Characterizing a National Community Web

DANIEL GOMES and MÁRIO J. SILVA
University of Lisbon

This article presents a characterization of the community Web of the people of Portugal. We defined criteria for delimiting this Web based on our past experience of crawling pages related to Portugal and collected over 3.2 million documents from 46,000 sites satisfying those criteria. Our characterization was derived from this crawl. We describe the rules that we established for defining the boundaries of this community Web and the methodology used to gather statistics. Statistics cover the number and domain distribution of sites; the number, type and size distribution of text documents; and the linkage structure of this Web. We also show how crawling constraints and abnormal situations on the Web can influence the statistics.

Categories and Subject Descriptors: H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*Collection*; C.2.5 [**Computer-Communication Networks**]: Local and Wide-Area Networks—*Internet*

General Terms: Measurement, Experimentation, Documentation

Additional Key Words and Phrases: Web characterization, Portuguese Web, Web measurements, Web communities

1. INTRODUCTION

A characterization of the Web is of great importance. It reflects technological and sociological aspects and permits us to understand how the Web has evolved. An accurate characterization of the Web enables improvements in the design and performance of applications that use the Web as a source of information (e.g. crawlers, proxies, search engines) [Cho and Garcia-Molina 2000].

The Web can be characterized from multiple perspectives using numerous metrics. This is a challenging task, mainly because of its large dimension and continual evolution [Leung et al. 2001]. Producing a feasible general characterization is hard, and some statistics derived from the analysis of the global

This study was partially supported by the FCCN-Fundação para a Computação Científica Nacional, FCT-Fundação para a Ciência e Tecnologia, under Grants POSI/SRI/40193/2001 (XMLBase project) and SFRH/BD/11062/2002 (scholarship).
Authors' address: Faculdade de Ciências da Universidade de Lisboa, Departamento de Informática, Campo Grande, 1749-016 Lisboa, Portugal; email: {dcg,mjs}@di.fc.ul.pt.

Web may not hold as we scale down to more restricted domains. The Web has partitions with specific characteristics that, given their small presence, do not become visible in a general Web characterization. However, these partitions can be of interest to relatively large communities such as those representing national or cultural groups. Additionally, characterizing a small partition of the Web is quite accessible and can be done with great accuracy.

In this article, we present a detailed characterization of a national community Web. This work was conducted in the context of a study of the Portuguese Web, broadly defined as the set of pages of cultural and sociological interest to the people of Portugal. The results are derived from a crawl performed by tumba!, a search and archival engine for the Portuguese Web [Silva 2003]. We focused our study on textual content available on the Portuguese Web, identifying metrics that would help us in the design and improvement of our system. The statistics themselves are interesting to anyone who manipulates this data or will compare it with our snapshot in the future. We compare our results with related work. However, we need to be cautious about the conclusions drawn because the results were gathered during different periods and using distinct methodologies which often are not detailed enough. The identification of the meaningful statistics for a community Web characterization and the methods used to gather and interpret the collected data could be useful to a wider audience. We detail our crawling policy and show how the crawling and data analysis processes can strongly influence the obtained statistics.

This article is organized as follows. The remainder of this section presents the adopted terminology. Section 2 presents our heuristics for defining the boundaries of the Portuguese Web. In the following 2 sections, we present the crawler configuration and the crawling results. In Sections 5, 6, and 7, we describe the statistics derived from the crawl related to Web sites, documents, and structure, respectively. Section 8 introduces related work. Finally, in Section 9, we draw our conclusions and present directions for future research.

1.1 Terminology

The concepts used in this study were adapted from the terminology proposed by the W3C [1999].

—Publisher: entity responsible for publishing information on the Web;

—Document: file resulting from a successful HTTP download;

—Page: HTML document;

—Web site: collection of documents referenced by URLs that share the same host name (a discussion about the definition of Web site can be found in O'Neill [1999]).

—Host page: document identified by a URL where the file path component is empty or a '/' only.

—Subsite: cluster of documents within a Web site maintained by a different publisher than that of the parent site.

—Host aliases: sites that have different names but are hosted on the same IP address and have the same host page.

2. IDENTIFYING THE BOUNDARIES OF A COMMUNITY WEB

The Web is designed to break all the geographical barriers and make information universally available. However, as the Web is the product of multiple user groups, it is possible to identify partitions within it containing the sites of interest to these groups. These are designated as community Webs and can be defined as the set of documents that refer to a certain subject or are of interest to a community of users. Detection of a community Web is not always obvious, despite various existing methods that can be used to identify its sites.

If we are interested in a small and static set of documents, then enumerating all the documents that compose the community Web can be adequate. However, it becomes very expensive to maintain the list of documents if it grows or changes frequently [Webb 2000].

We can also use the link structure [Flake et al. 2000] of the Web, but we'll have difficulties identifying documents loosely interlinked even if they refer to the same subject. For instance, the sites of several concurrent companies in the same business will not likely link to each other.

We can identify documents related to a country through a common country code Top Level Domain (ccTLD) [Postel 1994; Zabicka 2003]. In this case, we could exclude all the documents related to that country hosted under a domain outside the ccTLD. On the other hand, this rule could also include sites not related to the country but hosted under its ccTLD. For instance, multinational companies commonly register their name under many domains to protect their brands.

The language in which the documents are written is a good indicator of which country they are related to [Albertsen 2003]. However, problems arise if the language is not exclusive to a single country: we couldn't include all the documents written in English within a British community Web.

As a result, a precise definition of which documents should constitute a community Web is, in general, hard to obtain and is conditioned by the rules and resources used.

The community Web of our study is the Portuguese Web. We define it as the set of documents containing information related to Portugal or of major interest to the Portuguese people. Our first approach for establishing the boundary of the Portuguese Web outside the .PT domain was to harvest all the documents written in the Portuguese language. Soon, we found that this would require the downloading of a large number of documents, especially from Brazil, containing information not highly related to Portugal. As a defining rule, we consider as part of the Portuguese Web those documents that satisfy one of the following conditions:

(1) Hosted on a site under the .PT domain;
(2) Hosted on a site under the .COM, .NET, .ORG or .TV domains, written in the Portuguese language, and with at least one incoming link originating in a Web page hosted under the .PT domain.

This definition aims to be easily set as a crawling policy and guarantees that the crawler gets the best coverage of the Portuguese Web.

Condition (1) intends to include the sites that constitute the core of the Portuguese Web. A list of the most popular sites, accessed from the homes of a panel of Portuguese users during 2002 and 2003 [Marktest 2003], showed that 49.5% of the sites were hosted under the .PT domain. Based on this information, we considered the sites hosted under the .PT domain as the core of the Portuguese community Web.

Condition (2) intends to include the increasing number of Portuguese sites that are registered outside the .PT domain [Zook 2000]. Previous work [Flake et al. 2000; Gibson et al. 1998] showed that the link structure of the Web can be used to define communities. We assumed that the probability of a site hosted outside the .PT domain, belonging to the Portuguese Web community, decreases as the number of hops in the Web graph to the core increases. So, in order to restrict the inclusion of sites outside the .PT domain to the ones with the highest probability of being part of the Portuguese Web, we limited the number of hops to 1. Condition (2) imposes the rule that only documents directly linked from a site hosted under the .PT domain are part of the Portuguese community Web. However, Brazilian documents linked from the .PT domain and hosted under the allowed domains, such as .COM, will still be considered as part of the Portuguese Web.

