

Collecting Statistics about the Portuguese Web

Daniel Gomes
Mário J. Silva

DI-FCUL

TR-03-10

June 2003

Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1749-016 Lisboa
Portugal

Technical reports are available at <http://www.di.fc.ul.pt/tech-reports>. The files are stored in PDF, with the report number as filename. Alternatively, reports are available by post from the above address.

Collecting Statistics about the Portuguese Web

Daniel Gomes

Mário J. Silva

Universidade de Lisboa, Faculdade de Ciências

1700 Lisboa, Portugal

`dcgomes@xldb.di.fc.ul.pt`, `mjs@di.fc.ul.pt`

June 2003

Abstract

This report presents a characterization of text documents from the Portuguese Web. This characterization was produced from a crawl of over 4 million URLs and 131 thousand sites in 2003. We describe rules that we established for defining its boundaries and the methodology used to gather statistics. We also show how crawling constraints and abnormal situations on the Web can influence the results.

1 Introduction

The Web can be characterized from multiple perspectives using numerous metrics. Characterizing the Web is a challenging task [17]. Its characterization studies quickly become stale because of its permanent evolution. Its dimension limits the gathering of statistics to small samples of the WWW. The WWW is composed by partitions with common characteristics, which may not follow a general characterization [36]. A small partition may not be very relevant for a characterization of the global Web. However, its study can be very important for characterizing the Web of a given community.

The design and performance of applications that use the Web as a source of information is very sensitive to its characteristics [7]. We are studying the Portuguese Web, as a national Web having broadly defined the set of pages of cultural and sociological interest to the Portuguese people.

In this report, we present a characterization of the Portuguese Web. The results were derived from a crawl performed for tumbal!, a search and archival engine for the Portuguese Web [33]. We focused our study on textual contents available on the Portuguese Web, identifying metrics that will help us in the design and improvement of our system. In this study, we describe the methods used to gather statistics and interpret the results obtained.

This report is organized as follows: the remaining of this section presents the adopted terminology. Section 2, our heuristics for defining the boundaries of the Portuguese Web. In the following section, we present the crawler configuration. In section 4 we present the crawling results and in sections 5,6,7 we describe the statistics gathered on Portuguese sites, documents and web 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 [35].

- Publisher: entity responsible for publishing of information on the web;
- Document: file resultant 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 [24]).
- Host page: document identified by an URL where the file path component is empty or consists on a / only.
- Subsite: cluster of documents within a web site, maintained by a different publisher than that of the parent site.

2 Identifying the boundaries of the Portuguese Web

The Web is designed to break all the geographical barriers and make information available world-wide, independently from the physical location of the pages, we can find subsets of pages related to a country through a common country code Top Level Domain (ccTLD) [27] or the language spoken in the

country, but there isn't a definition of what a country Web is, and it is hard to come with a precise definition of what sites constitute a country web.

We define the Portuguese Web as the set of documents which contain information related to Portugal or of major interested to the Portuguese people. In practice, we consider documents that satisfy one of the following conditions:

- Hosted on a site under the .PT domain;
- Hosted on a site under the .COM, .NET, .ORG or .TV domains, written in Portuguese and with at least one incoming link originated in a web page hosted under the .PT domain.

This definition aims to be easily set as a crawling policy and guarantee that the crawler downloads documents that belong with high probability to the Portuguese Web. The objective of the first condition is to collect all the documents hosted under the ccTLD correspondent to Portugal. The second condition was motivated by the increasing number of Portuguese sites that are registered outside the .PT domain [37].

Our first approach for establishing the boundary of the Portuguese Web was to harvest all the documents written in the Portuguese language outside the .PT domain. However, this would require downloading a large number of pages, specially from Brazil, that have information not highly related to Portugal. The definition adopted mitigates this problem but it remains for Brazilian sites hosted under the allowed domains, such as .COM.

