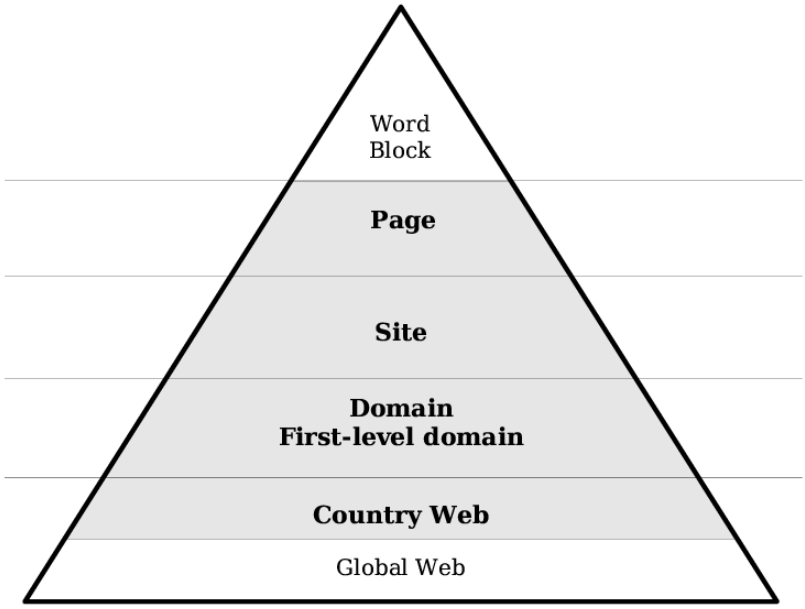

Figure 2: Levels of analysis for the Web; the marked levels are covered in this article.

As explained in Section 4, a domain name is a suffix of the site name; for instance `site1.example.com` and `site2.example.com` both belong to the same domain name "`example.com`". A top-level domain is usually a two-letter country code or one of the generic top-level domains such as `.com` or `.net`. The Web of a country is composed of sites under different top-level domains; for instance in the case of Spain, most of the pages are under `.es`, but there are also many pages under `.com` and other top-level domains.

The rest of this paper presents our observations about the Web of Spain at different levels: the page level is shown in Section 2, the site level in Section 3 and the domain level in Section 4. Finally, Section 5 presents our conclusions.

A more detailed version of this article in Spanish is available at `http://www.catedratelefonica.upf.es/webes/2005`.

2 Web Page Characteristics

In this section we present an analysis of the Web pages individually, without considering their grouping in sites or domains. We start with the metadata of the pages, such as the URL, title or size, and then we continue by analyzing the page contents and hyperlinks.

2.1 URLs

The address of a Web page is always written using a URL (*Uniform Resource Locator*) [Lee et al., 1994]. URLs have a double purpose, as at the same time they identify a resource and indicate how to locate it.

The URLs of Web pages use the Hypertext Transfer Protocol (HTTP). These URLs have the following form:

`http://`*host*`[:`*port*`]/`*directory*`/`*file*

For instance, `http://www.upf.edu/dtecn/docencia.html`, indicates that the host that has to be contacted is `www.upf.edu`, that the directory where the page is found is `dtecn/` relative to the root directory of the Web site, and that a filename called `docencia.html` has to be accessed.

In this section, we analyze some of the characteristics found on the URLs of the Web of Spain.

2.1.1 URL Length

The distribution of the length of the URLs is important because it can help in the development of compression schemes for caching or indexing the Web. In our collection, the average length of an URL, including the protocol specification `http://`, the server name, path and parameters, is 67 characters. This result can be compared with the average of 62 characters observed in Portugal [Gomes and Silva, 2003], or 50 characters in samples of the global Web [Suel and Yuan, 2001]. In this last reference, it was found that the URLs are very compressible; simply by exploiting the fact that many address share the same prefix they can be compressed to 13 bytes per URL, and with more sophisticated techniques, to 6 or 7 bytes.

About 80% of the URLs have between 40 and 80 characters. The distribution of URL lengths is shown in Figure 3, and it resembles a log-normal distribution with probability density function:

$$f(x) = \frac{e^{-((\log((x-\theta)/m)^2)/(2\sigma^2))}}{(x-\theta)\sigma\sqrt{2\pi}} \ ,$$

and estimated parameters $\theta = 13$ (position), $m = 47$ (scale) and $\sigma = 0.4$ (shape). By manual inspection of the longer URLs, we found that most of them fall into one of the following cases (mostly the first one):

- Repetition of elements, for instance: `http://www.comunitel.es/webs/ibooks/clsid: -D27CDB6EAE6D11cf96B8444553540000/clsid: D27CDB6EAE6D11cf96B8444553540000/clsid:...`
- Text parameters passed to a program, for instance: `http://fama.us.es/search*spi/ -tPropagaci{226}on+de+Ondas+Guiadas/ tpropagacion+de+ondas+guiadas/1,1,1,B/ frameset&-FF=tpropiedad+publica+organizacion...`

Most of the longer URLs do not seem to provide content that is not accessible using a shorter URL. It is likely that most URLs longer than 150 bytes are anomalies, so we recommend Web crawler authors to study the long URLs in their collection to detect patterns that can be discarded to avoid duplicates and save bandwidth.

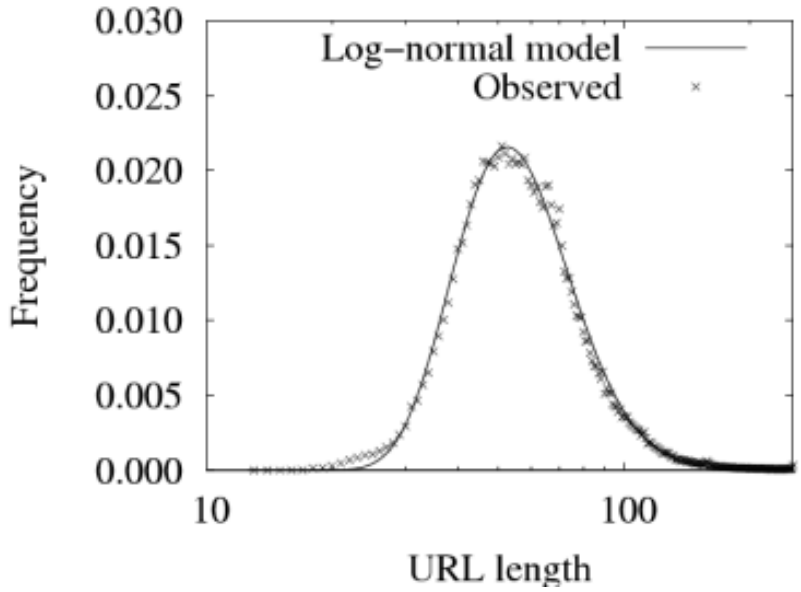

Figure 3: Distribution of URL lengths.

2.1.2 URL Depth

We studied the depth of pages inside Web sites. The page depth can be defined in two ways:

Logical depth: the starting page of a Web site is at depth 1, all the pages reachable directly from the home page are at depth 2, and so on. The logical depth measures the number of "clicks" needed to go from the front page of a site to a page.

Physical depth: the starting page of a Web site is at depth 1, pages with URLs of the form `http://site/pag.html` or `http://site/dir/` are at depth 2, and so on. In Web sites with static pages, the physical depth measures the organization of pages in files and directories.

Logical depth is in general hard to measure, as a Web site according to our definition can have several starting pages. In this study, we analyze the physical depth of pages, because it can be extracted directly from URLs. The distribution of this variable is shown in Figure 4, in which we have separated static and dynamic pages.

The distribution of pages follows a curve that reaches a maximum between the third and fourth levels. However, there is an important increase in the frequency in pages between 21 and 24 levels of depth –they constitute about 19% of the pages. By manual inspection we observed that for each one of these levels, there is a large group of pages inside a Web site (which is different for different levels). For instance, pages at level 21 correspond to a Web site for vacation rentals, and the "directories" observed in the URL are actually hidden parameters to an application, they are not physical directories in a file system. These pages are an anomaly and are not part of the real distribution; again, we recommend Web crawler authors to check sites having large numbers of URLs with more than 15 slashes to see if patterns can be detected to save the crawler's bandwidth.

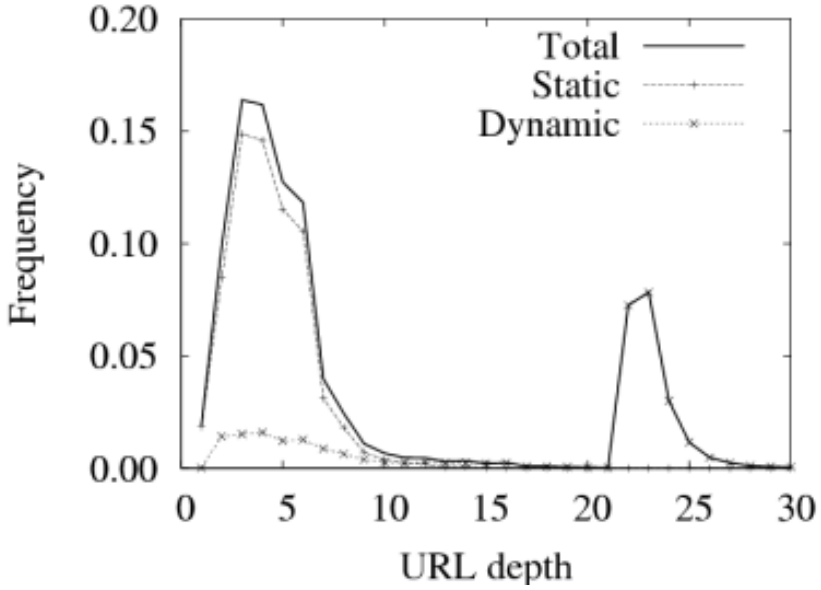

Figure 4: Distribution of the physical depth of Web pages.

2.2 Page Titles

We analyzed the title of pages and found that over 9% of the pages have no title, and 3% have a default title such as the Spanish equivalents of "*Untitled document*"[5] or "*New document 1*". When extracting page titles, we only stored the first 100 bytes, which is enough for most pages.

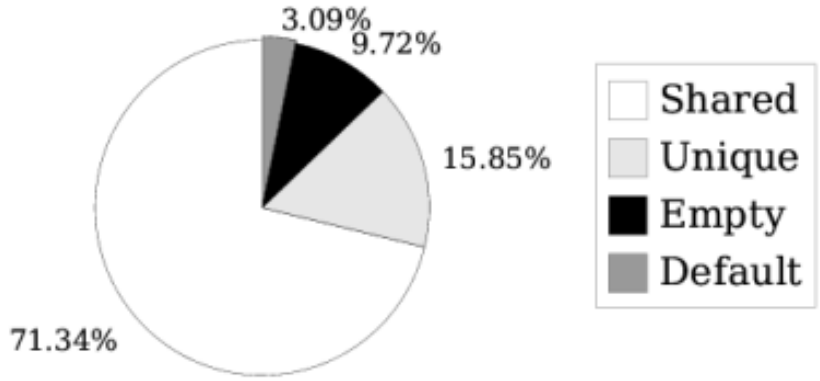

Figure 5: Distribution of the title of pages: shared (used by more than one page), unique (used exclusively by that page), empty or default.

Figure 5 shows the distribution of page titles. The most common case is that a page has a title, but this title is not unique in the collection of pages. It is unlikely that a title that is shared by many pages is a good description of the contents of a page. The titles of pages are repeated several times across pages, mainly because authors use just the name of the site as the title of many unrelated

[5]By searching in Google's collection from the global Web, we found nearly 40 million pages having this title.

pages. We observed that on average a title is shared by approximately 4 pages; this is similar to what was observed in the Web of Portugal: a different title each 5 pages [Gomes and Silva, 2003]. An analysis of the distribution of titles per domain is shown in Figure 33, and shows that titles are repeated frequently inside each domain.

Figure 6 shows the distribution of the title lengths in pages, excluding pages without title, or pages with a default title. Most of the titles are rather short, which reinforces the idea that they are probably not a good description of their contents. The distribution of the title lengths also follows a log-normal model with parameters $\theta = -10$, $m = 17$ and $\sigma = 0.2$.

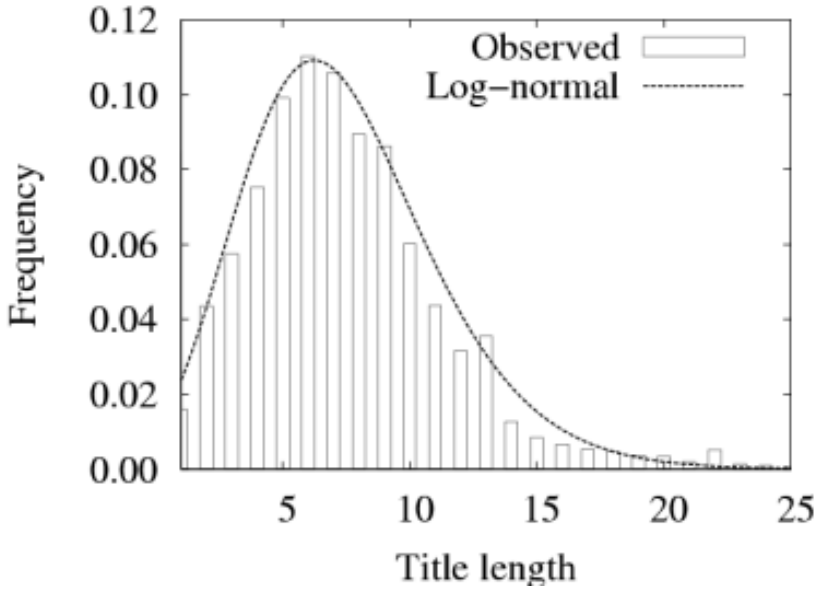

Figure 6: Distribution of the title lengths.

Using descriptive titles is an important element of Web usability [Nielsen, 2005], as it allows the visitors of a Web page to see the context of the page they are visiting and to store it with a descriptive name in their *bookmarks* or *favorites*. Furthermore, many search engines give more importance to keywords appearing in the title of the page than to the same keywords appearing in the main text of the page, so without a title or without good keywords in the title there are less possibilities of appearing in the top results of a search engine. Unfortunately, as titles are omitted or repeated across Web pages, less than 10% of the pages in the Web of Spain have titles that can be used by search engines; the usage of metadata such as keywords and description is probably even lower. It has been estimated that a critical mass of at least 16% of pages tagged with metadata is required in order to be able to propagate the metadata through links to provide "fuzzy metadata" for the rest of the Web [Marchiori, 1998].