The definition of the Portuguese community Web has an implicit geographical context. We ran an experiment with the purpose of comparing the coverage of the Portuguese community Web outside the .PT domain (Condition (2)) by different alternatives.

We gathered a list of 25 Portuguese sites hosted outside the .PT domain suggested by Portuguese users. Then we examined the suggested sites and verified that they were directly related to the Portuguese Web community. All the sites were written in Portuguese and referred to distinct subjects such as sports, humor, or radio. We compared our proposed defining criterion against 2 alternatives based on tools that provide geographic context data for Web sites.

In the first alternative, we used 2 commercial tools, Ip2location [Center 2003] and Maxmind [LLC 2003], and extracted a geographical location for each site on the list. We considered a site as part of the Portuguese Web if the tool returned that the site was located in Portugal. For 2 submissions of the same site, Maxmind returned different results. Except for this situation, both tools presented the same results which led us to believe that they are based on the same data.

Our second alternative was to access a whois database [Harrenstien et al. 1985] to identify the Portuguese sites hosted outside the .PT domain. For each site, we obtained the contact address of the correspondent domain registrant. If this address was located in Portugal, we considered the site as part of the Portuguese Web.

Finally, following our proposed definition, we checked if the sites had at least one link from a site hosted under the .PT domain. We used the search engines Google [Google 2003] and AllTheWeb [Overture Services 2003] to identify pages that link to the sites. We also tried to obtain geographic information through the DNS LOC record [Davis et al. 1996] but none of the domains had an associated record of this kind.

Table I. Comparison Between Alternative Definitions of the Portuguese Web

Definition	% Sites Identified	% Information Unavailable
Geographical tools	44	0
Whois registrant address	76	24
Linked from .PT	82	12

Table I presents the results obtained through the three definitions of the Portuguese Web. The geographical tools identified only 44% of the Portuguese sites, although they always returned an answer to the location requests. 76% of the Portuguese sites were identified through the registrant information. For 24%, of the sites, the whois database didn't contain the information regarding the requested domain. For some of these cases, we found the registrant information on another whois server. As there isn't a central whois database, the registrants information is distributed over the several registrars which causes inconsistencies among whois databases. The registrant address proved to be a precise method of identifying Portuguese sites outside the .PT domain. All the sites in our list which had a whois record available were correctly identified. Therefore, the whois databases could be the solution to the problem of distinguishing Brazilian sites from Portuguese sites outside the .PT domain. However, most of the whois databases are not publicly available or explicitly forbid their access by automated programs which conflicts with our purpose of having a definition of the Portuguese Web that can be implemented as a crawling policy. The existence of several record formats also makes it difficult to automatically process whois records. Additionally, companies that provide hosting services support several distinct sites identified by subdomains and the whois registries only keep information about second-level domains. This does not include the subdomains of hosted Web sites. This is a serious restriction if we consider, for instance, all the Portuguese blogs hosted under `blogspot.com`. If we had followed this approach, many Portuguese Web sites hosted outside Portugal would not be considered as part of the Portuguese Web.

We observed that 82% of the suggested sites would be included in the Portuguese Web using our criteria: they were written in the Portuguese language and had at least one link from a site hosted under a .PT domain. The results show that our proposed definition of the Portuguese Web provides the best coverage of the suggested sites.

3. CRAWLER CONFIGURATION

A crawler begins its task of harvesting the Web collecting the documents referenced by an initial set of URLs called the seeds. Then it iteratively extracts links to new URLs and collects their contents.

Crawlers are configured or developed according to the purpose of the data they gather. A crawler of a large scale search engine aims to collect pages with the highest PageRank [Cho et al. 1998; Brin and Page 1998]. On the other hand, archive crawlers focus on crawling the most pages on a given partition [Day 2003]. In our study, we configured Vúva Negra (VN) [Gomes 2003], the Web crawler of the tumba! search engine, to get the most information possible about the Portuguese Web. We initialized it with a set of 112,146 seeds gathered from

previous crawls and user registrations that included all the hosts registered under the .PT domain. We imposed on it the minimum constraints that ensure an acceptable performance of the crawler, considering the resources available and the need to make it robust against the usual anomalies in the traversed Web graph [Henzinger 2003]. A document was considered to be valid if it was part of the Portuguese Web as defined in the previous section. In addition, the following crawler conditions had to be met.

—*Multiple text types.* We considered not only documents of the text MIME type, but also documents of common MIME application types that we could convert to text. Accepted MIME types are text/html, text/richtext, text/tab-separated-values, text/plain, text/rtf, application/pdf, application/rtf, application/x-shockwave-flash, application/x-tex, application/msword, application/vnd.ms-excel, application/excel, application/mspowerpoint, application/powerpoint and application/vnd.ms-powerpoint.

—*URL depths less than* 6. The crawler followed at most 5 links in breadth-first search order from the seed of the site until it reached the referenced document. When crawling a site, we considered that any link found to a different site would be set as a seed to that site. This way, we guaranteed that any page with a link originated on a .PT domain would be visited, including Portuguese subsites hosted on foreign sites. Consider, for instance, the site `www.yahoo.com` and its Portuguese subsite `www.yahoo.com/users/myPortugueseSite/`. If the crawler had visited only the seed `www.yahoo.com`, it would had determined that the site was not part of the Portuguese Web and exited without finding the Portuguese subsite.

—*Documents downloaded in less than* 1 *minute.* This prevents very slow Web servers from blocking the progress of the crawl.

—*Document size under* 2*MB.* This prevents the download of very large files available on the Web such as database dumps.

3.1 Avoiding Traps

A crawler trap is a set of URLs that cause a crawler to traverse a site indefinitely. They are easily noticed due to the large number of documents discovered in the site [Heydon and Najork 1999]. In order to prevent the crawling of infinite sites, we set VN to visit a maximum of 8000 URLs per site. This turned out to be an acceptable limit, considering the dimensions of the Portuguese Web sites (see Section 5). This constraint reduced the number of unnecessary downloads and increased the robustness of the crawler but it wasn't enough to prevent traps from biasing a Web characterization. We found that most of the traps are unintentional, caused mainly by session identifiers embedded in the URLs or poorly designed HTTP Web applications that dynamically generate an infinite number of URLs that reference a small set of documents.

This raises the issue of how these documents should be considered in a characterization. They should not be excluded because they are available online and represent part of the Web. However, we cannot let them bias a characterization due to their infinite presence. We adopted the solution of setting VN as a very

patient Web surfer as a compromise. After seeing the same bitwise identical document 50 times, VN gives up on following links for that site, keeping all the information crawled up to that point. This limitation intends to avoid spider traps that always return the exact same content. If the trap generates slightly different contents, we identify it when the site reaches the maximum number of documents allowed. A criterion that identifies documents with distinct URLs and contents as being similar enough to be considered the same is highly subjective. If several documents are similar except in a banner ad that changes on every download, they could reasonably be considered the same. However, when the difference between them is only as short as the licitation value on an online auction, the small difference could be very significant. As a result, we considered the computation of partial similarity between documents to be too expensive and risky to be applied in the identification of spider traps during the crawling process.

We didn't identify any intentional trap in our crawl. However, they can be created (using DNS wildcarding [Barr 1996]) to resolve any possible host name within a domain to the same IP address, generating an infinite number of host aliases and giving Web crawlers the illusion that each site serves only a small number of pages. In order to mitigate this situation, VN avoids crawling on host aliases by identifying them through a precomputed list gathered from a previous crawl. We verified the host aliases list using online information (IP address and host pages) before starting the crawl so that it could be as accurate as possible. Hence, there are host aliases that have disappear from a previous crawl.