An alternative approach to find sites related to a country outside the ccTLD, is to include sites physically hosted in the country using the WHOIS database [29]. However, companies that provide hosting services support several distinct sites on the same machine and the WHOIS registries only keep information about the IP address of sub-networks, not of the web sites hosted. As a result, if we followed this approach, the Portuguese Web sites hosted outside of Portugal would not be considered as part of the Portuguese Web. This is a serious restriction if we consider, for instance, all the Portuguese users who have their homepages hosted in Yahoo!. WHOIS can then be an useful tool for a more accurate characterization of some of the sites of the Portuguese Web, but it is too weak to be used standalone to establish inclusion/exclusion criteria.

3 Crawler Configuration

A crawler begins its task of harvesting the Web collecting the contents of 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 Page Rank [8, 4]. On the other hand, archive crawlers focus on crawling the most pages on a given partition [9]. In our study, we configured Viúva Negra (VN)[13], the web crawler of the tumba! search engine to get the most information possible about the Portuguese Web and initialized it with a set of 112146 seeds gathered from previous crawls and user registrations. We imposed on it the minimum constraints that ensure an acceptable performance of the crawler, considering the resources available and the need to overcome existing pathological situations on the Web. 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 met:

- Documents of type text: we considered not only documents of the MIME type text but also documents of common application types that we could convert to text. The list of accepted MIME types is the following: 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 5: this means that the crawler followed at most 5 links from the seed of the site until it reached the 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 the Portuguese Web would be visited, including Portuguese subsites hosted on foreign sites. Consider for instance, the site www.yahoo.com and its subsite www.yahoo.com/users/-myPortugueseSite/. If the crawler visited only the seed www.yahoo.com it will identify that is not part of the Portuguese Web and exit without finding the Portuguese subsite;
- Each document was downloaded in less than 1 minute: prevents very slow web servers from blocking the progress of the crawl;
- Document size less than 2 MB: prevents the download of huge 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 crawl a site indefinitely. They are easily noticed due to the large number of documents discovered in the site[15]. 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 a posteriori 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, being caused mainly by session identifiers embedded in the URLs, or poorly designed HTTP web applications that dynamically generate infinite URLs and end up referencing a small set of documents.

This raises the issue of how should these documents be considered in a characterization. On the other hand, they should not be excluded because they are available online and represent part of the Web. However we can not let them bias a characterization due to its "infinite" presence. We adopted the solution of setting VN as a very patient web surfer as a compromise. After seeing the same document 50 times, VN gives up on following links for that site.

4 Crawling Statistics

The results presented were extracted from a crawl performed between the 1st of April and May, 15th 2003. VN visited a total of 131864 sites, processed over 4 million URLs and downloaded 78 GB of data.

Table 1 presents the statistics of the download status of crawled URLs. We gladly noticed that almost 84% of the requests resulted in a successful download and that 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 very frequent. There were over 6% of redirections and the crawler failed to process 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 (ex. files with extension jpeg) were not crawled. The UnknownHost error (0,5%) is caused by URLs referencing host names that no longer have an IP associated. We found that only 0,5% of the referenced files had a size bigger than 2MB and

State	Number	%
200	3235140	83,9
302	193870	5,0
404	132834	3,4
TimedOut (-8)	45486	1,2
301	39920	1,0
ExcludedByREP (-2)	35596	0,9
500	33247	0,9
NotAllowedType (-5)	25976	0,7
403	18598	0,5
UnknownHost (-14)	17842	0,5
SizeTooBig (-4)	17453	0,5
ConversionError (-11)	13986	0,4
Others	23244	0,6
Total	3856436	100,0

Table 1: 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.

the conversion to text was not possible in 0,4% of the cases. The remaining situations (0,6%) included other HTTP response codes, unidentified errors, socket and connection errors; each of them represented less than 0,1% of the total number of downloaded documents.

5 Site Statistics

We identified 46457 sites as being part of the Portuguese Web, 85% of them 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 sites begun their fully qualified name with "WWW".