2.3 Text of Pages

After extracting the text of the pages, we stored only the first 300 KB of each page. We depict the distribution of page sizes in Figure 7; there are many pages with a few bytes of text, and a few pages with a huge size.

A power-law distribution is a good model of the distribution of page sizes, with parameter 2.25. This figure can be compared with 2.75 in Chile [Baeza-Yates and Castillo, 2005a] or 2.84 in South

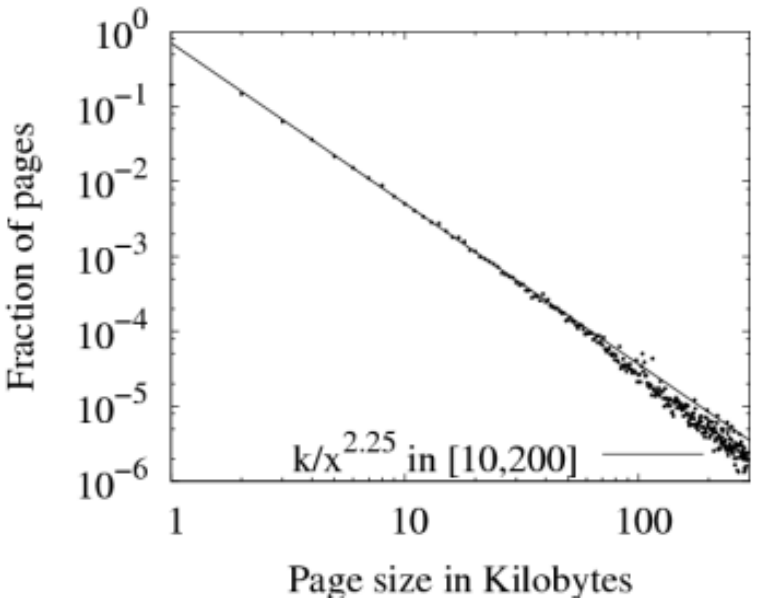

Figure 7: Distribution of page size.

Korea [Baeza-Yates and Lalanne, 2004]. These distributions can be used for optimizing large-scale storage systems to store Web pages.

To study the distribution of the sizes of the smaller pages, we draw an histogram with bins of exponential size, as shown in Figure 8. Most of the pages have between 256 B and 4 KiB of text, and the average is 2.9 KiB, which is almost equal to the Web of Portugal[Gomes and Silva, 2003] that has 2.8 KiB of text per page on average. Other authors have modeled page sizes using a log-normal distribution [Crovella and Bestavros, 1996], but our results do not fit with that distribution.

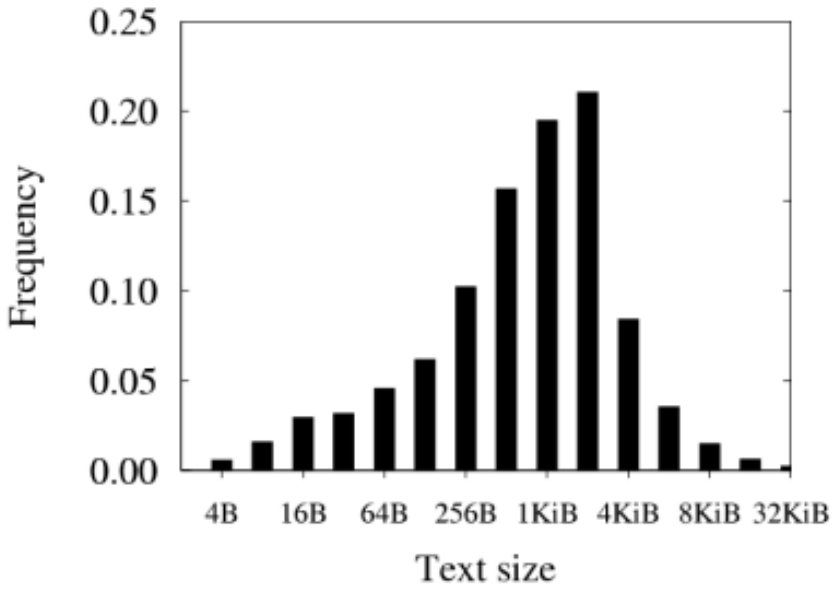

Figure 8: Distribution of page sizes for pages with less than 32 KiB of text.

When manually inspecting the pages, we notice that several of the pages that appear to have a very small text size are pages from Web sites built mostly with graphical elements such as images or animations, while bigger pages are either automatically generated indexes, or long texts covering diverse topics (legal, technical, etc.).

2.4 Language

We used an statistical text analysis system called Bow [Mccallum, 1996]. This application does, among other things, n-gram-based classification using Naïve Bayes. The system was trained with English documents, with documents in the official languages of the country: Spanish, Catalan, Galician and Basque, and with documents in other European languages. On the studied sample we were able to obtain the language with a high level of certainty for around 62% of the pages. The distribution for these pages is shown in Figure 9.

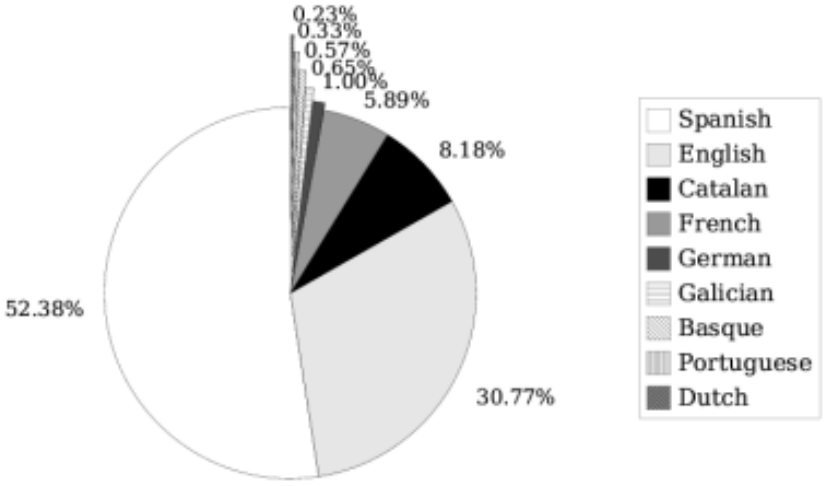

Figure 9: Distribution of the language in which pages are written.

The Spanish language is used by a little more than half of the pages, followed by English and Catalan. The fraction of pages written in the official languages of the country is around 62%. This is less than the 73% of pages in Portugal written in Portuguese [Gomes and Silva, 2003], 75% of Brazilian pages in Portuguese [Veloso et al., 2000], or approximately 90% of the Chilean pages in Spanish [Baeza-Yates and Castillo, 2000] and is related to the presence of a large group of pages in English, including pages related to tourism and technical documentation about computing.

There are other national domains in which English is the most used language on-line, such as the Web of Thailand (66% in English) [Sanguanpong et al., 2000] or the Web of several African countries (75% in English) [Boldi et al., 2002]. In the latter case we have to consider that the country with the larger number of pages in the African sample was South Africa, a country in which English is one of the official languages.

2.5 Vocabulary

The definition of word that we use is any alphanumeric sequence, including the accented characters in romance languages, of length equal or greater than 2. We analyzed 1 GB of the text extracted from pages in each of the three most frequent languages of the sample.

In Figure 10 the histogram with the word frequencies in the collection is shown. The distribution closely follows a Zipf's law with parameter 0.7 for English, and close to 0.8 for Spanish and Catalan.

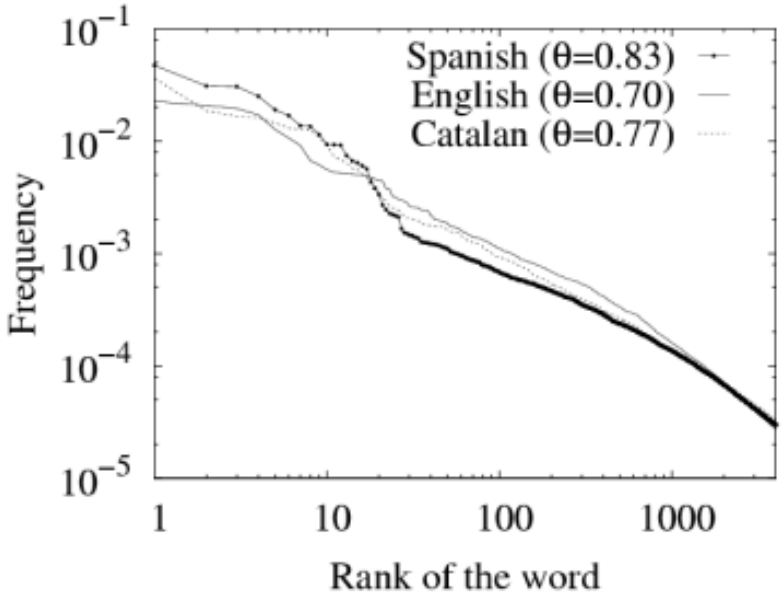

Figure 10: Distribution of the frequency of words in the collection.

The most frequent words are obviously mostly *stopwords*, carrying no meaning by themselves. The most frequent words in Spanish are practically the same as those appearing in 2002 [Baeza-Yates, 2003]. It is interesting to notice that the name of a country turned out to be a rather frequent term in this type of sample, as have been observed previously in Brazil [Veloso et al., 2000] and Chile [Baeza-Yates and Castillo,

Manual inspection of the most frequent words in the English pages indicates that most of these pages are technical documentation. The most frequent words in Catalan pages, on the other hand, indicate a strong presence of pages related to Universities or educational organizations. Manual inspection of the nouns with the higher frequency on Web pages is a good starting point for detecting the most represented topics on Web collections.

2.6 Dynamic Pages

A dynamic page is a page generated at the time of being requested, that did not exist previously; this is normal when there is a query to a database involved in the process of showing a page. Checking for the presence of a question mark "?" and of known extensions associated to dynamic pages, we found out that over 3.5 million pages of the Web of Spain (22%) were dynamic.

If we count by number of pages, the most used application for building dynamic pages is PHP[6], followed closely by ASP[7]. Other technologies that are used for many pages are Java (`.jhtml` and `.jsp`) and ColdFusion (`.cfm`). The distribution is shown in Figure 11.

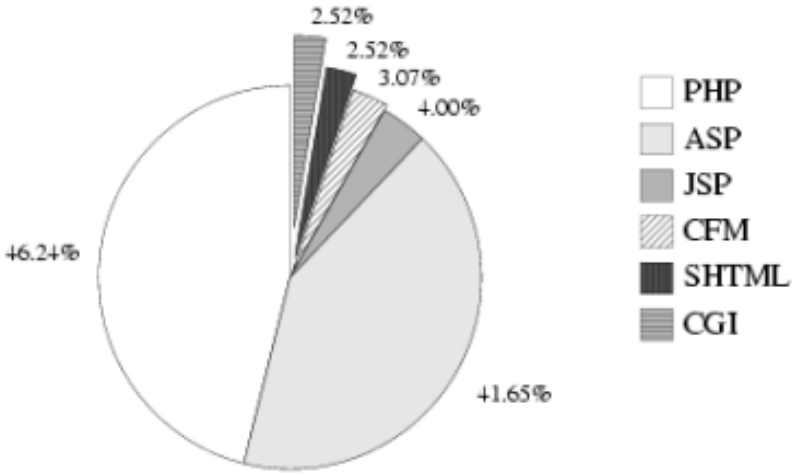

Figure 11: Distribution of technologies for dynamic pages.

Instead of using general-purpose programming languages, dynamic pages are built mainly using hypertext pre-processing techniques (PHP, ASP, JHTML, ColdFusion), in which commands to generate dynamic content are embedded in documents that are mostly HTML code. Programming languages for creating Web pages are always evolving, and the share of different technologies may change in the future. For instance, in the beginning of the Web most dynamic pages were written using Perl, which now is used only for a small fraction of pages. Also, the usage of XML and client-side transformation stylesheets may change the way in which dynamic pages are generated.

The share of different programming languages is related to the distribution of operating systems, as shown in Figure 29, given that ASP as a closed-source technology only works in certain platforms. On the other hand PHP, an open-source technology, clearly dominates the market, but not for a margin as wide as in Brazil (73% PHP) [Modesto et al., 2005] or Chile (78% PHP) [Baeza-Yates and Castillo, 2005a]. This situation is reversed in other countries in which ASP is more used than PHP, as in the samples from Africa (63% ASP) [Boldi et al., 2002] and South Korea (75% ASP) [Baeza-Yates and Lalanne, 2004].

It must be noted that some of the pages that seem to be static, even those with `.html` extension, are really generated automatically using batch processing and content management systems, so there are other dynamic content technologies that might be missing from this analysis.

[6]PHP: the hypertext pre-processor, available online at `http://www.php.net/`.
[7]ASP page in Microsoft Developer Network, available online at `http://msdn.microsoft.com/asp.net/`.

2.7 Documents that are not in HTML

We found approximately 200,000 links to document files that were not in HTML; this is a large collection of documents in absolute terms but represents only 1% of the pages on the Web. Plain text and Adobe PDF (*Portable Document Format*) are the most used format and comprise over 80% of the non-HTML documents. The distribution is shown in Figure 12.

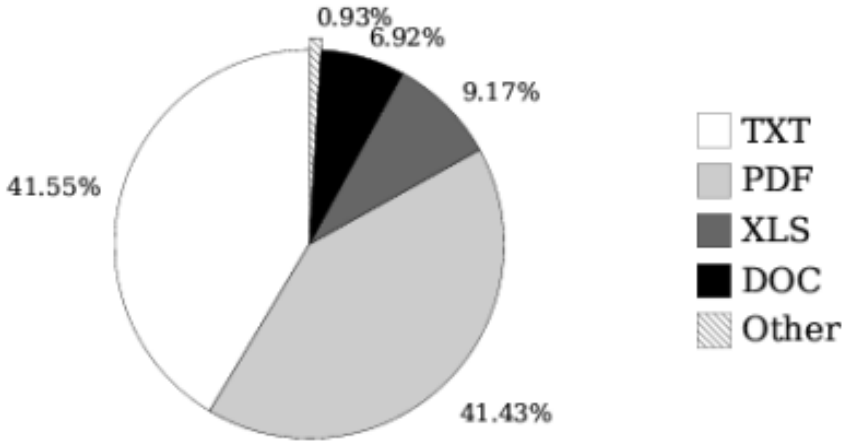

Figure 12: Distribution of links to documents, excluding links to HTML pages.