4. CRAWLING STATISTICS

The results presented were extracted from a crawl performed between the 1st of April and the 15th of May, 2003. VN visited a total of 146,076 sites, processed over 3.8 million URLs, and downloaded 78GB of data.[1]

Table II presents the statistics of the download status of crawled URLs. Almost 84% of the requests resulted in a successful download, and only 3.4% resulted in a 404 (File Not Found) response code which indicates that most of our seeds were valid and that broken links are not as frequent in this Web as reported in other studies [Najork and Heydon 2001; Spinellis 2003]. There were over 6% of redirections, and the crawler failed to fetch and parse a document within 1 minute in 1.2% of the requests. The Robots Exclusion Protocol prevented VN from downloading 0.9% of the URLs, and about the same number of URLs resulted in an Internal Server Error (500). The number of documents with a not allowed MIME type (0.7%) is underestimated because extracted links that had names hinting that the referenced content didn't belong to one of the allowed types (e.g., files with a .JPEG extension) were not crawled. The UnknownHost error (0.5%) is caused by URLs referencing host names that no longer have an associated IP. We found that only 0.5% of the referenced files had a size bigger than 2MB, and the conversion to text was not possible in

[1]The information gathered in this crawl is available for research purposes at `http://xldb.fc.ul.pt/linguateca/WPT_03.html`.

Table II. Summary of the Status Codes Associated to the URLs Visited (The positive numbers represent the HTTP response codes, and the negative numbers represent VN special codes that identify the reason why the contents referenced by the URLs were not collected.)

State	# URLs	%
200	3,235,140	83.9
302	193,870	5.0
404	132,834	3.4
TimedOut (-8)	45,486	1.2
301	39,920	1.0
ExcludedByREP (-2)	35,596	0.9
500	33,247	0.9
NotAllowedType (-5)	25,976	0.7
403	18,598	0.5
UnknownHost (-14)	17,842	0.5
SizeTooBig (-4)	17,453	0.5
ConversionError (-11)	13,986	0.4
Other	23,244	0.6
Total	3,856,436	100.0

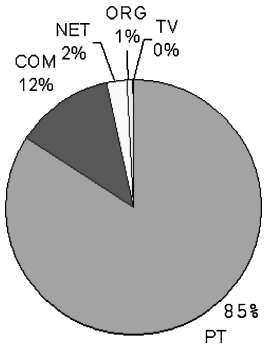

Fig. 1. Distribution of sites per top level domain.

0.4% of the cases. The remaining situations (0.6%) included other HTTP response codes, unidentified errors, socket and connection errors; each of these represents less than 0.1% of the total number of downloaded documents.

5. SITE STATISTICS

We considered that a site is part of the Portuguese Web if it hosted at least one document considered as part of the Portuguese Web. We identified 46,457 sites as being part of the Portuguese Web. 85% of the sites were under the .PT domain, 12% were under the .COM, 1% were under the .ORG domain, and just 3 sites were under the .TV (see Figure 1). 60% of the Web site names started with WWW. A Portuguese Web site has an average of 70 documents but the size distribution is very skewed, as shown in Figure 2. We were surprised by the high number (38%) of sites having a single document. We visited a random sample of these sites and observed that most warned readers that they were under construction or that the site moved to a different location. We also found a few cases where the host page was completely written using scripting

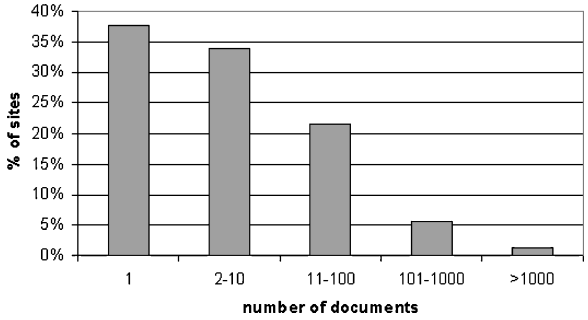

Fig. 2. Distribution of documents per site.

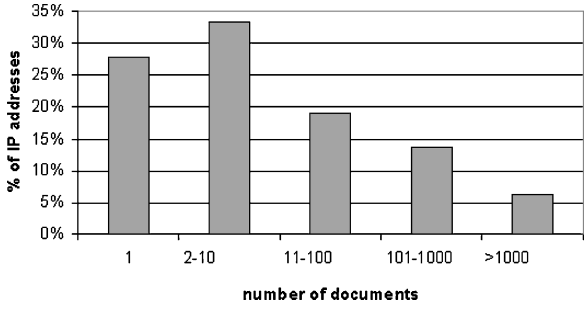

Fig. 3. Distribution of documents per IP address.

languages from which our parser couldn't extract links. A typical Web site had less than 101 documents (93%); 6% had between 101 and 1,000 documents, and only 1% of the sites had more than 1,000 documents. We identified just 577 sites that hosted more than the maximum number of 8,000 documents. We observed that most of these sites were huge database dumps available online through dynamically-generated Web pages. We concluded that, despite the restriction on the site size, we were able to exhaustively crawl 99% of the sites.

The distribution of documents per IP address is more uniform (see Figure 3). The percentage of IP addresses that host just one document is 28%. IP addresses that host 2 to 10 documents represent 33%, and those which host between 11 and 1,000 documents represent 33%. Only 6% host more than 1,000 documents.

Table III shows that over 32% of the IP addresses host more than 1 site. Each IP address hosts an average of 6.78 sites. Silva et al. [2002b] compared results from 2 crawls of the .PT domain performed in 2001 and 2002 and observed that the number of sites per IP address grew from an average of 3.78 to 4.57 sites per IP address. Our result suggests that this number continues to grow. There are 5 IP addresses that host more than 1,000 sites. These 5 IP addresses are from Web portals that offer their clients a virtual host under the portal domain, providing a proper host name for their site instead of having it as a subsite. Virtual hosts are very popular on the Portuguese Web: 82% of all sites are virtual hosts. It is important to distinguish host aliases from distinct virtual hosts. The first occur when multiple names refer to the same site, for instance, `http://xldb.fc.ul.pt` and `http://xldb.di.fc.ul.pt`. Distinct virtual hosts

Table III. Distribution of the Number of Sites Hosted per IP Address

# Sites per IP	# IP Addresses	%
1	4,643	67.7
2–10	1,931	28.2
11–100	247	3.6
101–1000	30	0.4
>1000	5	0.1
Total	6,856	100.0

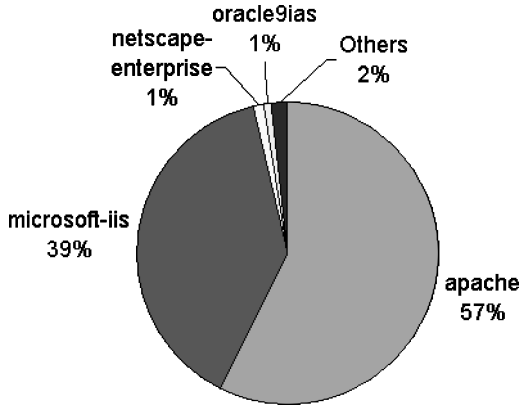

Fig. 4. Distribution of Web servers.

are distinct sites hosted on the same machine such as `http://xldb.di.fc.ul.pt` and `http://lasige.di.fc.ul.pt`. In our crawl, we found out that 8.5% of the virtual hosts were host aliases.