A Portuguese web site has an average of 69 documents but their distribution is very skewed, as shown in Figure 2. We were surprised by the high number of sites composed by a single document, over 38%, so we analyzed some of them and found out that they were mostly under construction, or they had a host page informing that the site has moved to a different location. We also found a few cases where the host page was completely written using scripting languages from which our parser couldn't extract links. A typical Portuguese web site has less than 100 documents (93%), 6% have

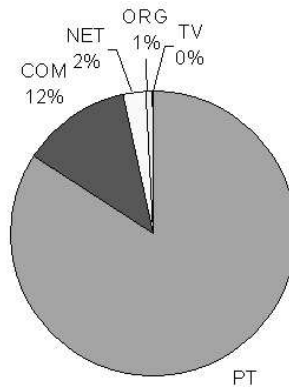


Figure 1: Distribution of domains.

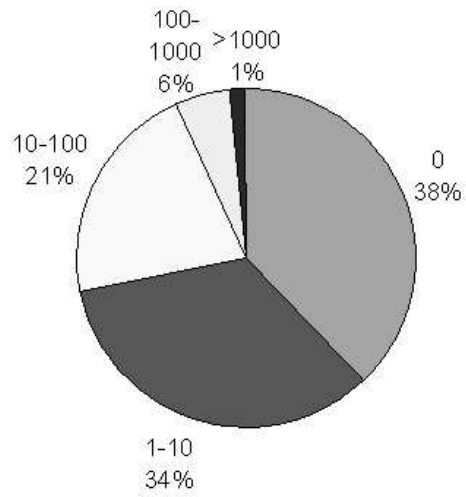


Figure 2: Distribution of documents per site.

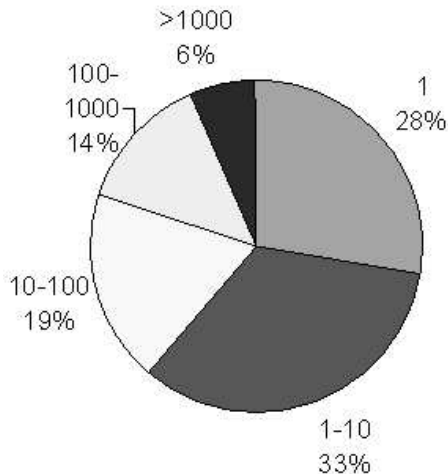


Figure 3: Distribution of documents per IP address.

Number of Sites	Number of IPs	% of IPs with number of sites
1	4643	67,7
2-10	1931	28,2
10-100	247	3,6
100-1000	30	0,4
>1000	5	0,1
Total	6856	100,0

Table 2: Distribution of hosts if by IP regarding the number of sites

between 10 and 100 documents and just 1% of the sites have more than 1000 documents.

The distribution of documents per IP address is more uniform as we can see in Figure 3, the percentage of IP addresses that host just one document is of 28%, IP addresses that host 1 to 10 documents represent 33%, 14% host 100 to 1000 and only 6% more than a 1000 documents.

Table 2 shows that over 32% of the IP addresses host more than 1 site. There are 5 IP addresses that host more than a 1000 sites, the last ones correspond to Web portals that offer their clients a host name under the portal domain, providing a proper host name for their site, instead of having it as a subsite. This feature is achieved through the usage of virtual hosts, which enable one web server to host several distinct sites. 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.

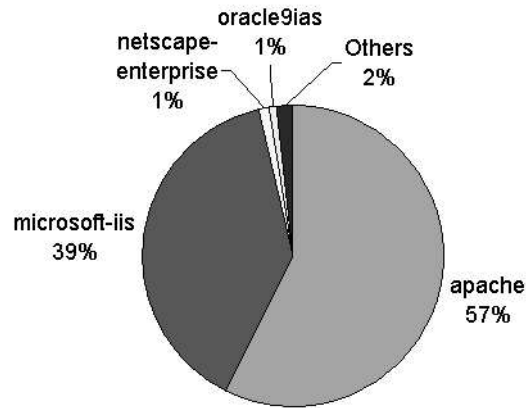


Figure 4: Distribution of web servers.

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`, while distinct virtual hosts are distinct sites hosted on the same machine, such as `http://www.tumba.pt` and `http://ul.tumba.pt`. We considered that if two sites identified by different names share the same IP address and have the same host page, they are host aliases. In our crawl we found out that 8,5% of the virtual hosts are host aliases.