The PDF format is the most used for documents that are not in HTML in Austria (54%) [Rauber et al., 2002], Brazil (48%) [Modesto et al., 2005], Chile (63%) [Baeza-Yates and Castillo, 2005a], Portugal (46%) [Gomes and Sil and South Korea (63%) [Baeza-Yates and Lalanne, 2004]. Despite the fact that Microsoft Windows is the most used operating system, the extensions associated to Microsoft Office applications such as Word or Excel comprise only around 16% of files.

The Web crawler was configured to download HTML pages and also plain text documents. The latter includes the source code of programs, and we found approximately 20,000 files with extensions of known programming languages. The distribution by file type is shown in Figure 13.

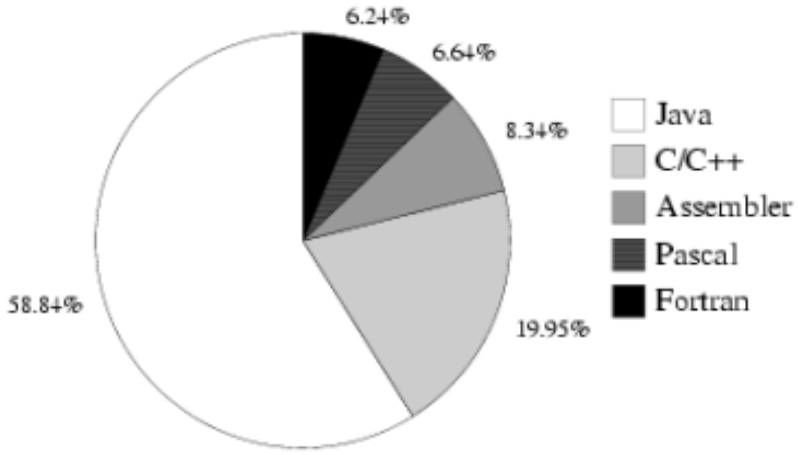

Figure 13: Distribution of links to source code and software.

2.8 Links to Web Pages

In this section, we consider the Web as a directed graph, in which each page is a node and each hyperlink is an edge. Using terminology from graph theory, the number of links received by a page is called its **internal degree**, and the number of links going out from a page is called its **external degree**. The distribution of both quantities is shown in Figure 14.

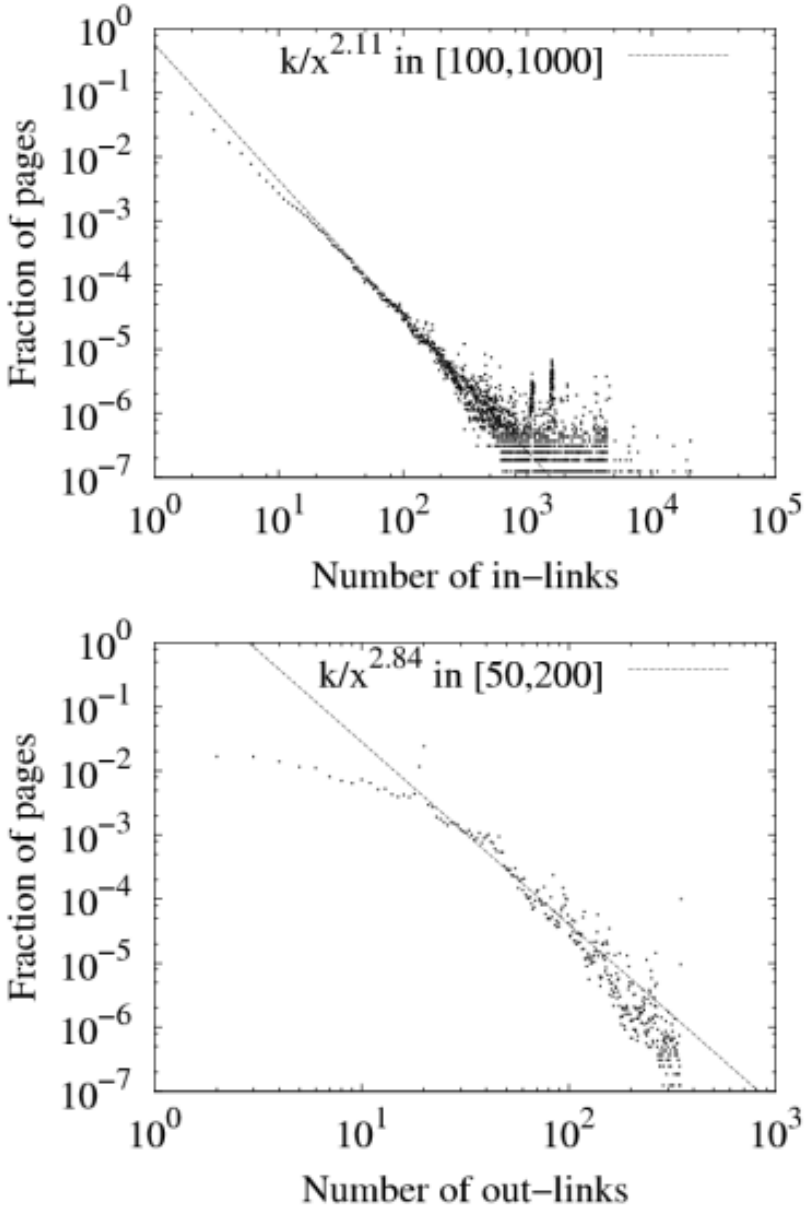

Figure 14: Distribution of internal and external degree of pages.

The internal degree of a page is a measure of its popularity on the Web, and it is beyond the control of the designer of a single Web site (except in the case of link farms). External degree is controlled by the designer of a Web site, and reflects how linked to the rest of the Web the author wants its page to be.

Adjusting a power-law distribution to the data, we obtain the parameters 2.11 for the internal degree and 2.84 for external degree. This is similar to the values that are observed for these parameters in another subsets of the Web, being the most usual values 2.1 and 2.7 [Pandurangan et al., 2002]. There are variations in this parameter among other national Web studies as in the sample of Africa (1.9; internal degree only) [Boldi et al., 2002], Brazil (1.6 and 2.6) [Modesto et al., 2005], Chile (1.8 and 4.1) [Baeza-Yates and Castillo, 2005a] and South Korea (2.0 and 3.3) [Baeza-Yates and Lalanne, 2004]. The Web graph is self-similar [Dill et al., 2002], so it is expected that the power-law distribution for the full Web can also be observed in smaller collections.

We can also see that there are certain anomalies: groups of pages sharing the same internal degree and the same external degree. These anomalies appear as increases in the frequency of pages having high degrees, and they are mostly due to spam pages, as has been observed previously [Fetterly et al., 2004, Thelwall and Wilkinson, 2003].

2.9 Ranking of Pages

There are several link analysis algorithms that attempt to infer how important a page is, by using information from its in-links. One of the most cited algorithm is Pagerank [Page et al., 1998]. PageRank can be understood in terms of persons browsing the Web in a random manner: every time they reach a pages, they decide whether to jump to a page at random (with probability ϵ) or to follow one of the links in the current page (with probability $1 - \epsilon$). In the latter case, any of the links in the page has the same probability of being chosen. The Pagerank algorithm simulates this process and returns the score of a page, which is the fraction of "time" that a user with this behavior would spend on each page.

In formal terms, this describes a Markovian process in which each page is a state and each hyperlink is a transition, and certain links between pages are added to avoid absorbing states. The Pagerank of a page is the probability of being on a page in the stationary state. The distribution of the scores obtained by applying the Pagerank algorithm to the Web pages of Spain is shown in Figure 15.

It is interesting to notice that because of the way in which Pagerank is calculated, using random jumps, even pages with very few in-links have a non-zero Pagerank value. The distribution of Pagerank scores also follows a power law, with parameter 1.96. For this parameter, a value of 2.1 has been observed in samples of the global Web [Dill et al., 2002], 1.9 in Brazil [Modesto et al., 2005], 1.9 in Chile [Baeza-Yates and Castillo, 2005a], and 1.8 in South Korea [Baeza-Yates and Lalanne, 2004].

In the figure, we can see an increase in the frequency of pages with a Pagerank value of about 2×10^{-5}. This is possibly due to Web page collusion [Baeza-Yates et al., 2005], a type of spam aimed at deceiving this type of link-analysis ranking.

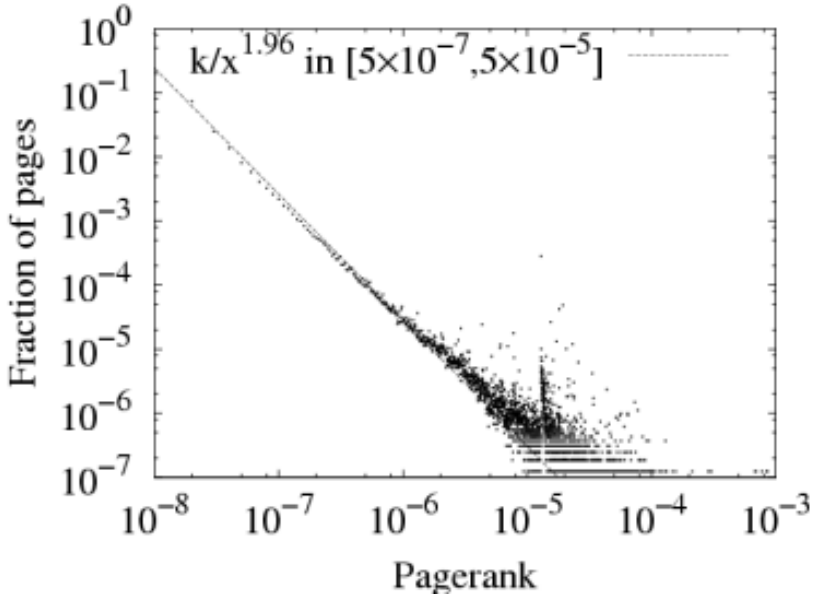

Figure 15: Distribution of Pagerank.

3 Web Site Characteristics

We define a Web site as a set of pages sharing the host-name part of the URL. Besides, we use the heuristic of considering both `http://www.site.ext/` and `http://site.ext/` as the same site.

3.1 Number of Pages

We observe an average of 52 pages per site. The distribution in the number of pages per Web site is very skewed, as shown in Figure 16.

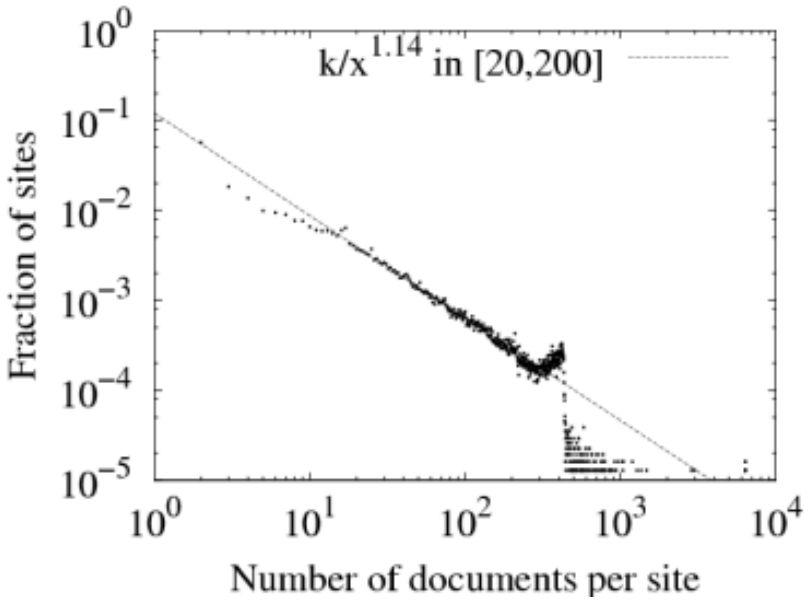

Figure 16: Distribution of the number of pages per Web site.

Close to 400 pages per site, there is a decrease in the frequency of the sites, as the crawler was configured to extract a maximum of 400 pages from the sites under `.com`.

Fitting a power-law to the central part of the distribution we obtain the parameter 1.14. This can be compared with 1.6 in Brazil [Modesto et al., 2005], 1.8 in Chile [Baeza-Yates and Castillo, 2005a] and 2.5 in South Korea [Baeza-Yates and Lalanne, 2004], meaning that in the Web of Spain there is relatively a smaller amount of larger sites. To better understand how skewed this distribution of pages per sites is, we can mention that just 27% of the sites have more than ten pages, 10% more than a hundred pages, and less than 1% more than a thousand pages.

3.1.1 Single-page Sites

There were 184,015 sites in which the crawler found only one Web page. This is a large fraction, about 60% of the sites, so we analyzed which was the reason for the crawler to not download more pages from those sites. We analyzed a sample of 30,000 sites and observed the following problems:

- The navigation of the site is built using Javascript, this is, it is necessary to parse and execute Javascript code to be able to browse the site.
- The starting page requires the Flash plug-in to be visualized, or relies on Java applets for navigation.

- The starting page is only a redirection to another site, or it uses frames and the `FRAME` HTML tag points to a different Web site, or it has only external links.
- The starting page has internal links, but these are malformed and could not be parsed by the crawler.

The fraction of sites in each case is shown in Figure 17. The fraction of sites that effectively have only one page, with no links, is close to 30%. Even the sites created only to reserve ("park" in the domain registry jargon) a certain address for a future Web site include some type of link for contact, or a link to the hosting provider.

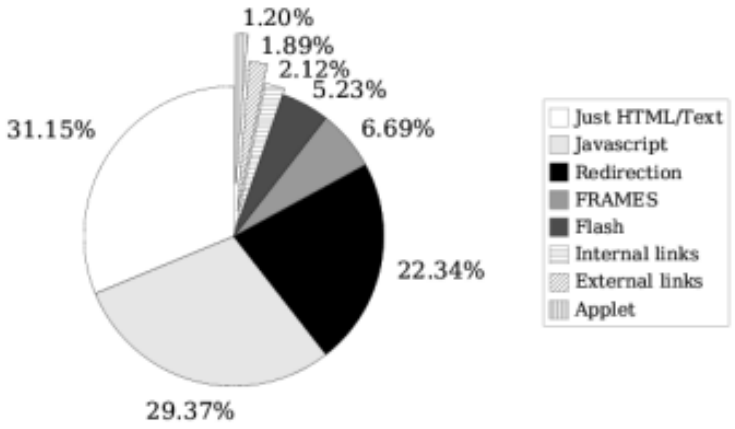

Figure 17: Distribution of the sites having only one page detected by the crawler.