5.1 Web Servers

We identified 172 distinct HTTP Web servers. Figure 4 presents their distribution. The Portuguese sites are mainly hosted at Apache (57%) and Microsoft IIS (39%) Web servers. The next two web servers (netscape-enterprise and oracle9ias) represent just 1% each, and the remaining just 2%. Statistic on the global Web present a similar percentage of Apache web servers (62.57%) but a considerably smaller percentage of Microsoft IIS servers (27.45%) [Netcraft Ltd. 2004]. On the other hand, our distribution of Web server software contrasts with the one obtained by Boldi et al. [2002] for the Africa Web in which there is a dominance of Microsoft IIS over Apache (56.1% against 37.95%).

Security experts encourage webmasters not to provide the Server HTTP field or to provide wrong answers in order to mislead possible attackers. From our experience, we believe that these recommendations are usually not followed by the Portuguese webmasters. However, if they become popular, it will be very difficult to correctly identify the distribution of Web server software.

6. DOCUMENT STATISTICS

In this section, we present metrics regarding the length of URLs, MIME types, size, language, and metadata of documents.

Fig. 5. Distribution of URL lengths.

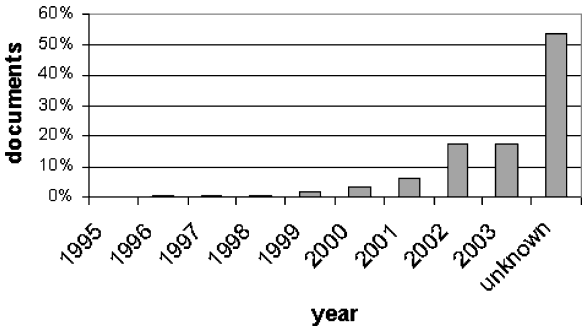

Fig. 6. Distribution of last-modified dates.

6.1 URLs

Every Web application must have some kind of data structure that maps into URLs. However, we didn't find in the literature a study discussing the lengths of the URLs. Today the size of URLs is, in practice, unlimited. We found valid URLs with lengths varying from 5 to 1,368 characters. Figure 5 shows the distribution of URL lengths (not considering the initial 7 characters of the protocol) over the number of documents. Most of the documents have a URL length between 20 and 100 characters, with an average value of 62 and a median value of 54. Analyzing the URLs, we found that 47.2% contained parameters suggesting that the referenced document had been dynamically generated.

6.2 Last Modified Date

HTTP provides a header field (Last-Modified Date) that should indicate the date of the last modification of documents. However, as shown in Figure 6, most of the documents (53.5%) returned an unknown value for this field. Plus, Mogul showed that even the returned values are very inaccurate due to incorrectly set Web server clocks (among other problems) [Mogul 1999b]. An analysis of the URLs with unknown values revealed that 82% of them had embedded parameters. We speculate that most of them are recent and would significantly increase the percentage of documents modified in the last months (see Figure 7)

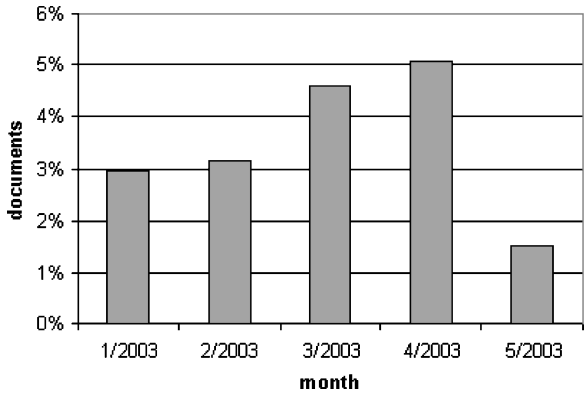

Fig. 7. Distribution of last-modified dates in the last 4 months.

Table IV. Number of Documents and Relative Presence on the Web for Each MIME Type Collected

MIME	# Documents	%
text/html	3,104,771	95.9
application/pdf	62,141	1.9
text/plain	33,091	1.0
application/x-shockwave-flash	17,598	0.5
application/msword	14,014	0.4
powerpoint	2,085	0.1
excel	915	0.0
application/x-tex	222	0.0
text/rtf	194	0.0
application/rtf	66	0.0
text/tab-separated-values	41	0.0
text/richtext	2	0.0
Total	3,235,140	100.0

since mechanisms to dynamically generate documents are usually used to reference short life contents such as news.

We believe that the last-modified header is a weak metric for evaluating changes and evolution of contents on the Web so metrics like these are meaningful only in the context of analysis of consecutive crawls [Wills and Mikhailov 1999; Fetterly et al. 2003].

6.3 MIMEs & Sizes

The rightmost column of Table IV shows the distribution of documents per MIME type, (we grouped all the MIME types used for Microsoft Powerpoint files under the name *powerpoint* and all the Microsoft Excel files under the name *excel*). The pre-dominant text format is text/html, present in over 95% of the collected documents, followed by application/pdf with just 1.9%.

In our first approach to determine the size of the documents, we analyzed the values of the HTTP header field Content-Length, but we noticed that 33% of the documents returned an unknown value. We then recomputed our results replacing the unknown sizes by the sizes of the documents. The differences on the average sizes between the results were insignificant except for text/html

Table V. Average Size, Extracted Text Size, Percentage of Extracted Text

MIME	Avg Doc Size(KB)	Avg Text Size(KB)	% Text
powerpoint	1054.9	7.0	1
text/rtf	475.6	1.2	0
application/pdf	207.4	13.6	7
application/rtf	121.3	4.7	4
application/msword	118.6	9.9	8
excel	50.4	21.9	43
application/x-shockwave-flash	43.9	0.3	1
text/html	20.5	2.5	12
text/richtext	16.3	16.2	99
application/x-tex	16.1	14.7	91
text/plain	10.5	7.8	74
text/tab-separated-values	3.9	3.8	97

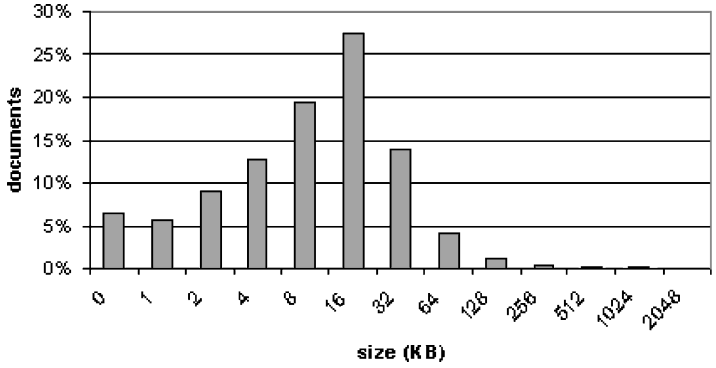

Fig. 8. Distribution of document sizes.

where the size grew from 12.2KB to 20.5KB. In Table V, the second and third columns show the average sizes of documents and corresponding extracted texts (without any formatting tags), and the fourth column presents the ratio between the length of the extracted text and document size. We can see that the size of the documents is almost inversely proportional to the size of the texts extracted. A curious fact is how documents of text/plain result in just 74% of text. We analyzed some of these documents and discovered that some Web servers return text/plain when the file type of the document is not recognized. Therefore, some PowerPoint Presentation files (.PPS) or Java Archives (.JAR) were incorrectly processed as text/plain, resulting in poor extraction of text from these files.