5.1 Web Servers

Figure 4 presents the distribuion of web servers. The Portuguese sites are mainly hosted in Apache (57%) and Microsoft IIS web servers (39%). The next two web servers (netscape-enterprise and oracle9ias) represent just 1% each and the remaining 168 web servers just 2%.

6 Document Statistics

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

6.1 URLs

Every web application must have some kind of data structures that maps into URLs. However, we didn't find in the literature a study discussing the lengths of the URLs. Nowadays, the size of URLs, is in practice unlimited.

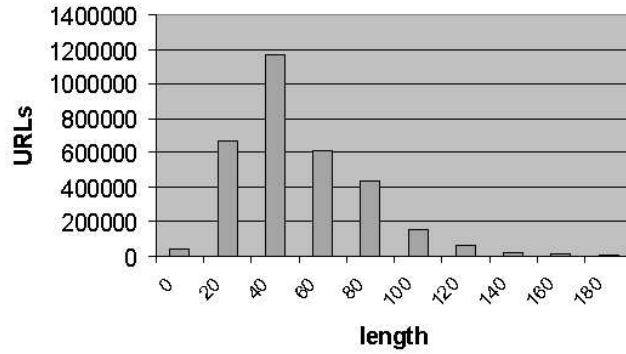


Figure 5: Distribution of URL lengths

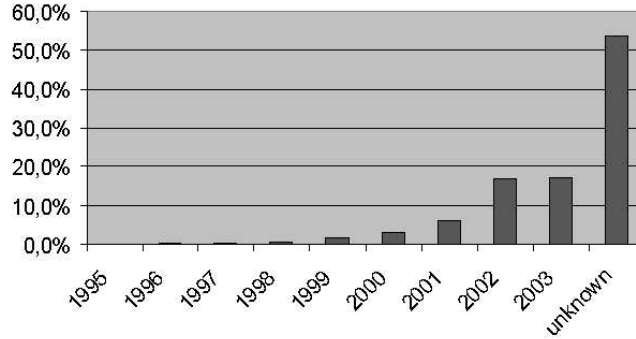


Figure 6: Distribution of Last-Modified Dates

We found valid URLs with lengths varying from 5 to 1368 characters. Figure 5 shows the distribution of URL lengths (not considering the initial 7 characters of the protocol) over the number of the documents. Most of the documents have an URL length between 20 and 100 characters, with an average value of 62 and median of 54. Analyzing the URLs we found that 2,3% contained parameters suggesting that the correspondent document had been dynamically generated.

6.2 Last Modified Date

HTTP provides a header field (Last-Modified Date) that should indicate the date of last modification of the document. However, as shown in Figure 6,

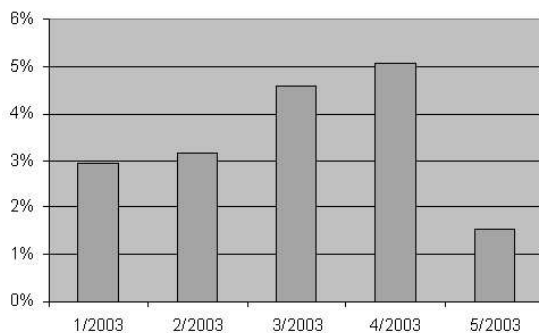


Figure 7: Distribution of Last-Modified Dates in the last 4 months

most of the documents (53,5%) returned an unknown value for this field. Plus, Mogul proved that even the returned values are many times inaccurate due to the web servers clocks are being set incorrectly (among other situations)[19]. An analysis of the URLs correspondent to the unknown values revealed that 82% of them had embedded parameters. We speculate that most of them are recent and they would significantly increase the percentage of documents modified in the last months (Figure 7), since mechanisms to dynamically generate documents are usually used to reference short life contents, such as news.

We think that the last-modified header is a weak metric for evaluating changes and evolution of contents on the Web, so metrics like these must be considered only in the context of analysis of consecutive crawls [11].