This specific data is difficult to compare with other countries, as it is very sensitive to the type of crawler being used. If the crawler can parse redirections or frames, or understands simple Javascript navigation commands, then the percentage of sites with only one page is lower. The important issue is that in the Web of Spain there are at least 90,000 sites that use only Javascript or only Flash for navigating from the start page and therefore they are difficult or impossible to index by most search engines: this is around 30% of the pages of the Web of Spain. A detailed analysis by components on the Web graph is shown in Figure 26.

3.1.2 Sites with many Pages

We analyze also the sites that appear to have many pages. We inspected manually the sites with the larger number of pages. Normally they correspond to one of the following categories:

- The site uses a system that automatically generates pages. This includes Web directories that are kept automatically, mailing list archives or document repositories using a content management system.
- The site has malformed links, in the sense that they create recursion of Web pages, possibly using a system for generating dynamic pages that "hides" the parameters in the URL. This can be done on purpose, as a technique for spamming search engines, or it might be due to a programming error.
- The site is a copy of documentation. In these cases they are mostly documentation of the API of Java classes generated with `javadoc`, or Linux documentation.

Copies ("mirrors") of documentation appear as Web sites having many pages and also a large amount of text, so they can be detected easily in collections of this size.

3.2 Size of the Pages in a Whole Web Site

We consider in this section only the text of the pages we collected. The average size of a whole Web site (consider only the text) is approximately 146 Kilobytes. This is only a small fraction of the total information available on Web sites, as HTML structural and formatting tags, plus images and other resources, constitute an important part of the information available.

The distribution of the total size of the pages by Web sites is very skewed, as shown in Figure 18, and follows a power law with parameter 1.15 in its central part.

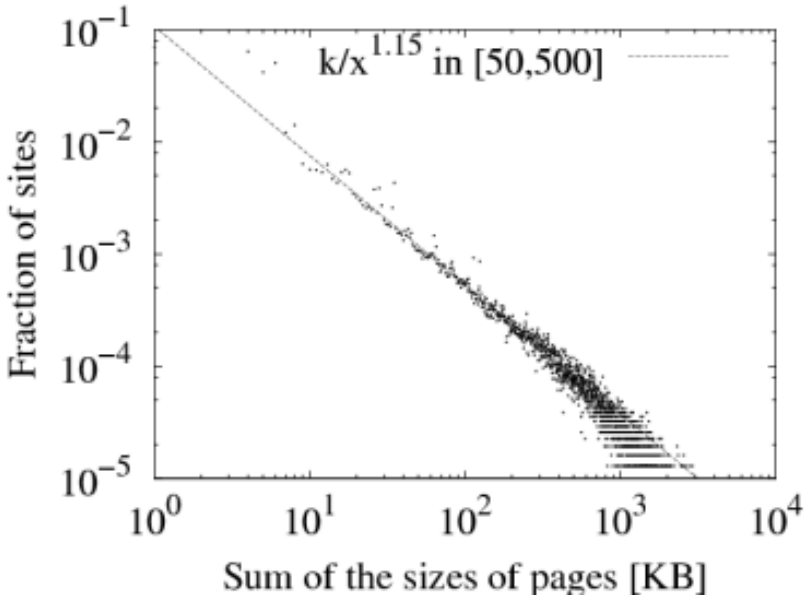

Figure 18: Total size per Web site, considering only the text.

.

Among the sites with the larger amount of textual contents, we found several replicas or *mirrors* of documentation. For instance, we found 6 copies of the "Request for Comments" technical notes `RFCXXXX`. We also found 7 copies of the LuCAS documentation ("*Linux en CAStellano*", "Linux in Spanish") , 30 copies of the Apache Tomcat documentation and 36 copies of the Linux Documentation Project (LDP), among others.

3.3 Internal Links

A link is considered **internal** if it points to another page inside the same Web site. The Web sites of Spain have on average 169 internal links. The distribution of the number of internal links per site is shown in Figure 19.

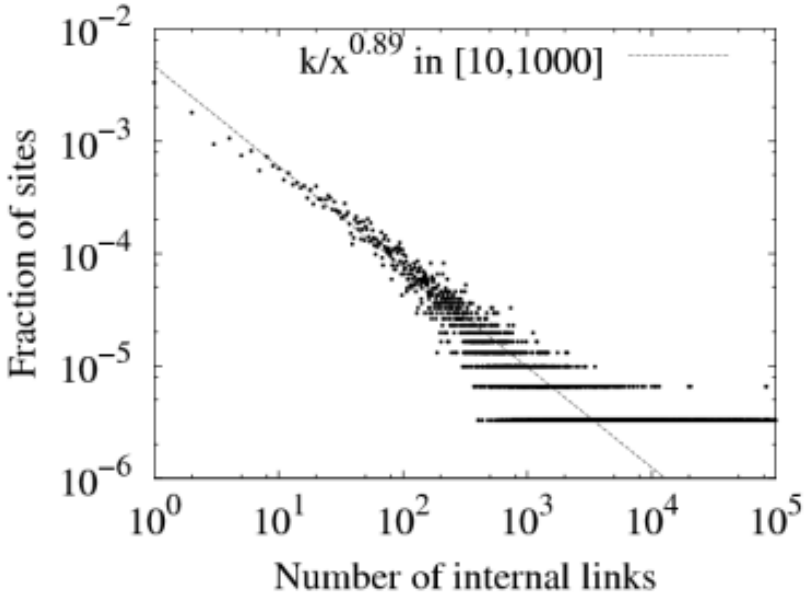

Figure 19: Distribution of the internal links per site.

This distribution is related to the distribution of pages per Web site; obviously, a Web site with very few pages cannot have too many internal links. For normalization, we calculated how many internal links *per page* each site has on average. The result is that an average Web site has approximately 0.15 internal links per page, or an internal link every 6 or 7 pages. There are many sites with an average number of internal links larger than this. This distribution is shown in Figure 20.

If we see the distribution of the number of internal links *per page*, there is no important correlation with the number of pages in a site, as shown in Figure 21. Different levels of internal connectivity are probably due to different reasons. For large Web sites, managing a large quantity of links might be difficult and require an automated system.

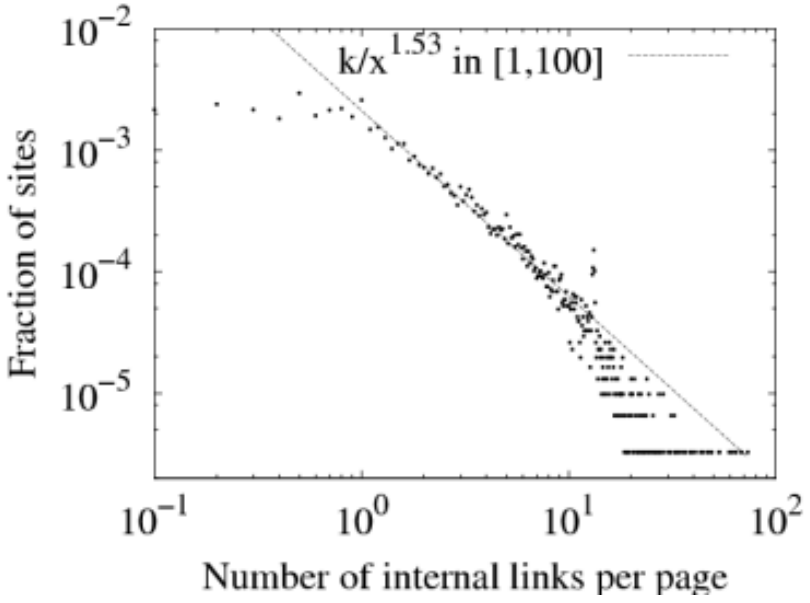

Figure 20: Distribution of the number of links per page.

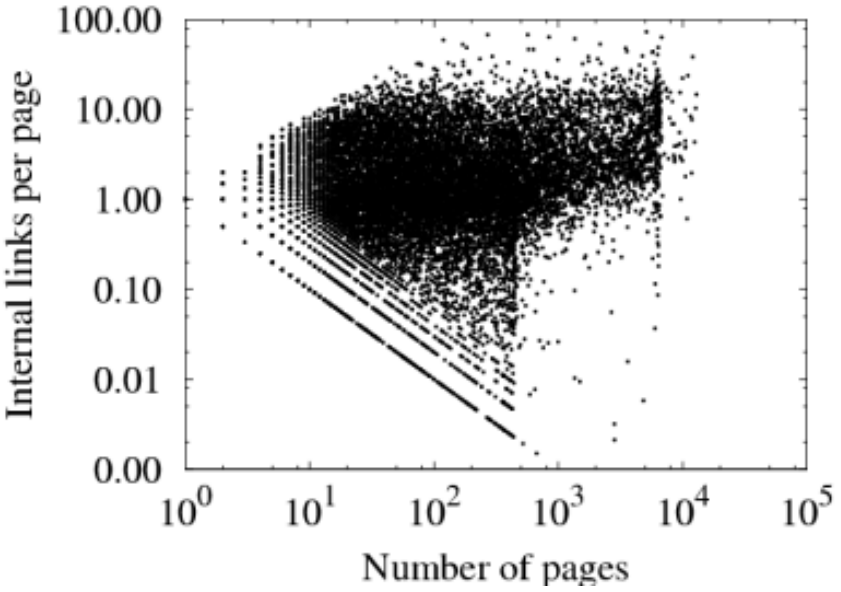

Figure 21: Distribution of the number of pages in a site against the average number of internal links per page in each site.

3.4 Links between Web Sites

In this section, we consider links between Web sites. A link between two Web sites represents one or several links between their pages, preserving the direction. This means that if there is a link, for instance, between `http://www.A.es/pageA.html` and `http://www.B.es/pageB.html`, then we say that there is a link between sites `www.A.es` and `www.B.es`; internal links are not considered. The resulting graph is also called the **hostgraph** [Dill et al., 2002].

To be fair when estimating the coverage of links to Web sites, we consider that it is better to discard one-page sites, as they might represent "under-construction" sites that are probably not worth linking. In the Web of Spain, there are 122,190 sites with more than one page. From them, 77,712 sites (63%) do not receive any reference from other site in Spain, and 109,787 (90%) have no link to other site in Spain. The distribution of the internal and external degree of Web sites also indicates a scale-free network, as shown in Figure 22.

The parameters obtained when fitting a power law are 1.82 and 1.34 for the internal and external degree respectively; this can be compared with 1.7 and 1.8 in Brazil [Modesto et al., 2005], 2.1 and 1.8 in Chile [Baeza-Yates and Castillo, 2005a], and 1.2 and 1.8 for South Korea [Baeza-Yates and Lalanne, 2004]. In the case of the global Web, an estimation of this parameter for the internal degree is 2.34 [Dill et al., 2002].

Among the sites with more in-links, we found mostly newspapers and universities. The sites with more out-links are mostly Web directories, and the coverage of these directories is very small: if we consider that there are over 300,000 sites and over 100,000 domains in the Web of Spain, then even the larger directories have a relatively small coverage (a maximum of about 5,000 sites is the largest observed value for outdegree in the Hostgraph); however, this can also be due to the fact that some directories are designed to avoid the downloading of a significant part of their collection of links by a Web crawler.

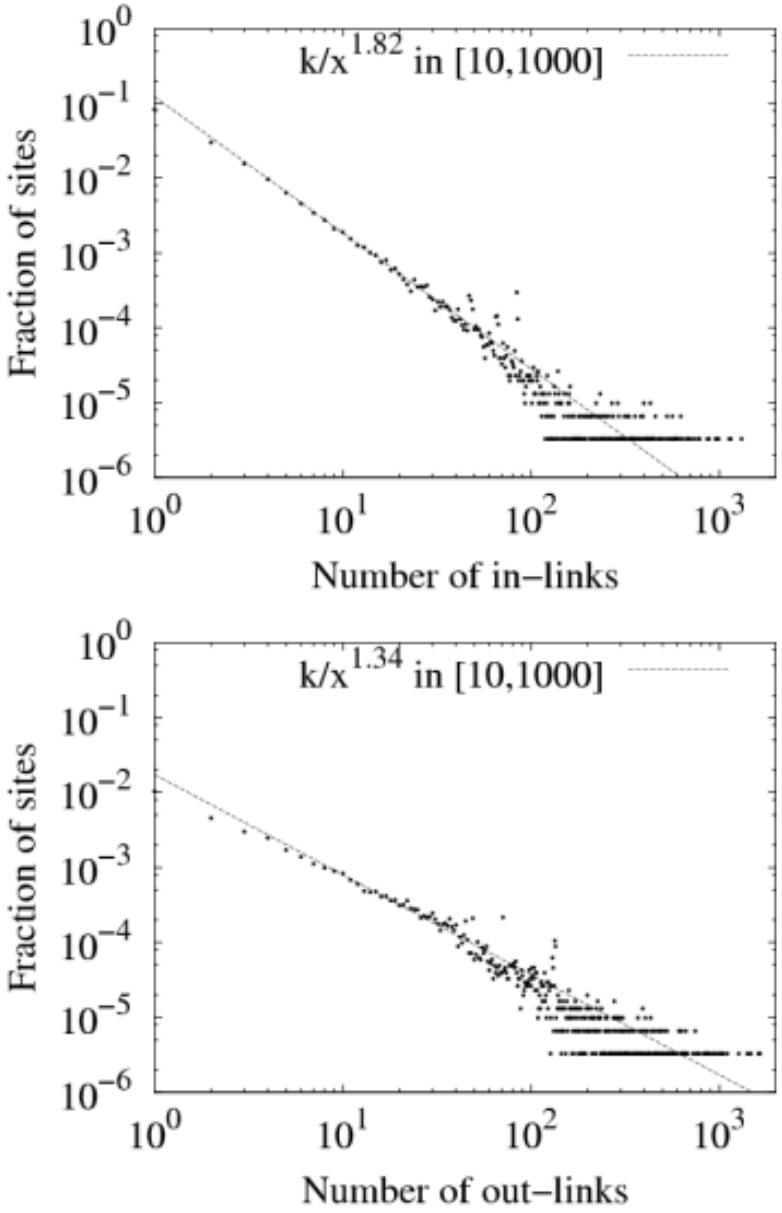

Figure 22: Distribution of the number of links between sites.