Figure 8 shows the general distribution of document sizes. Most documents have between 4 and 64KB. The mean size of a document is 32.4KB, and the mean size of the extracted texts is 2.8KB. The total size of the documents was 78.430GB, while the total size of extracted texts was just 8.791GB.

6.4 Language Distribution

Our crawler can identify the language of collected documents based on an idiom detector that implements an n-gram algorithm [Cavnar and Trenkle 1994].

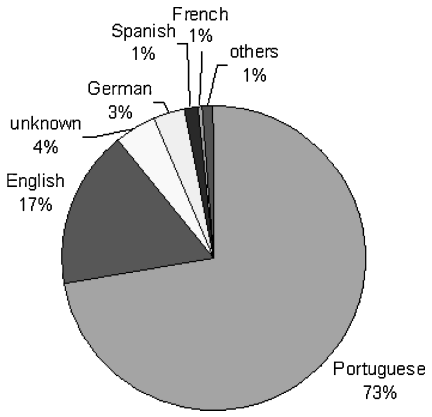

Fig. 9. Distribution of languages.

Figure 9 shows the distribution of languages on the documents of the Portuguese Web (including documents written in all languages hosted under the .PT domain): 73% of the documents were written in Portuguese, 17% in English, 3% in German, 1% in Spanish, 1% in French and 1% in other languages. According to O'Neill et al. [2003], on the global Web, 72% of the pages are written in English and only 2% are written in Portuguese.

Identifying the language of a document is sometimes a hard task because there are documents with short text or text written in several languages. The idiom detector couldn't identify the language of the document in 4% of the documents.

6.5 Metatags

We studied the usage of two important metatags supported by HTML: *description* and *keywords* [W3C 1999]. The description metatag provides a description of the page's content and the metatag keywords provides a set of keywords that search engines may present as a result of a search. We found that just 17% of the pages had the metatag description and that, among these, the use of this metatag didn't seem to be correct. We found only 44,000 distinct values for 555,000 description metatags. This means that 92% of the texts of the descriptions were repeated elsewhere. We identified a set of causes for this situation:

—the meta tag value is a default text inserted by a publishing tool;
—the publisher repeated the same text in all the pages of its site, although they are different;
—there are replicated pages on the Web.

The keywords metatag is present in 18% of the pages. A deeper analysis revealed that 91% of the pages that have the description metatag also had the keywords metatag. O'Neill et al. [2003] showed that the usage of metatags on the global Web has being increasing in the past years. In 2002, 70% of the pages included metadata [O'Neill et al. 2003]. Although our results focus on only two

Table VI. Distribution of Content with Replicas

Number of Replicas	Number of Contents	% of Contents
0	2,462,490	90.0
1	205,882	7.5
2	33,468	1.2
3	12,814	0.5
4	6,086	0.2
5	5,272	0.2
6–10	6,453	0.2
11–100	2,318	0.1
101–1000	154	0.0
>1000	5	0.0
Total	2,734,942	100.0

of the most popular metatags, we believe that the usage of metatags on the Portuguese Web is much less frequent than on the global Web.

The titles of the Web pages aren't very descriptive either. There were over 600,000 distinct titles for 3.1 million pages. The main reason we found for this observation is that the title of the site's host page is used as the title for all the pages in the site in most cases.

7. WEB STRUCTURE

7.1 Content Replication

To detect content replication, we compared the MD5 digest [Rivest 1992] obtained for each document. We identified 273,4942 distinct pieces of content. Table VI presents the replication distribution. We found that 15.5% of the content was referenced by several distinct URLs (replicas). Mogul identified only 5% of replicas when analyzing a client trace from WebTV [Kelly and Mogul 2002]. We believe that the difference between the two results is due to the distinct methodologies adopted in the respective analyses. A crawl-based approach analyzes all the documents available on the Web, while a client trace permits analyzes of only the content accessed by users. Most of the content (90%) is unique and 7.5% had exactly one replica. Content replicated more than 1,000 times is very rare. However, it was the cause of 13,146 downloads for just 5 distinct pieces of content. These situations are pathological for Web crawlers and also tend to bias the collection statistics. We observed these 5 cases and concluded that they were all caused by malfunctioning Web servers which always return the same error page for all the requests. Our measures against these traps (see Section 3.1) failed because all the links to documents with error messages were correctly extracted. When the crawler finally identified the trap, it already had numerous URLs to crawl even though it had stopped inserting new links.

Our measurements indicate that 42% of the replicas are duplicates of content hosted on the same site; 60% are duplicates of content hosted on a different site; 2% are duplicates of content hosted both in the same site and in another site.

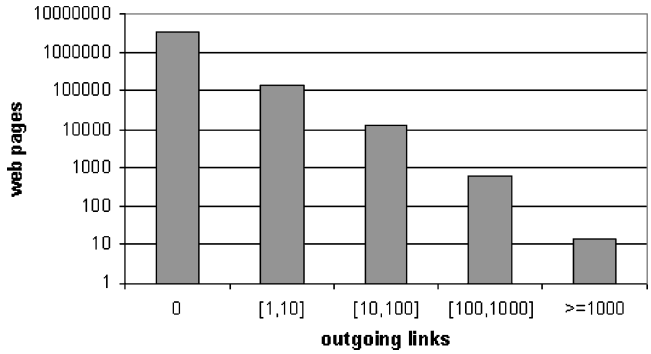

Fig. 10. Distribution of the number of outgoing links per Web page.

Table VII. The 10 documents with the Highest Number of Outgoing Links

	# links	URL
1	3,540	cpan.dei.uc.pt/modules/00modlist.long.html
2	2,425	ftp.ist.utl.pt/pub/rfc/
3	2,309	homepage.oninet.pt/095mad/bookmarks_on_mypage.html
4	1,632	www.fis.uc.pt/bbsoft/bbhtm/mnusbib3.htm
5	1,621	cpan.dei.uc.pt/authors/00whois.html
6	1,532	www.fba.ul.pt/links4.html
7	1,458	boa.oa.pt/bbsoft2/bbhtm/mnusbib3.htm
8	1,346	www.esec-canecas.rcts.pt/Educacao/Escolas.htm
9	1,282	pisco.cii.fc.ul.pt/nobre/hyt/bookmarks.html
10	1,181	www.fpce.uc.pt/pessoais/rpaixao/9.htm

7.2 Link Structure

The link structure of the Web can be represented as a graph. Nodes represent URLs and edges represent hypertext links. We restricted our analysis to links between distinct sites originated on Web pages hosted under the .PT domain and targeted to one of the accepted TLDs. The links to URLs that evidenced that the referenced content was not of one of the accepted types were excluded (e.g., URLs where the file part has a .jpg extension were not collected).

We observed that most of the Web pages (95%) didn't link to another Portuguese site (Figure 10). On average a Web page has 0.23 links to documents on another site. This is not a surprising result since links are usually internal to the site. However, we also found pages rich in outgoing links (Table VII).

We found that 66% of the links didn't point to a document on the Portuguese Web. This measure was made under the assumption that all the URLs hosted outside the .PT domain for which we couldn't determine the language were considered as outside the Portuguese Web. 40% of the links pointed to the host page of a site. We found that 3,189,710 documents (89%) were not referenced by a link originated in another Portuguese site. As observed on the global Web, here most links tend to point to a small set of pages (Figure 11).