6.3 MIMEs & Sizes

The right column of Table 3 shows the distribution of documents per MIME type, (we grouped all the MIME types correspondent to PowerPoint files under the name PPT and all the ones correspondent to EXCEL files under the name EXCEL). The predominant text format is text/html, present in over 95% of all 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, so we 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 where the size grew from 12,2 KB to 20,5 KB. In Table 4, the second and third columns show the average sizes of documents and corresponding extracted

MIME	Number of documents	Presence (% of docs)
text/html	3104771	95,9
application/pdf	62141	1,9
text/plain	33091	1,0
application/x-shockwave-flash	17598	0,5
application/msword	14014	0,4
PPT	2085	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	3235140	100,0

Table 3: Number of documents and relative presence on the Web for each MIME type collected

% text MIME	avg Doc Size(KB)	avg Text Size(KB)	% text
PPT	1054,9	7	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

Table 4: Average size, extracted text size, percentage of extracted text

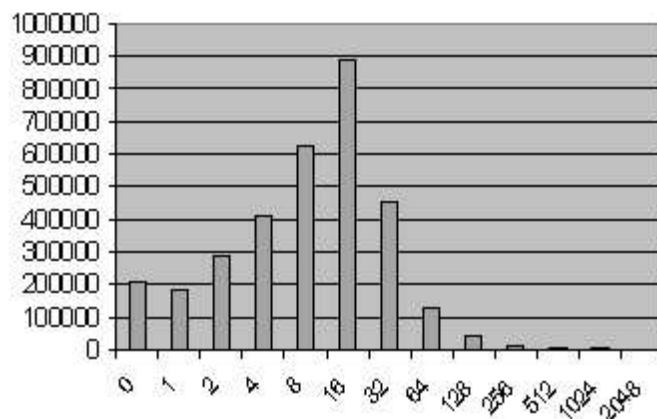


Figure 8: Distribution of sizes considering all files

texts (without any formatting tags), 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, when don't recognize the file type return text/plain. This resulted that 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 64 KB. The mean size of a document is 32,4 KB and the mean size of the extracted texts is 2,8 KB. The total size of the documents was 78,430 GB, while the total size of extracted texts was just 8,791 GB.

6.4 Language Distribution

VN can identify the language of collected documents based on an idiom detector that implements an n-gram algorithm [6]. 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. Identifying the language of a document is sometimes a hard task because there are documents very poor in text or written in several languages. In 4% of the documents

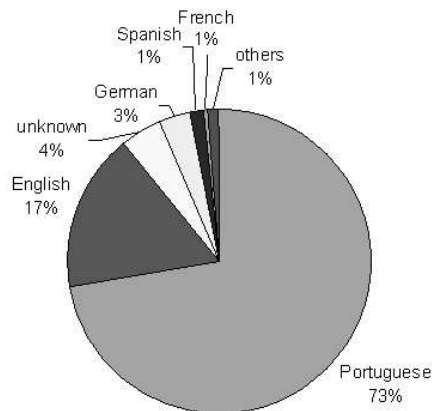


Figure 9: Distribution of Languages

the idiom detector couldn't identify the language of the document.

6.5 Meta-tags

We studied the usage of two important meta-tags supported by HTML: *description* and *keywords* [34]. The description meta-tag gives a content's description of the page and the meta-tag keywords provides a set of keywords that describe the contents of the site. We found that just 17% of the pages had the meta-tag description and that among these the usage of this meta-tag doesn't seem to be correct. We found out only 44 thousand distinct values for 555 thousand description meta-tags. This means that 92% of the texts of the descriptions are 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 there are different;
- There are replicated pages on the web.

The keywords meta-tag is present in 18% of the pages, a surprising result because it is recommended that the tag should be present only on the host page of each site. It should then be less frequent than the description tag. A deeper analysis revealed that 91% of the pages that have the description meta-tag also had the keywords meta-tag.