3.5 Summation of the Scores in Link-Based Ranking

A rather direct interpretation of Pagerank is that it represents the fraction of time that would be spend in each page by a person browsing the Web at random. As shown in Figure 15, this distribution is very skewed. It is natural to ask which would be the fraction of time this person would spend in each site, which corresponds to the sum of the Pagerank scores assigned to each page in a Web site. The resulting distribution is shown in Figure 23.

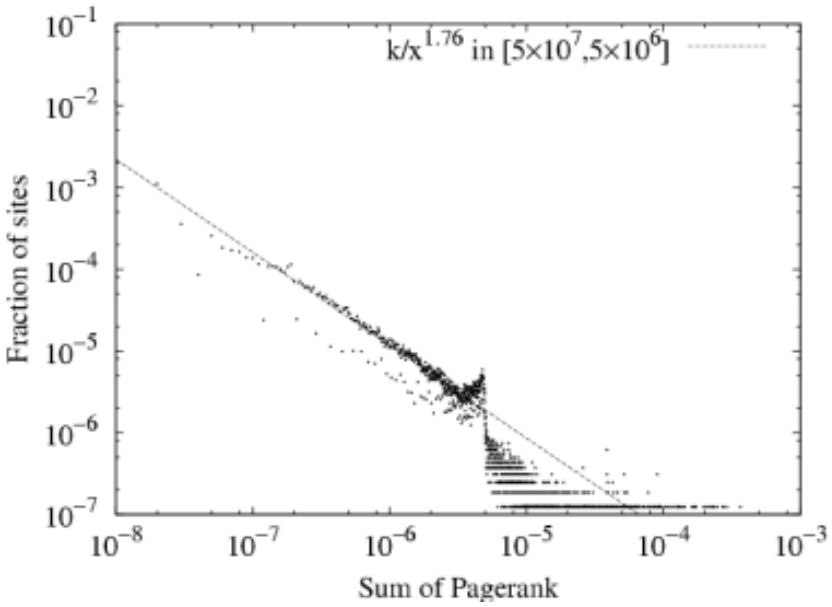

Figure 23: Distribution of the sum of Pagerank per Web site.

The distribution follows a power law with parameter 1.76. In our previous studies about the Web, we did not find a power-law as clear as in this case, probably because the collections were smaller, so probably the sum of Pagerank per sites requires at least 100,000 Web sites to be modeled properly by a power-law.

3.6 Strongly Connected Components

In a graph, it is said that a subset of the graph is a connected component if it is possible to go from each node in that subset to another node in the same subset by following links (in any direction). The subset is called a strongly connected component if this is possible by respecting the direction of the links. Not all of the Web of Spain –and not all of the Web of the world– is strongly connected.

We study the distribution of the sizes of the strongly connected components (SCCs) in the graph of Web sites. The distribution of the sizes of the components is presented in Table 2. A giant strongly connected component appears, as was observed by Broder et al. [Broder et al., 2000]. This is a typical signature of scale-free networks.

In this table, we consider in the components of size 1 only the sites that have at least one incoming or outgoing link. We note that there are four components having between 20 and 49 sites each that are probably link farms, but there is clearly a giant strongly connected component of more than 8,000 sites, this is about 15.1% of the nodes, and is very similar to the figure in Chile (15.3%) [Baeza-Yates and Castillo, 2005a] and South Korea (15.1%) [Baeza-Yates and Lalanne, 2004], and smaller than the observed in Brazil (23.3%) [Modesto et al., 2005].

Component size	Number of components
1	47,724
2	107
3	19
4	13
5	4
6	2
9	1
20	1
40	1
47	1
49	1
8,518	**(Giant component) 1**

Table 2: Size of the strongly connected components in the hostgraph.

When plotting the sizes of the components a power-law is observed, with parameter 3.84, as shown in Figure 24. This parameter can be compared with 4,23 observed in Chile [Baeza-Yates and Castillo, 2005a], 2,60 in South Korea [Baeza-Yates and Lalanne, 2004] and 2,81 in a sample of the global Web [Dill et al., 2002].

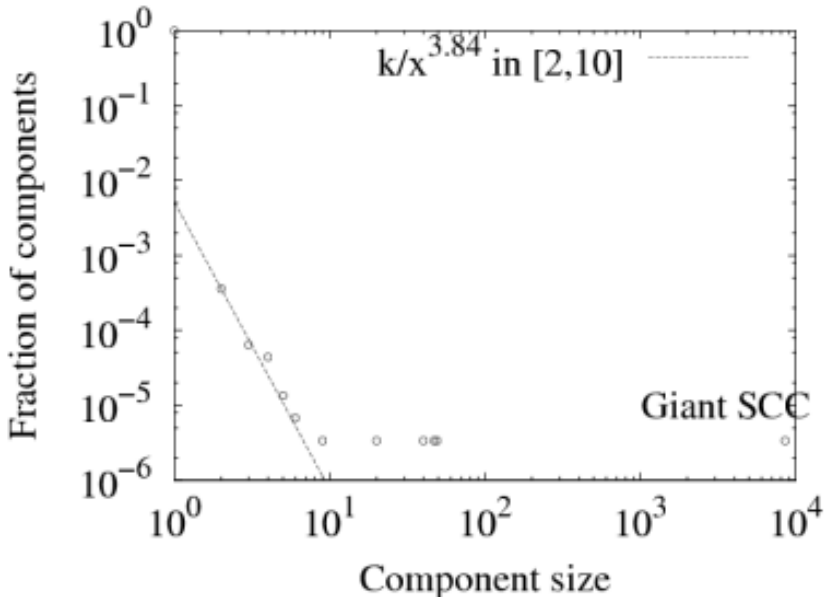

Figure 24: Distribution of the sizes of the strongly connected components in the hostgraph.

3.7 Link Structure among Web Sites

The giant strongly connected component appearing in Table 2 can be used as the starting point to distinguish certain structural components on the Web [Broder et al., 2000, Björneborn, 2004]:

(a) MAIN, sites on the strongly connected component;

(b) OUT, sites that are reachable from MAIN, but have no links towards MAIN;

(c) IN, sites that can reach MAIN, but that have no links from MAIN;

(d) ISLANDS, sites that are not connected to MAIN;

(e) TENDRILS, sites that only connect to IN or OUT, but following the links in the reverse direction;

(f) TUNNEL, a component joining OUT and IN without going through MAIN.

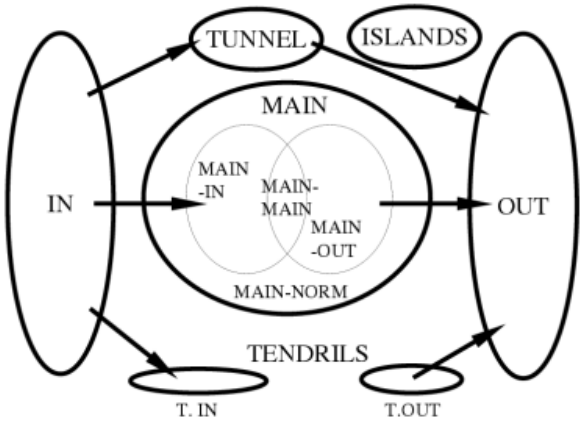

Figure 25: Graphical depiction of the macroscopic structure of the Web.

.

In [Baeza-Yates and Castillo, 2001] we extended this notation by separating MAIN into the following sub-components:

(g) MAIN-MAIN, sites that can be reached directly from IN, or that can reach directly sites in OUT;

(h) MAIN-IN, sites that can be reached directly from IN but are not in MAIN-MAIN;

(i) MAIN-OUT, sites that can reach directly OUT but do not belong to MAIN-MAIN;

(j) MAIN-NORM, sites that do not belong to the sub-components defined previously.

The distribution of Web sites in components is shown in Table 3. Note that the Web sites in the components IN and ISLANDS can only be found if the address of the starting pages of these sites are known beforehand, as these sites cannot be reached by following links. Also, we

Table 3: Distribution of sites in the components of the Web. The distribution of pages indicates the percentage of pages in the sites of each component.

	Sites		Pages	
Name of the component	Total	Only sites with links	Total	Only sites with links
MAIN-IN	0.15%	0.80%	0.83%	1.15%
MAIN-MAIN	0.77%	4.15%	18.54%	25.64%
MAIN-NORM	0.56%	2.99%	2.45%	3.39%
MAIN-OUT	1.31%	7.07%	19.49%	26.95%
MAIN (total)	2.79%	15.01%	41.31%	57.13%
IN	0.48%	2.59%	2.29%	3.17%
OUT	13.77%	74.05%	26.32%	36.41%
T. IN	1.09%	5.86%	1.18%	1.63%
T. OUT	0.18%	0.98%	0.39%	0.55%
TUNNEL	0.06%	0.30%	0.39%	0.55%
ISLANDS	81.63%	1.21%	28.12%	0.56%

give percentages over the total of sites, as well as only over sites with at least one in- or out-link. Finally, we also include the distribution of the number of pages in the sites in each component.

The distribution of sites on the components of the Web graph shows an important correlation with the distribution of other characteristics of the sites [Baeza-Yates and Castillo, 2001]. For instance, we studied the sites with only one page in Figure 17; now we can relate those sites with the components of the Web graph, as seen in Figure 26. In the component MAIN there are very few sites with only one page, while in the ISLANDS component they are approximately 50%.

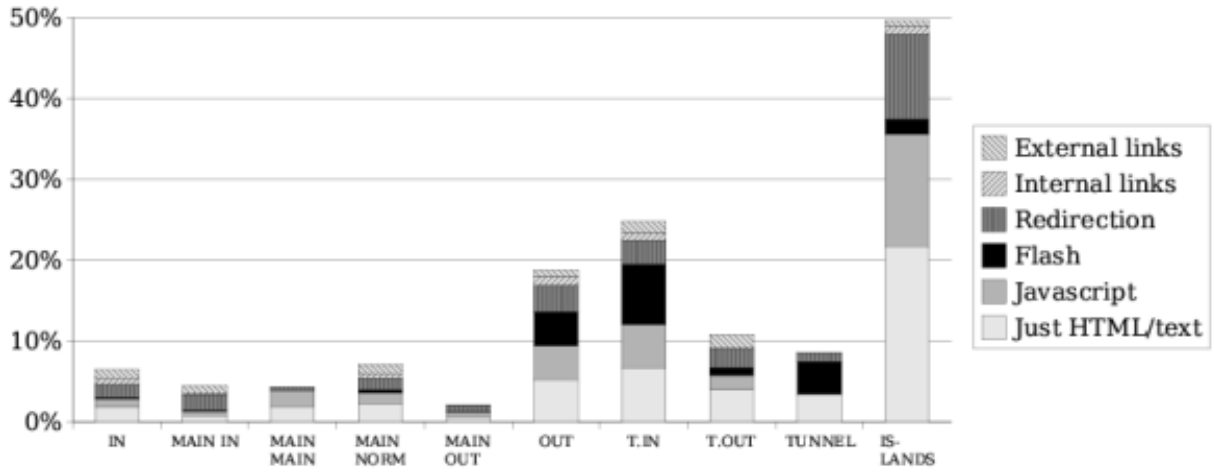

Figure 26: Distribution of sites with only one page according to the macroscopic structure of the Web.

.

Another characteristic that we study is in which top-level domains are the sites on each component. The result is shown in Table 4; we highlight that **all of the sites in MAIN** are under `.es`,

while in other top-level domains the component OUT is the most common. Also, the ISLANDS we found are roughly evenly split between `.es` and `.com`, while the latter has much more sites, so probably our starting URLs represent better the `.es` domain.

Table 4: Distribution of the domains in which the sites on each component are.

Component	Total	ES	COM	NET	ORG	Other
IN	2.59%	100.00%	0.00%	0.00%	0.00%	0.00%
MAIN (total)	14.01%	100.00%	0.00%	0.00%	0.00%	0.00%
OUT	74.05%	23.15%	55.04%	7.63%	11.88%	2.29%
T.IN	5.86%	24.68%	61.44%	6.85%	3.67%	3.37%
T.OUT	0.98%	100.00%	0.00%	0.00%	0.00%	0.00%
TUNNEL	0.30%	100.00%	0.00%	0.00%	0.00%	0.00%
ISLANDS	1.21%	46.86%	44.67%	4.96%	1.31%	2.19%

4 Domain Characteristics

We define the domain of a page as a suffix of its Web site name, using the following rule: if a site name has the form `www.A.es` or `www.xxx.A.es`, then the domain name is `A.es`.

For cases in which it is mandatory to register third or fourth level domains, we made an exception. This occurs with two providers of free sub-domains under `.es.vg` and `.es.fm`. Also, the domain `uk` does not allow direct registration, so sites owners have to choose a third-level domain under, for instance, `.co.uk`; there are a number of sites that use this extension, not only because of the commercial and diplomatic ties between Spain and the United Kingdom, but also due to the presence of Web sites from Gibraltar. Finally, the domain `eu.int` is used by a number of institutions within the European Union that are located in Spain. For the purposes of this study, these domains **are considered first-level domains**. For instance, if the site name is of the form `www.A.co.uk`, then the domain is `A.co.uk`.

In total, we found 118,248 different domains for the Web sites of Spain.

4.1 IP Address and Hosting Provider

We made a DNS search in the IP address of each of the studied sites, obtaining about 88% of the IP addresses. The addresses that were not found are sites that no longer exist or that were not reachable at the time of our experiment.

We grouped these addresses by domain, to count for how many different domains the same IP address was used. The distribution of the number of different domains per IP is shown in Figure 27.