7.2.1 *Document Importance.* The importance of a document can be determined through the analysis of the Web graph. In order to achieve a meaningful

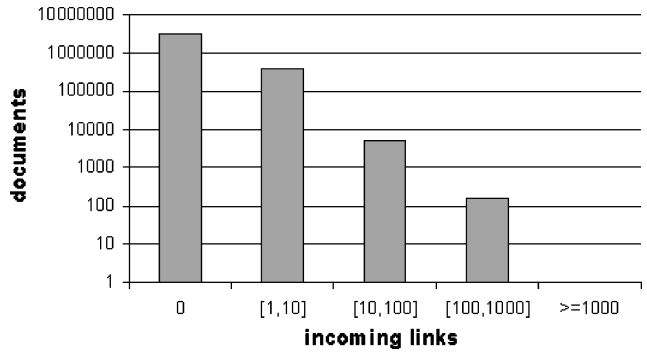

Fig. 11. Distribution of the number of incoming links per document.

Table VIII. The 10 Documents with the Highest Number of Incoming Links

	# Incoming links	URL
1	6,862	www.fccn.pt/
2	754	clanhosted.clix.pt
3	688	www.sapo.pt
4	606	www.publico.pt
5	522	www.infocid.pt
6	448	paginasbrancas.pt
7	423	www.dn.pt
8	413	www.sapo.pt/
9	361	security.vianetworks.pt
10	350	www.uminho.pt

ranking of the relative importance of documents, we handled links to replicas and HTTP redirects differently.

—Links to replicas cause the splitting of the number of links to a document among the several URLs that refer it. In the presence of replicated content, we elected the document with the smallest URL as the common reference. We then erased the replicated pages from the Web graph and retargeted the links to the replicas to the URL used as common reference.

—HTTP redirects are almost invisible to Web surfers. Involuntarily, publishers link to the URL of the redirect instead of the URL of the document. This causes a split in the number of links between the redirects and the document. We followed each redirect until we found a non-redirect URL. Then we replaced the redirect nodes in the graph by the correspondent non-redirect URLs.

Table VIII presents the 10 documents that received most incomings links on the Web graph obtained after the previously modifications to handle replicas and redirects were applied.

Despite our efforts to eliminate pathological situations in the analyzed graph, we still observe some anomalies on the most ranked document lists. In positions 3 and 8 of Table VIII, the number of incoming links was spread between 2

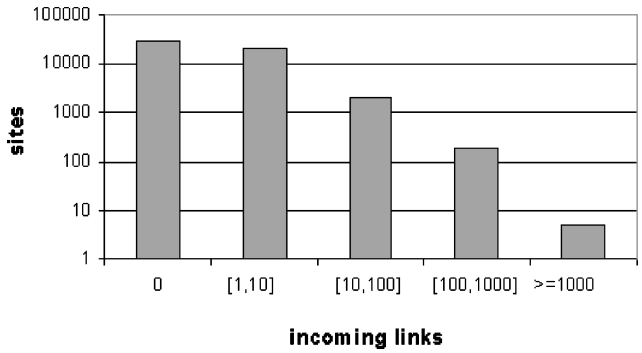

Fig. 12. Distribution of the number of incoming links per site.

different URLs although we know that they both refer to the same document. The problem was that we identified the two URLs by their string representation and, between the crawl of the first and the second URL, the content referenced by them changed. Sometimes the change on the content is very small. In our second example, the change was just a link to an advertisement.

7.2.2 *Site Importance.* The importance of a site can be derived from the total number of incoming links. A highly-ranked site might not host highly-ranked documents. For instance, some online newspapers receive a large number of incoming links to many distinct news pages but as news is interesting and is in many cases, available for only a short period of time, they never get to be highly-ranked documents.

Broder et al. [2000] analyzed the graph structure of the Web through 2 large crawls of 200 million pages each. They considered each page as a node and each hypertext link as an edge on the graph. They found that 91% of the pages were reachable from one another by following either forward or backward links after computing an algorithm that finds weak components in the graph. Our study followed a different methodology.

We considered only the links between distinct sites and didn't detect weak components in the graph. We generated a graph where each Portuguese site is a node and each link between documents on two different sites an edge.

We analyzed the graph as being undirected by following links in both directions and found that 73% of the sites connected to another site. This result contrasts with the one obtained by Broder et al. [2000] (91%) and shows that the connectivity of the graph decreased on a smaller partition of the Web such as the Portuguese Web.

Then we analyzed the graph as being directed, following links only in their real direction. We found that only 45% of the sites were reachable from one another site which leaves us with a majority of sites (55%) that are never linked (orphan sites).

Figure 12 presents the distribution of the number of incoming links per site. Table IX presents the 10 Portuguese sites that received most incoming links.

We obtained a list of 495 selected sites, accessed from the homes of a panel of Portuguese users, during the period we performed the crawl [Marktest 2003].

Table IX. The 10 Sites that Received Most Incoming Links

	# Incoming links	Site
1	7,109	www.fccn.pt
2	1,881	br.weather.com
3	1,617	images.clix.pt
4	1,601	www.sapo.pt
5	1,481	www.clinicaviva.pt
6	7,94	www.depp.msst.gov.pt
7	777	www.infocid.pt
8	721	www.fct.mct.pt
9	652	ultimahora.publico.pt
10	615	www.miau.pt

Table X presents the 40 sites that received the most distinct users. We can observe that the majority (27) of these popular sites are hosted under the .PT domain as we assumed. We noticed a high number of accesses to sites that are automatically accessed by tools. For instance, when a user types a URL of a site that is not found, Internet Explorer automatically redirects his request to auto.search.msn.com by default. These sites appear as overrated in usage statistics. We noticed that 50% of the sites accessed were part of the Portuguese Web. We found a correlation of 0.527 between the number of users and links to the Portuguese sites. This shows that the most linked sites are also most often visited by the users of this community Web.

At first sight, it's surprising that the site www.fccn.pt, which occupies the first position in Table IX, is not present in the list of the 495 sites accessed by the users. A deeper analysis revealed that 96% of the links to the FCCN (National Foundation for Scientific Computing) site were originated on sites hosted under the .RCTS.PT domain and almost all of them (99%) pointed to the host page (www.fccn.pt/). The RCTS network (Network for Science, Technology and Society) is also managed by FCCN. It is composed of over 11,000 sites from several public institutions and specially schools, hosted under the .RCTS.PT domain. We found that FCCN automatically generated a site on the RCTS network for every school in the country, initially composed by a single Web page containing its address, email, and a link to www.fccn.pt/. The content of these sites was supposed to be replaced by content produced by the school but, in most cases, this didn't happen. As a result, the default site prevailed, generating a high number of links to the FCCN site from other sites.

8. RELATED WORK

Web characterization has been done from different perspectives through the years almost since the beginning of the Web [Pitkow 1998]. The Web Characterization Project has been a great contributor to research in Web characterization [OCLC 2003; O'Neill et al. 2003].