Number of replicas	Number of Contents	% of contents
0	2462490	90,0
1	205882	7,5
2	33468	1,2
3	12814	0,5
4	6086	0,2
5	5272	0,2
6-10	6453	0,2
10-100	2318	0,1
100-1000	154	0,0
>1000	5	0,0
Total	2734942	100,0

Table 5: Distribution of contents with replicas

The titles of the web pages aren't very descriptive either. There were over 600 thousand 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

We found out that 15,5% of the downloaded documents corresponded to a content already downloaded under a different URL (replicas). We identified 2734942 distinct contents, (Table 5 describes the replication distribution). Most of the contents are unique (90%) and 9.96% had at least one replica. Contents replicated more than 1000 times are very rare. However, they caused 13146 downloads for just 5 distinct contents. These situations are pathological for web crawlers and also tend to bias the collected statistics. We manually analyzed these 5 cases and concluded that they were all caused by mal-functioning of web servers that always return an error page for all the requests. Our measures against these traps (see Section 3.1) failed because all the links that originated the error messages were extracted from correct pages. When the crawler 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 contents had replicas inside the same site, 60% had replicas outside the site and 2% had replicas inside and outside the site.

Number of outlinks	Number of Pages	% of docs
0	3043636	98
1	44255	1,4
1-10	14677	0,5
10-100	2186	0,1
>100	17	0,0
Total	3104771	100,0

Table 6: Distribution of outlinks

Number of inlinks	Number of docs	% of docs
0	3172317	98,1
1	44511	1,4
1-10	16789	0,5
10-100	1479	0,0
>100	44	0,0
Total	3235140	100%

Table 7: Distribution of inlinks

7.2 Link Structure

Our link analysis was quite limited because we only gathered information about links between distinct sites within the Portuguese Web. We found that 98% of the Portuguese pages don't have a link to another site within the Portuguese Web (Table 6) and that 98,1% of the documents are not referenced by a link originated in another Portuguese site (Table 7). This suggests that the Portuguese sites are weakly inter linked. However, we found some pages which are rich in links to Portuguese sites (hubs), as we can see in Table 8. In Table 9 we show the top 20 more linked sites within the Portuguese Web (authorities).

8 Related Work

Web characterization has been done from different perspectives through the years almost since the beginning of the Web [26]. The Web Characterization Project has been a great contributor for research in Web characterization [23, 25].

Najork and Heydon performed a large scale web crawl from which they gather statistics regarding the outcome of download attempts, distribution

Pos.	URL	Out-Links
1	cpan.dei.uc.pt/modules/00modlist.long.html	2567
2	ftp.ist.utl.pt/pub/rfc/	2425
3	homepage.oninet.pt/095mad/bookmarks_on_mypage.html	2309
4	cpan.dei.uc.pt/authors/00whois.html	1621
5	www.fba.ul.pt/links4.html	1532
6	www.esec-canecas.rcts.pt/Educacao/Escolas.htm	1346
7	pisco.cii.fc.ul.pt/nobre/hyt/bookmarks.html	1287
8	www.fpce.uc.pt/pessoais/rpaixao/9.htm	1181
9	www.jn.sapo.pt/Cyber/urls/portugal.htm	1099
10	www.deb.uminho.pt/Fontes/enviroinfo/publications/default.htm	1060
11	free.clix.pt/ports/devel.html	1013
12	www.indeks.pt/0001.htm	990
13	ftp.ist.utl.pt/pub/CPAN/modules/00modlist.long.html	973
14	groups.google.pt/intl/pt/options/universities.html	787
15	jaamaro.pt/default.asp?news=15	764
16	www.di.fc.ul.pt/jbalsa/univs/U.html	709
17	www.animais.jcle.pt/classificados/vende/	665
18	www.sacaparte.pt/bi_ct_00.asp?idioma=1&local=70200&menu=89	641
19	www.eb23-cercal-alentejo.rcts.pt/Cercal/links.htm	639
20	alfa.ist.utl.pt/~farsilva/.private/PEANUTS_index-e.html	619