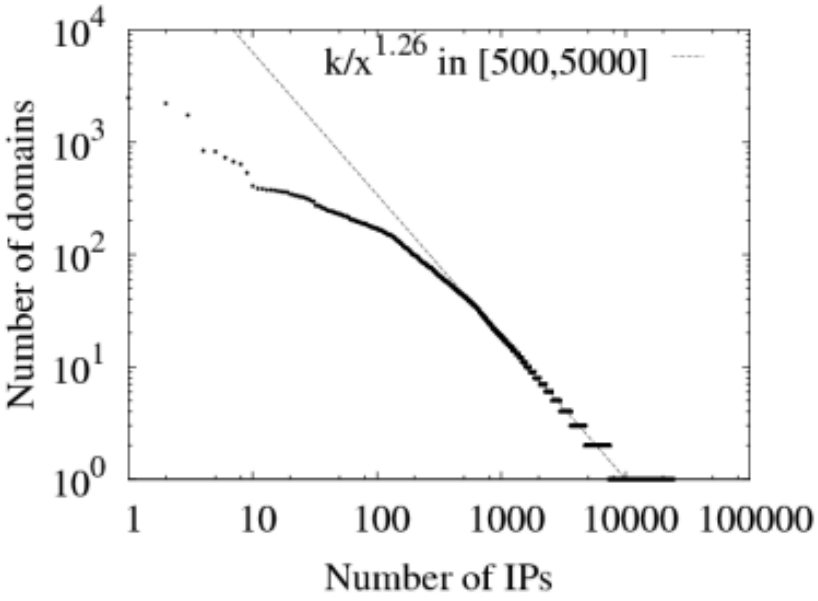

Figure 27: Distribution of the number of domains per IP address.

In total, there are about 24,000 IP addresses for the 118,000 domains, this means that each address has on average five domains, however, the distribution is very skewed: there are four IP addresses with more than 1,000 domains each, and 16,565 IP addresses with only one domain. A similar distribution, in terms of a few IP address concentrating most of the domains, was observed in the Web of Portugal [Gomes and Silva, 2003]. The power-law parameter 1.26 shows a distribution

that is much less skewed than, for instance, the Web of South Korea, in which the parameter is 2.76 [Baeza-Yates and Lalanne, 2004]. This means that there is more diversity and competition in terms of providers offering hosting in the Web of Spain.