Najork and Heydon [2001] performed a large scale Web crawl from which they gathered statistics regarding the outcome of download attempts, distribution of types, size of the documents, and replication. They found that the distribution

Table X. The 40 Most Accessed Sites by Portuguese Users. (courtesy 2003 Marktest Lda)

	# Users	Site
1	779,000	www.sapo.pt
2	580,000	www.microsoft.com
3	560,000	pesquisa.sapo.pt
4	548,000	loginnet.passport.com
5	540,000	www.clix.pt
6	538,000	www.google.pt
7	480,000	www.geocities.com
8	477,000	login.passport.net
9	471,000	www.terravista.pt
10	463,000	www.iol.pt
11	408,000	windowsupdate.microsoft.com
12	405,000	v4.windowsupdate.microsoft.com
13	380,000	www.msn.com
14	311,000	pesquisa.clix.pt
15	290,000	ww2.hpg.ig.com.br
16	247,000	Webmail.iol.pt
17	244,000	www.mytmn.pt
18	227,000	Webmail.sapo.pt
19	224,000	www.aeiou.pt
20	223,000	www.google.com
21	219,000	www.cidadebcp.pt
22	215,000	planeta.clix.pt
23	203,000	www.yahoo.com
24	202,000	Webmail.clix.pt
25	193,000	caixadirecta.cgd.pt
26	191,000	netcabo.sapo.pt
27	189,000	dossieriraque.clix.pt
28	185,000	tsf.sapo.pt
29	185,000	www.cgd.pt
30	182,000	login.passport.com
31	181,000	www.msn.com.br
32	180,000	bandalarga.netcabo.pt
33	177,000	www.dgci.gov.pt
34	173,000	www.abola.pt
35	171,000	auth.clix.pt
36	165,000	pwp.netcabo.pt
37	159,000	www.tvi.iol.pt
38	156,000	netbi.sapo.pt
39	156,000	www.record.pt
40	156,000	download.com.com

of pages over Web servers follows a Zipfian distribution. Lawrence and Giles [1999] studied the accessibility of information on the Web and drew conclusions about the size, extracted text, and usage of metadata in HTML pages.

Boldi et al. [2002] studied the structural properties of the African Web, analyzing HTTP header fields and content of HTML pages, and Punpiti et al. [2000] presented quantitative ,measurements and analyses of documents hosted under the .th domain.

Replication on the Web has been studied in several works through the syntactically clustering of documents [Broder et al. 1997], the study of the existence of near-replicas on the Web [Shivakumar and Garcia-Molina 1998], and different

levels of duplication between hosts and mechanisms to detect them [Bharat and Broder 1999]. The study of gateway and proxy traces also found replication on the Web and identified that a few Web servers are responsible for most of the duplicates [Douglis et al. 1997; Mogul 1999a]. A large client trace gathered from the WebTV networks evidenced the existence of URL aliasing and its implications to Web caching systems [Kelly and Mogul 2002].

On language analysis, the authors propose a technique for estimating the size of language-specific corpus and used it to estimate the usage of English and non-English language on the WWW [Grefenstette and Nioche 2000]. Funredes [2001] presented a study on the presence of Latin languages on the Web. Aires and Santos [2002] measured the Web, written in the Portuguese language.

The notion of hostgraph and connectivity of Web sites and country domains was presented in Bharat et al. [2001].

A first effort to characterize the Portuguese Web defined a set of metrics to describe the Web within the RCCN network (network that connected several Portuguese academic institutions) [Nicolau et al. 1997]. The Netcensus project aims at periodically collecting statistics regarding all type of files hosted under the .PT domain [Silva et al. 2002a, 2002b]. In our previous work, we presented a system for managing the deposit of digital publications and characterized a restricted set of Portuguese online publications, exposing the most common formats and file sizes [Noronha et al. 2001].

The statistics we gathered are sometimes significantly different from those presented in the bibliography. This is not a surprising result, since they are based in different and heterogeneous partitions of the Web, using distinct methodologies and obtained from different dates.

9. CONCLUSION AND FUTURE WORK

This article describes our work in identifying, collecting, and characterizing the Portuguese Web. We propose defining criteria that cover this Web with high precision and are simultaneously easy to configure when setting-up the harvesting policies on a crawler.

We observed that most of the sites are small, virtual hosts under the .PT domain. The number of sites under construction is very high. The use of appropriate or descriptive metatags is still insignificant on the Portuguese Web. We identified situations on the Web that may bias the results and proposed solutions, showing that Web characterization depends on the crawling technology that is used.

This study is interesting to others who need to characterize community Webs and may help in the design of software systems that operate over the Web. Web archivers can better estimate necessary resources and delimit partitions interesting for archival. Web proxies can be more accurately configured by administrators, crawlers can be improved through the definition of adequate architectures, and crawling policies of Web search engines can be used to improve their coverage of the Web, leading to better search results.

In future work, we will extend the characterization of the Portuguese Web to other MIME types and gather new metrics that will enable us to monitor

the evolution of the Web and its linkage structure. We also intend to improve the crawler performance so that statistics can be gathered in a shorter period of time. A major issue to be studied in the future is to develop a more accurate and efficient definition of the Portuguese Web. The current definition demands the download of large numbers of documents hosted outside the .PT domain in order to identify a very small percentage written in Portuguese. This is highly inefficient and makes it difficult for us to distinguish the documents of interest to our application domain from others. The difficulty is particularly high for sites hosted under general purpose TLDs. In future work, we intend to combine crawling policies with the usage of geographical tools such as Gtrace [Periakaruppan and Nemeth 1999] in order to obtain a more precise definition.

ACKNOWLEDGMENTS

We thank Miguel Costa and Bruno Martins for the discussions and development of software components that we used to extract the results presented in this article. We thank Marktest for giving us, the access to the Netpanel statistics.

REFERENCES

AIRES, R. AND SANTOS, D. 2002. Measuring the Web in Portuguese. In *Proceedings of the Euroweb Conference.* B. Matthews, B. Hopgood, and M. Wilson, Eds. Oxford, UK, 198–199.

ALBERTSEN, K. 2003. The paradigma Web harvesting environment. In *Proceedings of 3rd ECDL Workshop on Web Archives*. Trondheim, Norway.

BARR, D. 1996. RFC 1912. IETF.

BHARAT, K. AND BRODER, A. 1999. Mirror, mirror on the Web: A study of host pairs with replicated content. In *Proceedings of the 8th International Conference on the World Wide Web*. Elsevier, 1579–1590.

BHARAT, K., CHANG, B.-W., HENZINGER, M. R., AND RUHL, M. 2001. Who links to whom: Mining linkage between Web sites. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE Computer Society, 51–58.

BOLDI, P., CODENOTTI, B., SANTINI, M., AND VIGNA, S. 2002. Structural properties of the African Web. In *Proceedings of the 11th International World Wide Web Conference*. Honolulu, Hawaii.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst. 30*, 1–7, 107–117.

BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A., AND WIENER, J. 2000. Graph structure in the Web. In *Proceedings of the 9th International World Wide Web Conference on Computer Networks*. North-Holland Publishing Co., 309–320.

BRODER, A. Z., GLASSMAN, S. C., MANASSE, M. S., AND ZWEIG, G. 1997. Syntactic clustering of the Web. In *Proceedings of the 6th International Conference on the World Wide Web*. Elsevier, 1157–1166.

CAVNAR, W. AND TRENKLE, J. 1994. N-gram-based text categorization. In *the 3rd Annual Symposium on Document Analysis and Information Retrieval*. 161–175.

CENTER, H. S. D. 2003. Geo targeting IP address to country city region ISP latitude longitude database for Internet developers—ip2location. Available at `http://www.ip2location.com/`.