Table 8: Top 20 hubs

Position	URL	InLinks
1	www.fcn.pt/	7005
2	www.sapo.pt	961
3	www.infocid.pt/	489
4	www.dn.pt	383
5	www.fct.mct.pt	339
6	www.ist.utl.pt	337
7	www.uminho.pt	309
8	security.vianetworks.pt/	305
9	www.caleida.pt/	294
10	www.parlamento.pt	260
11	www.paginasamarelas.pt	222
12	tucows.telepac.pt	173
13	www.cp.pt	161
14	www.expresso.pt/	145
15	www.iapmei.pt	144
16	www.record.pt	140
17	www.abola.pt	124
18	www.rtp.pt	119
19	www.cgd.pt	118
20	www.up.edu.pt	115

Table 9: Top 20 Authorities

of types and size of the documents, replication and they also witnessed that the distribution of pages over web servers follows a Zipfian distribution [20]. Lawrence and Giles studied the accessibility of information on the Web and draw conclusions about the size, extracted text and usage of meta-data in HTML pages [16].

Boldi et al. studied the structural properties of the African Web analyzing HTTP header fields and contents of HTML pages [3] and Punpiti et al. presented quantitative measurements and analyses of documents hosted under the .th domain [28].

Replication on the web as been studied in several works, through the syntactically clustering of documents [5], the study of the existence of near-replicas on the Web [31] and different levels of duplication between hosts and mechanisms to detect them [1]. A study of analysis 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 [10, 18].

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 [14]. Funredes presented a study on the presence of latin languages on the Web [12].

The notion of hostgraph and connectivity of web sites and country domains was presented in [2].

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) [21]. The Netcensus project aims to periodically collect statistics regarding all type of files hosted under the .PT domain [32, 30]. 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 [22].

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

9 Conclusion and Future Work

This report described our work in identifying, collecting and characterizing the Portuguese Web. We propose a criteria for definition of the Portuguese Web, which corresponds to a precise coverage of this Web and is simultaneously easy to configure when setting-up the harvesting policies on a crawler.

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 meta-tags 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 used crawling technology.

This study 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.

As future work, we intend to extend the characterization of the Portuguese Web to other MIME types and gather new metrics of characterization, that would enable us to study 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 define a more accurate and efficient definition of the Portuguese Web, hence the current definition demands the the downloading of large numbers of documents hosted outside the .PT domain, to identify the small percentage of them written in Portuguese. This is highly inefficient, and makes it difficult for us unable to distinguish the documents of interest to our application domain, namely those in Portuguese and Brazilian sites, when they are hosted in general purpose TLDs.

10 Acknowledgements

We thank Bruno Martins for the discussions and development of software components that we used to extract the results presented in this report.

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 (project XMLBase) and SFRH/ BD/ 11062/ 2002 (scholarship).

References

- [1] K. Bharat and A. Broder. Mirror, mirror on the web: a study of host pairs with replicated content. In *Proceedings of the eighth in-*

- ternational conference on World Wide Web*, pages 1579–1590. Elsevier North-Holland, Inc., 1999.
- [2] K. Bharat, B.-W. Chang, M. R. Henzinger, and M. Ruhl. Who links to whom: Mining linkage between web sites. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 51–58. IEEE Computer Society, 2001.
 - [3] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Structural Properties of the African Web. In *Proceedings of the 11th International World Wide Web Conference*, Honolulu, Hawaii, May 2002.
 - [4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
 - [5] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. In *Selected papers from the sixth international conference on World Wide Web*, pages 1157–1166. Elsevier Science Publishers Ltd., 1997.
 - [6] W. Cavnar and J. Trenkle. N-gram-based text categorization. In *Third Annual Symposium on Document Analysis and Information Retrieval*, pages pages 161–175, 1994.
 - [7] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, pages 200–209, 2000.
 - [8] J. Cho, H. García-Molina, and L. Page. Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1–7):161–172, 1998.
 - [9] M. Day. Collecting and preserving the world wide web. http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf, 2003.
 - [10] F. Douglass, A. Feldmann, B. Krishnamurthy, and J. C. Mogul. Rate of change and other metrics: a live study of the world wide web. In *USENIX Symposium on Internet Technologies and Systems*, 1997.
 - [11] D. Fetterly, M. Manasse, N. M., and J. Wiener. A Large-Scale Study of the Evolution of Web Pages. In *Proceedings of the 12th International World Wide Web Conference*, Budapest, May 2003.