4.2 Software used as Web Server

We checked for each IP address, which is the Web server software that is used and which is the operating system. This is done by issuing a `HEAD` HTTP request that receives a response such as this:

```
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 response is very comprehensive, including the server name (Apache), the version (1.3.33), the operating system (Linux) and the installed extensions (PHP and SSL). The distribution of server software is shown in Figure 28.

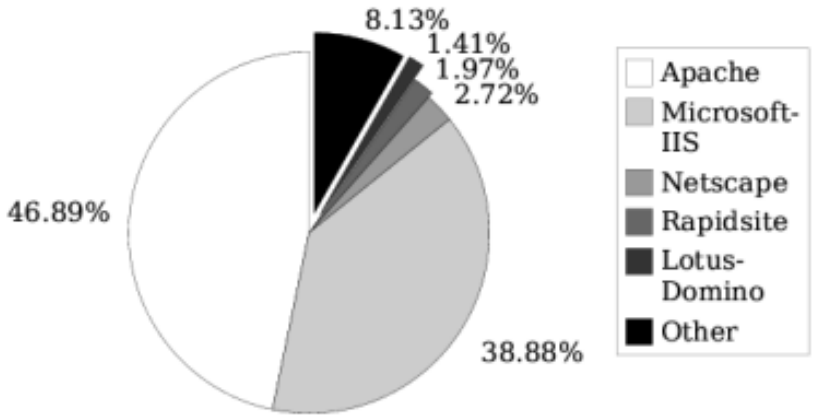

Figure 28: Distribution of the software used for Web server.

The two dominant software brands are Apache and Microsoft IIS (Internet Information Server), in that order. The data suggests that the market share of Microsoft IIS is larger in the Web of Spain that in the global Web: according to Netcraft[8], the proportion is 69% for Apache and 21% for Microsoft IIS.

We also studied the version of the Web server software that is most used, and the result is shown in Table 5.

Table 5: Distribution of the version of the most used Web server softwares.

Microsoft-IIS		Apache	
70.44%	5.0	57.95%	1.3
21.18%	6.0	27.86%	2.0
6.64%	4.0	1.18%	1.1
1.39%	5.1	0.32%	1.2
0.27%	3.0	0.04%	1.0

[8]Netcraft's Web server survey, available online at `http://news.netcraft.com/`.

The most modern stable versions, Apache 2.0 and Microsoft IIS 6.0 have been available during the last 2 or 3 years, and the fact that they still have a share of less than 30% in their user base indicates that the life cycle of Web server software is larger than other programs, such as Web browsers. A possible interpretation is that Web server administrators are much more conservative when updating their programs, specially when they keep several Web sites at once, so they prefer using older, more stable versions.

In regards to the operating system, we noted that in about 16% of the cases the Web server response does not include an indication about the operating system, possibly because of security concerns. The distribution of operating systems is shown in Figure 29.

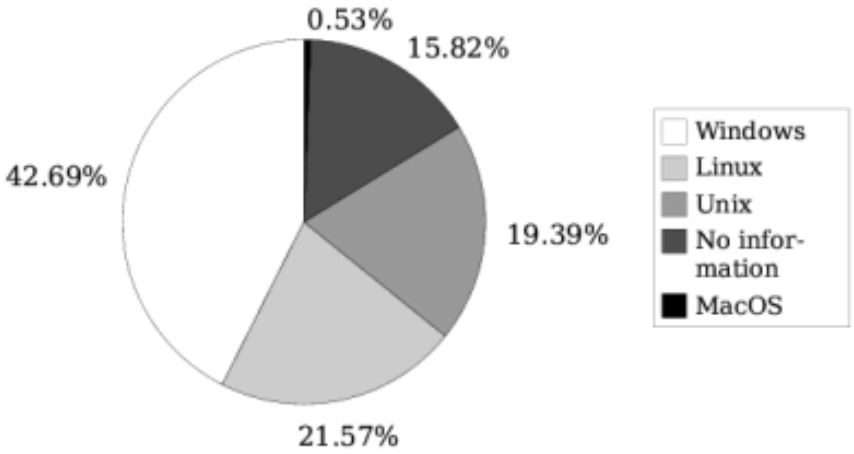

Figure 29: Distribution of the operating system by IP address.

The most used operating system used for Web servers in Spain is Windows (43%), followed closely by operating systems based in Unix (41%); this means that at least 15% of the servers based in Windows prefer Apache as a Web server. In the case of Chile, the relative positions of the usage of Web servers are inverted, with 31% for Windows and 57% for Unix [Baeza-Yates and Castillo, 2005a].

4.3 Number of Sites per Domain

On average we found 2.55 sites per domain, but there are several very large domains. For instance, we found almost 30 domains with more than 1,000 sites in each one. On the other hand 111,415 domains (about 92%) have only one site.

The distribution of the number of sites for each of the 10,000 larger domains is shown in Figure 30.

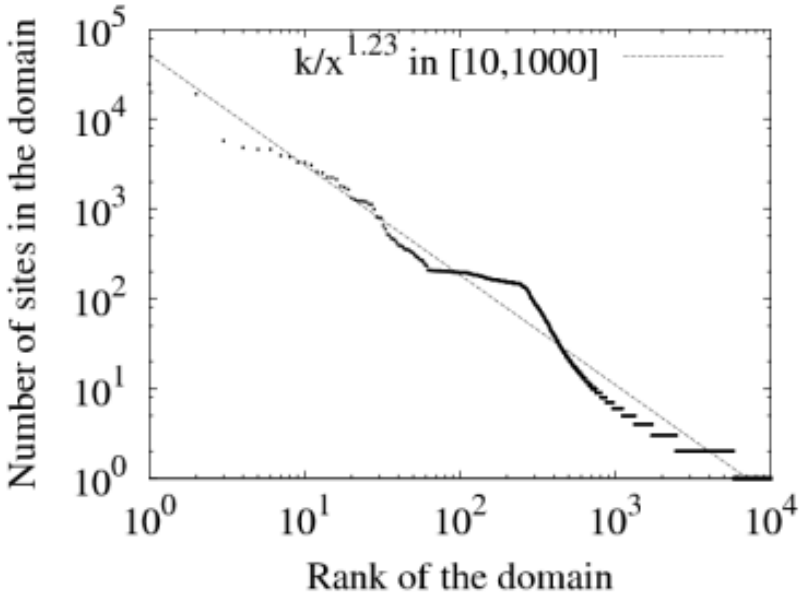

Figure 30: Distribution of the number of pages per domain.

Practically all of the 30 larger domains (which we inspected manually) have its domain server configured to use *DNS Wildcarding* [Barr, 1996], this is, they are configured to reply with the same IP address no matter which domain name is used. For instance, `http://X.bcnlink.com/`, for each string "`X`", always returns the same IP address, and the resulting Web page is always the same.

4.4 Number of Pages per Domain

There is an average of 133 pages per domain. The distribution of the number of pages per domain exhibits a power-law with parameter 1.18 in its central part. Figure 31 shows the distribution of this variable, which is very similar to the distribution of pages per site shown in Figure 16, which has an average of 52 pages per site.

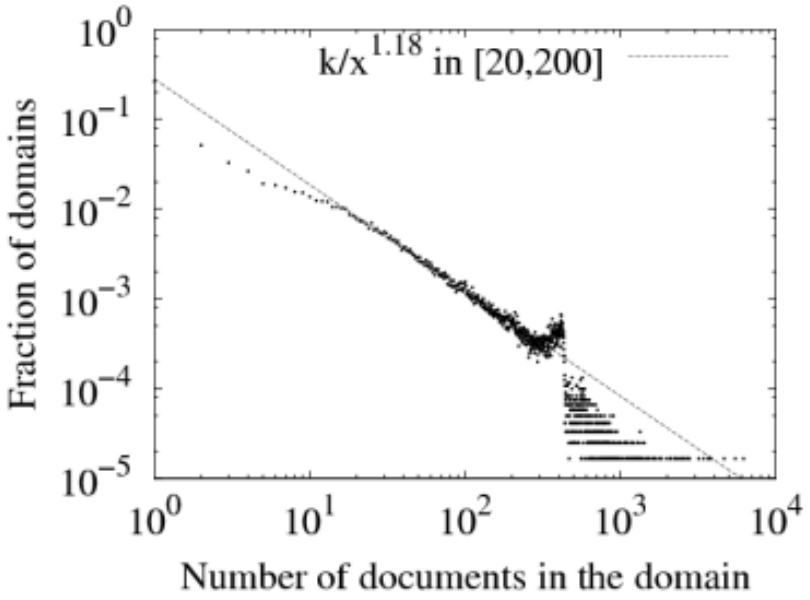

Figure 31: Distribution of the number of pages per domain.

There are 32,008 domains with only one Web page, which represents only 26% of the domains. This number is much lower than the 60% of the sites that have only one page; possibly this is due to the fact that creating a new site once you own the domain name has no cost, while purchasing a new domain site has a cost.

4.5 Total Size of the Domains

The average size of a Web domain, considering only text, is of approximately 373 Kilobytes. The distribution of the total size of pages per domain is shown in Figure 32, and follows a power-law with parameter 1.19 in its central part.

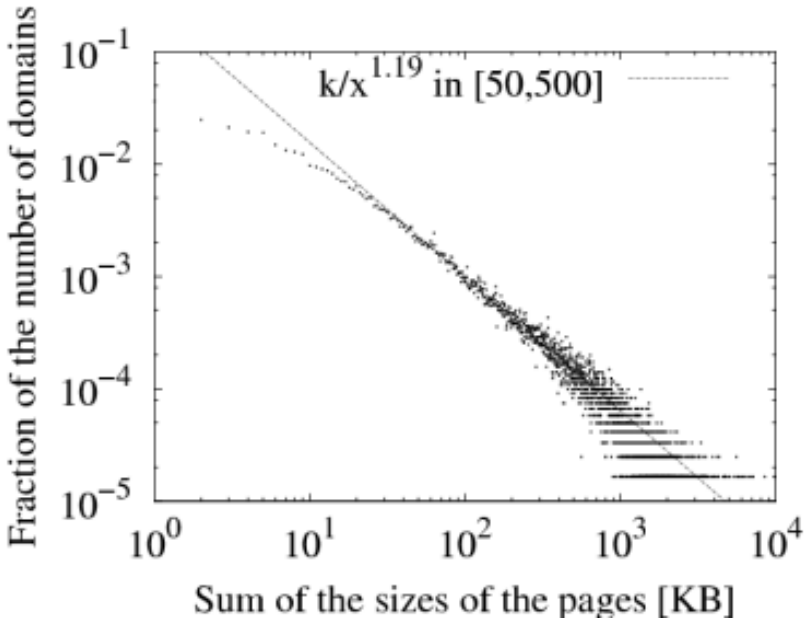

Figure 32: Total size per domain, considering only the text.

.

Most of the large domains in terms of text are universities, research centers and databases for academic use. This is similar to the case of Chile [Baeza-Yates and Castillo, 2005a] and Thailand [Sanguanpong et al., 2000] in which there is also a strong presence of academic Web sites; on the contrary, for instance, in South Korea, the majority of Web sites is of commercial type [Baeza-Yates and Lalanne

4.6 Page Titles inside a Domain

In Figure 5 we show that only 16% of the page titles are unique, and there is a large amount of repeated titles and untitled pages. We now focus on the distribution of different titles per Web site.

For this, we measure the ratio between the number of titles and the number of pages on a site. For instance, if a Web site has 10 pages and 4 different titles, then the value of this parameter is 0.25. In the Figure 33 we study if it is related to the size of the Web sites.

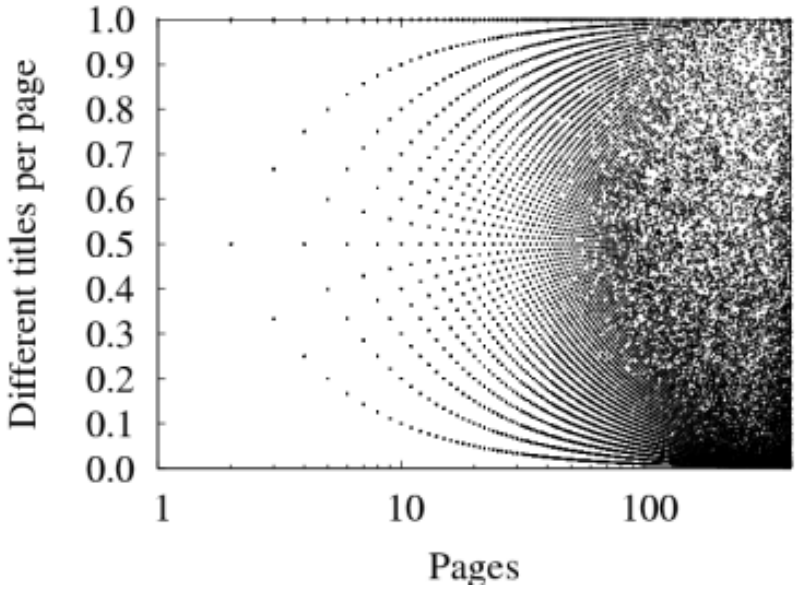

Figure 33: Distribution of the number of different titles versus the number of pages in each domain.

In general, we do not see a significant correlation between these two variables: a large site can have the same proportion of different titles as a small one, as this parameter depends more on the quality of the design of a Web site that on its magnitude. However, the density is higher towards the lower part, meaning that it is slightly more difficult in large domains to keep several different titles for the pages.

4.7 Links between Domains

Next, we measure the number of links between domains, with the purpose of obtaining a graphical representation of the relationship between domains. In Figure 34 we have included the 50 domains that receive more links in the Web of Spain. In relation to the year 2002 [Baeza-Yates, 2003], we see more government-related sites with a large number of references.

We used the program `neato` of the `graphviz` package[9]. Using a spring model and an iterative algorithm, this program finds a low energy configuration for the graph. The program's input includes the minimal length for each edge, than in our case is inversely proportional to the number of links between the domains that particular edge connects.

Besides, we have divided the domains in three classes: commercial (rectangles), educational (ellipses) and government (diamonds). In this graph, a thicker line represents a larger number of

[9]GraphViz, a graph visualization software, available online at `http://www.graphviz.org/`.

links, and we can clearly see that domains on the same class tend to group together, even if we consider than in some cases it is not usual to link to similar sites; for instance, between competing newspaper usually there are very few links.

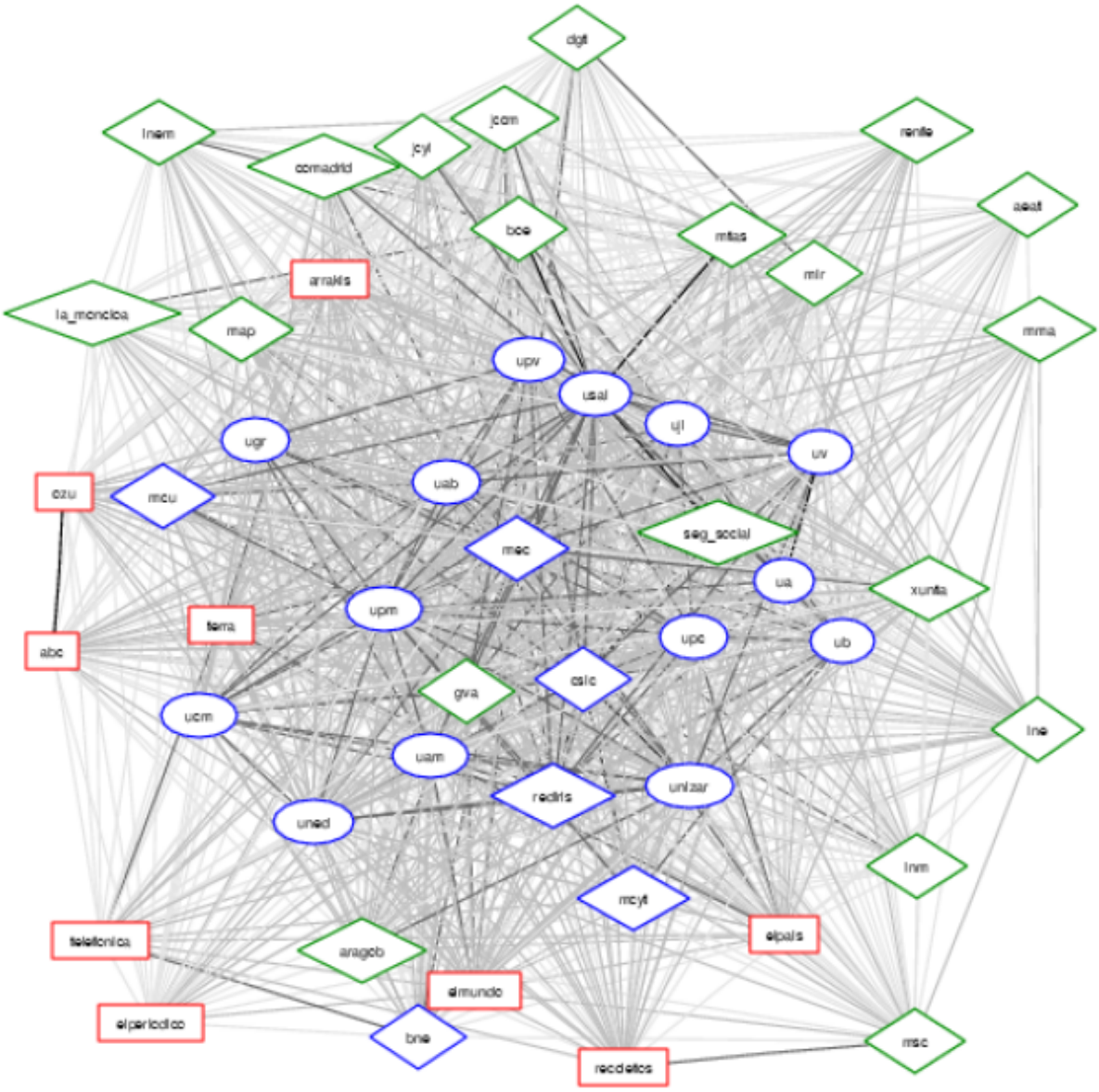

Figure 34: Graphical depiction of the links between domains.

4.8 First-level Domains of the Web Sites of Spain

Our collection of pages includes servers that are physically located in Spain; but this does not always mean that they are under the `.es` top-level domain. In Table 6 we show the distribution of these domains.

Naturally the largest top-level domains: `.com`, `.org`, `.net`, etc. are the most used. It is interesting that the even the "recent" generic top-level domains as `.info` and even `.aero` are frequently used. On the other hand, there are domains such as `.tv` or `.fm` that are often used because they are easy to remember for TV or radio stations.

Table 6: Most used top-level domains (TLDs) in the Web of Spain.

TLD	Name	% domains	% pages
`com`	Commercial (generic)	65.026%	31.436%
`es`	Spain	15.965%	56.033%
`org`	Organization (generic)	7.581%	5.950%
`net`	Network (generic)	7.387%	4.954%
`es.vg`	Hosting provider, Virgin Islands	1.784%	0.027%
`info`	Information (generic)	0.816%	0.690%
`es.fm`	Hosting provider, Micronesia	0.306%	0.002%
`biz`	Business (generic)	0.290%	0.105%
`tv`	Tuvalu	0.144%	0.076%
`to`	Tonga	0.088%	0.017%
`us`	United States of America	0.053%	0.046%
`ws`	Western Samoa	0.050%	0.024%
`pt`	Portugal	0.046%	0.039%
`cc`	Cocos Islands	0.046%	0.025%
`edu`	Educational (generic)	0.039%	0.122%
`ad`	Andorra	0.037%	0.016%
`as`	American Samoa	0.031%	0.027%
`co.uk`	Commercial, United Kingdom	0.028%	0.046%
`coop`	Cooperatives (generic)	0.024%	0.019%
`de`	Germany	0.019%	0.009%
`fm`	Micronesia	0.017%	0.009%
`cu`	Cuba	0.017%	0.061%
`nu`	Niue	0.016%	0.024%
`cl`	Chile	0.015%	0.013%
`name`	Person (generic)	0.013%	0.012%
`bz`	Belize	0.013%	0.010%
`it`	Italy	0.012%	0.011%
`nl`	Netherlands	0.010%	0.008%
`fr`	France	0.009%	0.008%
`tk`	Tokelau	0.008%	0.010%

4.9 External top-level Domains

We found links to approximately 50 million different sites outside Spain. For each of the external sites found, we extracted its top-level domain. The top 30 most linked top-level domains are shown in Table 7. This distribution is similar to the one of the global Web for `.com`, `.net` and `.org`[10]; in the second column, we show the global ranking of each domain in terms of its number of servers. For instance, the `.de` domain is the 5^{th} in terms of receiving links from the Web of Spain, and the 7^{th} in terms of number of sites in the global Web.

Table 7: Fraction of links to the top 30 most referenced top-level domains (TLDs).

Ranking				
Spain	Global	TLD	Nam	Percent of sites
1	2	`com`	Commercial (generic)	49.99%
2	25	`org`	Organization (generic)	8.69%
3	1	`net`	Network (generic)	6.07%
4	176	`tk`	Tokelau	3.25%
5	7	`de`	Germany	3.13%
6	5	`edu`	Educational (generic)	2.71%
7	10	`co.uk`	Commercial U.K.	2.31%
8	4	`it`	Italy	1.85%
9	8	`fr`	France	1.20%
10	12	`ca`	Canada	0.91%
11	6	`nl`	Netherlands	0.90%
12	21	`ch`	Switzerland	0.82%
13	3	`jp`	Japan	0.79%
14	16	`us`	U.S.A.	0.67%
15	14	`se`	Sweden	0.58%
16	42	`cl`	Chile	0.57%
17	10	`ac.uk`	Academic U.K.	0.49%
18	17	`be`	Belgium	0.48%
19	19	`dk`	Denmark	0.47%
20	37	`pt`	Portugal	0.44%
21	9	`au`	Australia	0.42%
22	31	`gov`	Government U.S.A.	0.42%
23	74	`info`	Information (generic)	0.42%
24	27	`ru`	Russia	0.41%
25	10	`org.uk`	Organizations U.K.	0.38%
26	23	`at`	Austria	0.37%
27	13	`pl`	Poland	0.33%
28	26	`no`	Norway	0.32%
29	67	`biz`	Business (generic)	0.32%
30	11	`br`	Brazil	0.28%

[10]Internet Domain Survey from the Internet Systems Consortium, available online at `http://www.isc.org/ds/`.

Half of the external sites linked from the Web of Spain are located in the `.com` domain, as shown in Table 7. The generic top-level domains `.org`, `.info` and `.biz` appear with much more frequency that expected by the number of host names in each of these domains.

A similar connectivity study that was made at the level of Web sites and involved several countries [Bharat et al., 2001] showed that the most referenced sites from the `.es` domain in 2001 were in Germany, the United Kingdom, France and the `.int` domain of international organizations. This is consistent with our findings, and the most referenced domains have cultural, economical or geographical ties with Spain.

Figure 35 shows the distribution of links to external domains. A power-law with parameter 1.80 can be obtained, even when the top 10 more important domains do not fit well to the model. Note that the graph continues beyond the 200 or 300 existent domains as there are many typographical errors in domain names, for instance "`.orq`" or "`.con`".

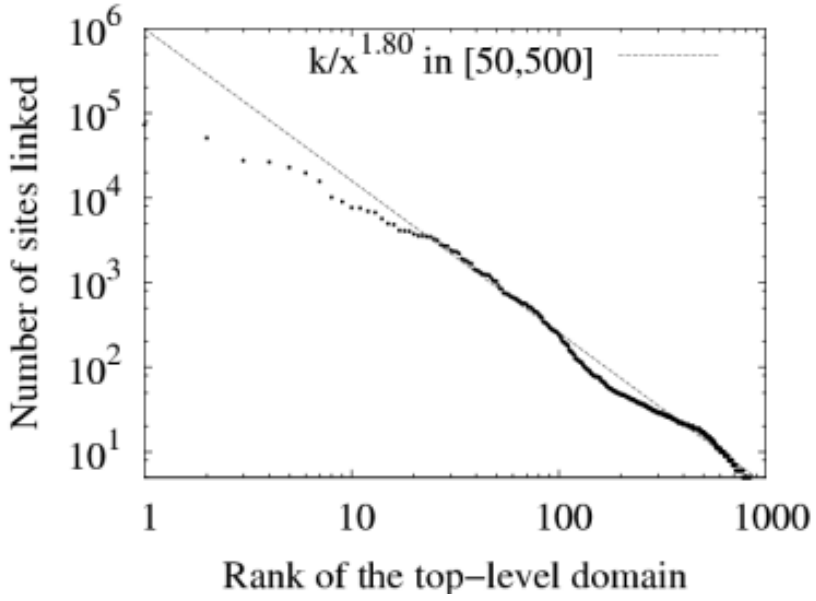

Figure 35: Frequency of links to external sites, grouped by top-level domain.

5 Conclusions

Our collection from the Web of Spain has over **300,000** Web sites, and these sites contain more than **16 million pages**. With respect to the Web graph, we obtained statistics that are very similar to the ones from other samples, which indicates that from a subset of the Web we can obtain a good approximation of the characteristics of the global Web graph.

Our analysis also demonstrates the heterogeneity of the Web, which from the user's point of view is positive due to its diversity in terms of topics, authorships, genres, etc. but at the same time negative due to its quality. We found many sites that were isolated, had very small textual contents, very few references, broken links, large fractions of duplicated content, among other issues.

A study about the Web of a country has many applications. The most obvious one is to help in the development of better search engines, in particular, in the development of data structures for storing information about the Web and to rank search results. The main findings of our research on the Web of Spain are summarized below:

The country-code top-level domain While the domain suffix assigned to Spain is `.es`, there is a large quantity of Web sites that do not use the top level domain of the country `.es`, but prefer to use `.com` or `.net`. The top level domain where most of the Web sites of Spain are located is `.com` (66%), followed by `.es` (16%). However, if we count the number of pages, we have 31% for `.com` and 56% for `.es`. Web sites in `.es` have more content per site, are better connected and have much less spam than the sites of Spain in other domains.

This means that, while the chief constituent of the Web of Spain is the `.es` domain, there are many sites that also belong to Spain but are outside this domain, and those sites have to be taken into account for characterizing this collection. It is likely that the same is true for other national Webs that cannot be defined only by their corresponding country-code top-level domain.

The fact that several requisites are requested for obtaining a `.es` domain has kept the domain less used and relatively free from bad practices such as link spam or "cybersquatting" (registering a domain name with the intent of selling it to its rightful owner). This eventually could be irrelevant for the users of the Web of Spain, as our impression is that very few people verify the browser's address bar to see if the host part ends in `.com` or `.es`. We do not have data related to the number of visits received by each site, but we can infer that due to the fact that the Web sites of Spain under `.es` have more content and are in the better-connected parts of the Web, they probably receive a larger share of visits.