CHO, J. AND GARCIA-MOLINA, H. 2000. The evolution of the Web and implications for an incremental crawler. In *Proceedings of the 26th International Conference on Very Large Data Bases.* (Sept.) 10–14, 200–209.

CHO, J., GARCIA-MOLINA, H., AND PAGE, L. 1998. Efficient crawling through URL ordering. *Comput. Netw. ISDN Syst. 30*, 1–7, 161–172.

DAVIS, C., VIXIE, P., GOODWIN, T., AND DICKINSON, I. 1996. A means for expressing location information in the domain name system. RFC 1876. IETF.

DAY, M. 2003. Collecting and preserving the World Wide Web. Available at `http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf`.

DOUGLIS, F., FELDMANN, A., KRISHNAMURTHY, B., AND MOGUL, J. C. 1997. Rate of change and other metrics: A live study of the World Wide Web. In *the USENIX Symposium on Internet Technologies and Systems*.

FETTERLY, D., MANASSE, M., NAJORK, M., AND WIENER, J. L. 2003. A large-scale study of the evolution of Web pages. In *Proceedings of the 12th International World Wide Web Conference*. Budapest, Hungary.

FLAKE, G., LAWRENCE, S., AND GILES, C. L. 2000. Efficient identification of Web communities. In *the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Boston, MA. 150–160.

FUNREDES. 2001. The place of latin languages on the Internet. Available at `http://www.funredes.org/LC/english/L5/L5index_english.html`.

GIBSON, D., KLEINBERG, J. M., AND RAGHAVAN, P. 1998. Inferring Web communities from link topology. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*. Pittsburgh, PA. 225–234.

GOMES, D. 2003. Vúva negra. Available at www.tumba.pt/english/crawler.html.

GOOGLE. 2003. Google Web search features. Available at `www.google.com/help/features.html#link`.

GREFENSTETTE, G. AND NIOCHE, J. 2000. Estimation of english and non-english language use on the WWW. In *Proceedings of RIAO'2000—Content-Based Multimedia Information Access*. Paris, France. 237–246.

HARRENSTIEN, K., STAHL, M. K., AND FEINLER, E. J. 1985. NICNAME/WHOIS. RFC 954. IETF.

HENZINGER, M. 2003. Algorithmic challenges in Web search engines. *J. Internet Math. 1*, 1, 115–126.

HEYDON, A. AND NAJORK, M. 1999. Mercator: A scalable, extensible Web crawler. *World Wide Web 2*, 4, 219–229.

KELLY, T. AND MOGUL, J. 2002. Aliasing on the World Wide Web: Prevalence and performance implications. In *Proceedings of the 11th International World Wide Web Conference*. Honolulu, Hawaii.

LAWRENCE, S. AND GILES, C. L. 1999. Accessibility of information on the Web. *Nature 400*, 107–109.

LEUNG, S.-T. A., PERL, S. E., STATA, R., AND WIENER, J. L. 2001. Towards Web-scale Web archeology. Tech. rep. 174, (Sept.) Compaq Research Center, Paolo Alto CA.

LLC, M. 2003. Maxmind: How to locate your Internet visitors geotargeting IP address to country state city ISP organization latitude longitude. Available at `http://www.maxmind.com/`.

MARKTEST. 2003. Netpanel. Available at netpanel.marktest.pt/.

MOGUL, J. 1999a. A trace-based analysis of duplicate suppression in HTTP. Tech. rep. 99/2, (Nov.) Compaq Computer Corporation, Western Research Laboratory.

MOGUL, J. 1999b. Errors in timestamp-based HTTP header values. Tech. rep. 99/3, (Dec.) Compaq Computer Corporation, Western Research Laboratory.

NAJORK, M. AND HEYDON, A. 2001. On high-performance Web crawling. SR, Tech A68 Compaq Research Center, Palo Alto, CA.

NETCRAFT LTD. 2004. Netcraft: April 2003 archives. Available at `http://news.netcraft.com/archives/2003/04/index.html`.

NICOLAU, M. J., MACEDO, J., AND COSTA, A. 1997. Caracterização da informação WWW na RCCN. Tech. Rep., Universidade do Minho, Portugal.

NORONHA, N., CAMPOS, J. P., GOMES, D., SILVA, M. J., AND BORBINHA, J. 2001. A deposit for digital collections. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*. Springer-Verlag, 200–212.

OCLC. 2003. Web characterization. Available at `http://wcp.oclc.org/`.

O'NEILL, E. T. 1999. Web sites: Concepts, issues, and definitions. Available at `http://wcp.oclc.org/ pubs/rn1-websites.html`.

O'NEILL, E. T., LAVOIE, B. F., AND BENNETT, R. 2003. Trends in the evolution of the public Web. *D-Lib Magazine 9*, 4 (April).

OVERTURE SERVICES, I. 2003. Alltheweb.com: Frequently asked questions—URL investigator. Available at `www.alltheweb.com/help/faqs/url_investigator`.

PERIAKARUPPAN, R. AND NEMETH, E. 1999. GTrace: A graphical traceroute tool. In *Proceedings of the 13th USENIX Conference on System Administration*. 69–78.

PITKOW, J. E. 1998. Summary of WWW characterizations. *Comput. Netw. ISDN Syst. 30*, 1–7, 551–558.

POSTEL, J. 1994. Domain name system structure and delegation. RFC 1591. IETF.

PUNPITI, S. S. 2000. Measuring and analysis of the Thai World Wide Web. In *Proceedings of the Asia Pacific Advance Network*. 225–230.

RIVEST, R. 1992. The MD5 message-digest algorithm. RFC 1321. IETF.

SHIVAKUMAR AND GARCIA-MOLINA. 1998. Finding near-replicas of documents on the Web. In *the International Workshop on the World Wide Web and Web Databases*. Lecture Notes in Computer Science, vol. 1590, 204–212.

SILVA, L. O., MACEDO, J., COSTA, A., BELO, O., AND SANTOS, A. 2002a. Netcensus: Medição da evolução dos conteúdos na web. Tech. rep. Departamento de Informática, Universidade do Minho, Portugal.

SILVA, L. O., MACEDO, J., COSTA, A., BELO, O., AND SANTOS, A. 2002b. Obtenção de estatísticas do www em Portugal. Tech. rep. Universidade do Minho, Portugal.

SILVA, M. J. 2003. The case for a portuguese Web search engine. In *Proceedings of IADIS International Conference WWW/Internet*. Algarve, Portugal.

SPINELLIS, D. 2003. The decay and failures of Web references. *Comm. ACM 46*, 1, 71–77.

W3C. 1999. HTML 4.01 specification. Available at `http://www.w3.org/TR/html401/`.

W3C. 1999. Web characterization terminology and definitions sheet. Available at `http://www.w3.org/1999/05/WCA-terms/`.

WEBB, C. 2000. Towards a preserved national collection of selected Australian digital publications. In *Proceedings of the Preservation Conference*. York, UK.

WILLS, C. E. AND MIKHAILOV, M. 1999. Towards a better understanding of Web resources and server responses for improved caching. *Comput. Netw. 31*, 11–16, 1231–1243.

ZABICKA, P. 2003. Archiving the Czech Web: Issues and challenges. In *Proceedings of the 3rd ECDL Workshop on Web Archives*. Trondheim, Norway.

ZOOK, M. 2000. Internet metrics: Using host and domain counts to map the Internet. Telecomm. Policy, *24*, 6/7, 613–620.

Received December 2003; accepted June 2004