- [12] Funredes. The place of latin languages on the internet, 2001.
- [13] D. Gomes and M. J. Silva. The gorky details behind a web crawler. Technical report, Universidade de Lisboa, Faculdade de Ciências, Departamento de Informática, 2003. in preparation.
- [14] G. Grefenstette and J. Nioche. Estimation of english and non-english language use on the WWW. In *Proceedings of RIAO'2000, Content-Based Multimedia Information Access*, pages 237–246, Paris, 12–14 2000.
- [15] A. Heydon and M. Najork. Mercator: A Scalable, Extensible Web Crawler. *World Wide Web*, 2(4):219–229, 1999.
- [16] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 400:107–109, 1999.
- [17] S.-T. A. Leung, S. E. Perl, R. Stata, and J. L. Wiener. Towards Web-Scale Web Archeology. Research Report 174, Compaq Reseach Center, Paolo Alto CA, September 2001.
- [18] J. Mogul. A trace-based analysis of duplicate suppression in HTTP. Technical Report 99/2, Compaq Computer Corporation Western Research Laboratory, November 1999.
- [19] J. Mogul. Errors in timestamp-based HTTP header values. Research Report 99/3, Compaq Computer Corporation Western Research Laboratory, December 1999.
- [20] M. Najork and A. Heydon. On high-performance web crawling. Src research report, Compaq Systems Research Center, 2001.
- [21] M. J. Nicolau, J. Macedo, and A. Costa. Caracterização da informação www na rccn. Technical report, Universidade do Minho, 1997.
- [22] N. Noronha, J. P. Campos, D. Gomes, M. J. Silva, and J. Borbinha. A deposit for digital collections. In *Proc. 5td European Conf. Research and Advanced Technology for Digital Libraries, ECDL*, pages 200–212. Springer-Verlag, 2001.
- [23] OCLC. Web characterization. <http://wcp.oclc.org/>.
- [24] E. T. O'Neill. Web Sites: Concepts, Issues, and Definitions, 1999.
- [25] E. T. O'Neill, B. F. Lavoie, and R. Bennett. Trends in the Evolution of the Public Web. *D-Lib Magazine*, 9(4), April 2003.

- [26] J. E. Pitkow. Summary of WWW characterizations. *Computer Networks and ISDN Systems*, 30(1–7):551–558, 1998.
- [27] J. Postel. Domain name system structure and delegation. Available at URL <http://www.ietf.org/rfc/rfc1591.txt?number=1591>, 1994.
- [28] S. S. Punpiti. Measuring and Analysis of the Thai World Wide Web. In *Proceedings of the Asia Pacific Advance Network*, pages 225–230, August 2000.
- [29] RIPE. Query the RIPE whois database. <http://www.ripe.net/db/whois/whois.html>.
- [30] A. Santos, A. Costa, J. Macedo, O. Belo, and L. Silva. Obtenção de estatísticas do www em portugal. Technical report, OCT and DI, Universidade do Minho, 2002.
- [31] Shivakumar and Garcia-Molina. Finding near-replicas of documents on the web. In *WEBDB: International Workshop on the World Wide Web and Databases, WebDB*. LNCS, 1999.
- [32] L. Silva, J. Macedo, and A. Costa. NetCensus: Medição da evolução dos conteúdos na web. Technical report, Departamento de Informática, Universidade do Minho, 2002.
- [33] M. J. Silva. The case for a portuguese web search engine. DI/FCUL TR 03–03, Department of Informatics, University of Lisbon, March 2003.
- [34] W3C. HTML 4.01 Specification. <http://www.w3.org/TR/html401/>.
- [35] W3C. Web Characterization Terminology and Definitions Sheet. <http://www.w3.org/1999/05/WCA-terms/>, 1999.
- [36] C. E. Wills and M. Mikhailov. Towards a better understanding of Web resources and server responses for improved caching. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1231–1243, 1999.
- [37] M. Zook. Internet Metrics: Using Host and Domain Counts to Map the Internet., 2000.