Languages Approximately 50% of the pages in Spain are in Spanish, followed by 30% in English and 8% in Catalan. The contents in Galician and Basque (the other co-official languages in Spain) only comprise around 2% of the pages. The large amount of English pages is explained partially because of tourism Web sites, and partially because of large collections of technical documentation in English.

During this study, we discovered over 250,000 pages in Catalan, that we used to create CucWEB[11], a corpus of Catalan pages on the Web annotated with linguistic information. We are also obtaining and processing a corpus of pages in spanish, as we consider that the Web can be used as a linguistic corpus as long as one can understand that it has some drawbacks. In particular, "the Web is not representative of anything else. But neither are other corpora, in any well-understood sense" [Kilgarriff and Grefenstette, 2004].

[11] CucWEB: a Catalan corpus from the Web, available online at `http://www.catedratelefonica.upf.es/`.

Contents and popularity The majority of the most referenced domains belong to governmental or academic sources, and this is particularly true for Catalan pages. This might be due to the fact that they were established earlier than other sites, but can also reflect that both universities and government sites are very important in terms of number of pages and information content. Indeed, a large fraction of the information available on the Web of Spain (except for duplicates) is generated by these types of sites. Newspapers also have an important share in both the number of pages and in number of references.

Relationship with other countries From Spain, the most referenced country-code domains from Spain are Germany, the United Kingdom, Italy, France and Canada. These relationships express cultural, geographical and economical ties. Further work is needed to compare this list sistematically with "real world" data such as the volume of commercial trade or travel to those countries to and from Spain.

Web server technologies Operating systems for Web servers are divided in 43% of Windows-based operating systems, and 41% of Unix-based, including Linux. Although the Apache Web server is the dominant application for serving Web pages, as it is in the global Web, the share of Microsoft's Web server (IIS) is larger in the Web of Spain, which reflects a larger share of Microsoft in the software used by hosting providers than in the full Web.

The most used programming language for dynamic pages is PHP with a 46% share, followed by ASP with 41% of pages. Other programming languages, such as general-purpose programming languages like perl or even Web specific languages such as Java Server Pages (`.jsp`) are much less used.

File formats While most of the documents on the Web are written in HTML, there are also other document formats. The most important ones are Adobe Portable Document Format (PDF) and plain text, each one with about 40% of share. Open, non-proprietary formats for documents are prefered on the Web of Spain.

Findability of information About 60% of the sites on the Web of Spain have only one indexable Web page following regular links, and about half of them have other pages, but those pages are difficult or impossible to access by current Web search engines. Search engines have to be able to parse at least trivial Javascript code to be able to find more pages.

As for Web directories, 63% of the studied Web sites are not linked to by other Web site in Spain, which makes them harder to find; we also found that no directory of pages in Spain has a large coverage, in terms of linking to a significant fraction of different domains inside the Web of Spain.

Finally, most Web pages had repeated or default titles, with only about 10% of the pages having an unique title. It is likely that an even smaller fraction of pages have metadata associated to them.

Acknowledgements

Maria Eugenia Fuenmayor and Paulo Golgher managed the Web crawler during the process of page downloading. The classification of pages by languages was made by Bárbara Poblete, Gemma Boleda, Stefan Bott and Toni Badia. We also thank anonymous reviewers for their comments and suggestions.

This project was funded by Cátedra Telefónica de Producción Multimedia, Universitat Pompeu Fabra.

References

[Alonso et al., 2003] Alonso, J. L., Figuerola, C. G., and Zazo, á. F. (2003). *Cibermetría: nuevas técnicas de estudio aplicables al Web*. Ediciones TREA, Spain.

[Arasu et al., 2001] Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., and Raghavan, S. (2001). Searching the web. *ACM Transactions on Internet Technology (TOIT)*, 1(1):2–43.

[Baeza-Yates, 2003] Baeza-Yates, R. (2003). The web of spain. *UPGRADE*, 3(3):82–84.

[Baeza-Yates and Castillo, 2000] Baeza-Yates, R. and Castillo, C. (2000). Caracterizando la web chilena. In *Encuentro chileno de ciencias de la computación*, Punta Arenas, Chile. Sociedad Chilena de Ciencias de la Computación.

[Baeza-Yates and Castillo, 2001] Baeza-Yates, R. and Castillo, C. (2001). 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. IEEE CS Press.

[Baeza-Yates and Castillo, 2004] Baeza-Yates, R. and Castillo, C. (2004). Crawling the infinite web: five levels are enough. In *Proceedings of the third Workshop on Web Graphs (WAW)*, volume 3243 of *Lecture Notes in Computer Science*, pages 156–167, Rome, Italy. Springer.

[Baeza-Yates and Castillo, 2005a] Baeza-Yates, R. and Castillo, C. (2005a). Características de la web chilena 2004. Technical report, Center for Web Research, University of Chile.

[Baeza-Yates and Castillo, 2005b] Baeza-Yates, R. and Castillo, C. (2005b). Characterization of national web domains. Technical report, Universitat Pompeu Fabra.

[Baeza-Yates et al., 2005] Baeza-Yates, R., Castillo, C., and López, V. (2005). Pagerank increase under different collusion topologies. In *First International Workshop on Adversarial Information Retrieval on the Web*.

[Baeza-Yates and Lalanne, 2004] Baeza-Yates, R. and Lalanne, F. (2004). Characteristics of the korean web. Technical report, Korea–Chile IT Cooperation Center ITCC.

[Baeza-Yates and Poblete, 2003] Baeza-Yates, R. and Poblete, B. (2003). Evolution of the chilean web structure composition. In *Proceedings of Latin American Web Conference*, pages 11–13, Santiago, Chile. IEEE CS Press.

[Baeza-Yates et al., 2003] Baeza-Yates, R., Poblete, B., and Saint-Jean, F. (2003). Evolución de la web chilena 2001–2002. Technical report, Center for Web Research, University of Chile.

[Barabási, 2001] Barabási, A. L. (2001). The physics of the web. *PhysicsWeb.ORG, online journal*.

[Barabási, 2002] Barabási, A.-L. (2002). *Linked: The New Science of Networks*. Perseus Books Group.

[Barábasi et al., 2001] Barábasi, A.-L., Ravasz, E., and Vicsek, T. (2001). Deterministic scale-free networks. *Physica A*, 299(3-4):559–564.

[Barr, 1996] Barr, D. (1996). RFC 1912: Common DNS operational and configuration errors. http://www.ietf.org/rfc/rfc1912.txt.

[Benczúr et al., 2003] Benczúr, A. A., Csalogány, K., Fogaras, D., Friedman, E., Sarlós, T., Uher, M., and Windhager, E. (2003). Searching a small national domain – a preliminary report. In *Poster Proceedings of Conference on World Wide Web*, Budapest, Hungary.

[Bharat et al., 2001] Bharat, K., Chang, B. W., Henzinger, M., and Ruhl, M. (2001). Who links to whom: Mining linkage between Web sites. In *International Conference on Data Mining (ICDM)*, San Jose, California, USA. IEEE CS.

[Bharat and Henzinger, 1998] Bharat, K. and Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 104–111, Melbourne, Australia. ACM Press, New York.

[Björneborn, 2004] Björneborn, L. (2004). *Small-World Link Structures across an Academic Web Space – a Library and Information Science Approach*. PhD thesis, Royal School of Library and Information Science, Copenhagen, Denmark.

[Björneborn and Ingwersen, 2004] Björneborn, L. and Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14):1216–1227.

[Boldi et al., 2002] Boldi, P., Codenotti, B., Santini, M., and Vigna, S. (2002). Structural properties of the African Web. In *Proceedings of the eleventh international conference on World Wide Web*, Honolulu, Hawaii, USA. ACM Press.

[Broder et al., 2000] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web: Experiments and models. In *Proceedings of the Ninth Conference on World Wide Web*, pages 309–320, Amsterdam, Netherlands. ACM Press.

[Caldarelli et al., 2002] Caldarelli, G., Capocci, A., , and Muñoz, M. A. (2002). Scale-free networks from varying vertex intrinsic fitness. *Phys Rev Lett*, 89(25).

[Cho et al., 1999] Cho, J., Shivakumar, N., and Garcia-Molina, H. (1999). Finding replicated web collections. In *ACM SIGMOD*, pages 355–366.

[Crovella and Bestavros, 1996] Crovella, M. E. and Bestavros, A. (1996). Self-similarity in world wide web traffic: evidence and possible causes. In *SIGMETRICS '96: Proceedings of the 1996 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, volume 24, pages 160–169, New York, NY, USA. ACM Press.

[da Silva et al., 1999] da Silva, A. S., Veloso, E. A., Golgher, P. B., , Laender, A. H. F., and Ziviani, N. (1999). Cobweb - a crawler for the brazilian web. In *Proceedings of String Processing and Information Retrieval (SPIRE)*, pages 184–191, Cancun, MÃ©xico. IEEE CS Press.

[Davison, 2000] Davison, B. D. (2000). Topical locality in the web. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*, pages 272–279, Athens, Greece. ACM Press.

[Dill et al., 2002] Dill, S., Kumar, R., Mccurley, K. S., Rajagopalan, S., Sivakumar, D., and Tomkins, A. (2002). Self-similarity in the web. *ACM Trans. Inter. Tech.*, 2(3):205–223.

[Efthimiadis and Castillo, 2004] Efthimiadis, E. and Castillo, C. (2004). Charting the Greek Web. In *Proceedings of the Conference of the American Society for Information Science and Technology (ASIST)*, Providence, Rhode Island, USA. American Society for Information Science and Technology.

[Fetterly et al., 2004] Fetterly, D., Manasse, M., and Najork, M. (2004). Spam, damn spam, and statistics: Using statistical analysis to locate spam Web pages. In *Proceedings of the seventh workshop on the Web and databases (WebDB)*, Paris, France.

[Fetterly et al., 2005] Fetterly, D., Manasse, M., and Najork, M. (2005). Detecting phrase-level duplication on the world wide web. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 170–177, New York, NY, USA. ACM Press.

[Gomes and Silva, 2003] Gomes, D. and Silva, M. J. (2003). A characterization of the portuguese web. In *Proceedings of 3rd ECDL Workshop on Web Archives*, Trondheim, Norway.

[Gulli and Signorini, 2005] Gulli, A. and Signorini, A. (2005). The indexable Web is more than 11.5 billion pages. In *Poster proceedings of the 14th international conference on World Wide Web*, pages 902–903, Chiba, Japan. ACM Press.

[Gyöngyi and Garcia-Molina, 2005] Gyöngyi, Z. and Garcia-Molina, H. (2005). Web spam taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web*.

[Kilgarriff and Grefenstette, 2004] Kilgarriff, A. and Grefenstette, G. (2004). Introduction to the special issue on the Web as corpus. *Computational Linguistics*, 29(3):333–348.

[Koster, 1996] Koster, M. (1996). A standard for robot exclusion. http://www.robotstxt.org/wc/exclusion.html.

[Lee et al., 1994] Lee, B. T., Masinter, L., and Mccahill, M. (1994). RFC 1738: Uniform resource locator (URL). http://www.ietf.org/rfc/rfc1738.txt.

[Marchiori, 1998] Marchiori, M. (1998). The limits of web metadata, and beyond. In *WWW7: Proceedings of the seventh international conference on World Wide Web 7*, pages 1–9, Amsterdam, The Netherlands, The Netherlands. Elsevier Science Publishers B. V.

[Mccallum, 1996] Mccallum, A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. <http://www.cs.cmu.edu/∼mccallum/bow/>.

[Menczer, 2004] Menczer, F. (2004). Lexical and semantic clustering by web links. *Journal of the American Society for Information Science and Technology*, 55(14):1261–1269.

[Modesto et al., 2005] Modesto, M., Pereira, á., Ziviani, N., Castillo, C., and Baeza-Yates, R. (2005). Um novo retrato da web brasileira. In *Proceedings of XXXII SEMISH*, pages 2005–2017, SÃ£o Leopoldo, Brazil.

[Nielsen, 2005] Nielsen, J. (2005). Alertbox: Jakob nielsen's column on web usability. <http://www.useit.com/alertbox/>.

[Page et al., 1998] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project.

[Pandurangan et al., 2002] Pandurangan, G., Raghavan, P., and Upfal, E. (2002). Using Pagerank to characterize Web structure. In *Proceedings of the 8th Annual International Computing and Combinatorics Conference (COCOON)*, volume 2387 of *Lecture Notes in Computer Science*, pages 330–390, Singapore. Springer.

[Pitkow, 1999] Pitkow, J. E. (1999). Summary of WWW characterizations. *World Wide Web*, 2(1-2):3–13.

[Rauber et al., 2002] Rauber, A., Aschenbrenner, A., Witvoet, O., Bruckner, R. M., and Kaiser, M. (2002). Uncovering information hidden in Web archives. *D-Lib Magazine*, 8(12).

[Sanguanpong et al., 2000] Sanguanpong, S., Nga, P. P., Keretho, S., Poovarawan, Y., and Warangrit, S. (2000). Measuring and analysis of the Thai World Wide Web. In *Proceeding of the Asia Pacific Advance Network conference*, pages 225–230, Beijing, China.

[Suel and Yuan, 2001] Suel, T. and Yuan, J. (2001). Compressing the graph structure of the Web. In *Proceedings of the Data Compression Conference DCC*, Snowbird, Utah, USA. IEEE CS Press.

[Thelwall, 2002] Thelwall, M. (2002). Conceptualizing documentation on the web: An evaluation of different heuristic-based models for counting links between university web sites. *Journal of the American Society for Information Science and Technology*, 53(12):995–1005.

[Thelwall, 2004] Thelwall, M. (2004). *Link Analysis: An Information Science Approach (Library and Information Science)*. Academic Press.

[Thelwall and Wilkinson, 2003] Thelwall, M. and Wilkinson, D. (2003). 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.

[Veloso et al., 2000] Veloso, E. A., de Moura, E., Golgher, P., da Silva, A., Almeida, R., Laender, A., Neto, R. B., and Ziviani, N. (2000). Um retrato da Web Brasileira. In *Proceedings of Simposio Brasileiro de Computacao*, Curitiba, Brasil.

[Zipf, 1949] Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley, Cambridge, MA, USA.