Trends in Web characteristics

João Miranda
Foundation for National Scientific Computing
1708-001 Lisboa
Portugal
joao.miranda@fccn.pt

Daniel Gomes
Foundation for National Scientific Computing
1708-001 Lisboa
Portugal
daniel.gomes@fccn.pt

ABSTRACT

The Web is a hypertextual environment that is on permanent evolution. There are new technologies and Web publishing behaviors that emerge everyday. It is important to track trends on its evolution to develop efficient tools to process its data. This study presents trends on the evolution of the Web derived from the analysis of the evolution of a Web portion by comparing characterizations performed within 5 years of interval. The Web portion used as a case study was the Portuguese Web. Several metrics regarding content and site characteristics were analyzed. We believe that the obtained trends are representative of the evolution of the Web in general.

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

Trends, Statistics, Web, Internet

Keywords

Web trends, Web characterization, Web measurements

1. INTRODUCTION

In Web's early days only computer experts were able to publish a Web page. In 2009, any common Internet user with few technical skills can become a mass Web publisher using free tools and services, such as blog platforms, wikis, content management systems or Google sites. Therefore, the Web is prone to suffer significant changes on its characteristics within a few years, affecting, for instance, the media types commonly used for publication. It is important to track trends on the evolution of the Web to develop efficient tools to process its data. Web characterization is a research field that contributes to model its characteristics across time [28]. However, it is impossible to gather an instant snapshot of the whole Web. Therefore, Web characterization studies are limited to the analysis of selected Web portions. Web characterization studies can be performed using distinct methodologies to select a Web portion for analysis. However, to derive Web evolution trends, it is important that a similar methodology is applied on the Web characterizations performed across time. For instance, a Web characterization extracted from a university proxy log should not be compared with one gathered from Web crawling. The proxy log characterization reflects a Web portion composed only by the contents that were accessed by the proxy users, while the Web crawl contains a broad scope of the information available on the Web. Nonetheless, a crawler iteratively harvests contents from the Web by following links on pages and it is unable to find contents that do not receive any links.

This study compares the obtained results from a characterization of a Web portion performed in 2008 with previous studies to derive evolution trends [11, 13]. The Web portion used as a case study was the Portuguese Web. A national Web contains a broad scope of publication genres including most of those present on the global Web, such as news, blogs or commercial sites. Although a national Web may present peculiar characteristics, such as language dominance, there are prevalent characteristics across Web portions. According to Baeza-Yates et al. the results obtained for several national Web characterizations show that there are characteristics shared across countries and valid on the global Web, such as URL length or HTTP responses distributions [1]. Thus, we believe that the obtained results from the Portuguese Web for the presented metrics reflect the trends of the global Web. The Portuguese Web was also chosen because it is relatively small and can be exhaustively harvested. Plus, it was thoroughly characterized in the past using a similar methodology, which enables comparisons from several perspectives. The main contribution of this study is an analysis of the trends of Web characteristics.

This paper is organized as follows. Section 2 presents the related work. Section 3 presents the methodology adopted. Section 4 presents statistics related to contents collected. Section 5 describes the results obtained regarding the characteristics of sites. Section 6 presents the conclusions and future work.

2. RELATED WORK

Web characterization has been a subject of several studies. Pitkow presented a summary of the first efforts to characterize the Web [26]. Najork and Heydon performed a large scale crawl producing several statistics [20]. Boldi et al. presented structural information on the African Web, including structure of pages and most used technologies [5]. The characterization of national webs has received a lot of attention from the research community. Baeza-Yates et al. characterized and compared the Korean, Chilean and Greek webs showing similarities that contribute to validate general models for Web characteristics [4, 2]. They also published an in-depth study of the Spanish Web [3] and performed a comparison of the results of 12 Web characterization studies comprising over 120 million pages from 24 countries, unveiling similarities and differences between the collections [1]. Tolosa et al. presented the characteristics of the Argentinian Web from a crawl over 10 million pages from 150 000 sites performed in 2006 [27]. Zabicka and Matejka analysed the Czech Web archive, performing a characterization of the archived contents [29].

The Web Characterization Project analyzed the trends in the size and content of the Web [24]. Modesto et al. characterized the evolution of the Brazilian Web between 2000 and 2005, making a comparison with the results previously obtained [18]. O'Neill et al. presented key trends in the evolution of the public Web from 1998 to 2002, analyzing its total size, growth, internationalization and metadata usage [25]. Funredes and Union Latine have been studying the presence of languages and cultures on the Web since 1996 [9]. Lasfargues et al. presented a characterization of the French Web based on a crawl performed in 2007 and its evolution based on annual crawls using different methodologies performed since 2004 [17].

There were previous studies that contributed to characterize the Portuguese Web. Nicolau et al. defined a set of metrics to characterize the Web within the national scientific community network [21]. Noronha et al. presented a system for supporting the archive of Web publications in a digital library. They performed a crawl of selected publications and characterized the obtained collection [23]. Gomes et al. produced two previous characterizations of the Portuguese Web that will be used as baseline in this study to derive trends on Web characteristics [11, 13].

3. METHODOLOGY

The following terminology was adopted in this study. A *crawler* is a program that iteratively downloads contents and extracts links to find new ones. A *seed* is a URL used in the set of initial addresses to visit when starting a crawl. A *site* is identified by a fully qualified domain name. For instance, `www.fccn.pt` and `arquivo-web.fccn.pt` are two different sites. Each different subdomain of a second (third, fourth...) level domain is assumed to be a different site. A *content* is a file resulting from a successful HTTP download: a request returned with a 200 response code (Successful - OK). The amount of information published here is in decimal multiples: 1 KB = 10^3 bytes [15].

The Portuguese Web Archive (PWA) project aims to automatically gather and preserve the information published on the Portuguese Web [12]. The most recent Web characterization results presented in this study were extracted from a crawl of the Portuguese Web performed by the PWA in 2008, that included all media types, which we named **allmedia08**. There were two previous studies that will be used as baseline to derive trends through their comparison with the results obtained in the crawl performed in 2008 by the PWA. The first study presented a thorough characterization of the Portuguese Web derived from a crawl of 3.2 million textual contents performed in 2003 [13], which we will henceforth refer to as **textual03**. The second presented the most prevalent media types on the Portuguese Web, based on a crawl performed in 2005 [11], which we will henceforth call **allmedia05**.

Status Code	% allmedia08	% textual03	Description
200	85.2%	88.4%	OK
302	7.2%	5.3%	Found
404	5.1%	3.6%	Not Found
301	1.3%	1.1%	Moved Permanently
303	0.4%	0.0%	See Other
403	0.2%	0.5%	Forbidden
500	0.2%	0.9%	Internal Server Error
400	0.2%	0.1%	Bad Request
401	0.2%	0.1%	Unauthorized
503	0.1%	0.0%	Service Unavailable
Other	0.0%	0.0%	Other codes

Table 1: The ten most common response codes logged while harvesting the Portuguese Web in allmedia08 and comparison with textual03.

The allmedia08 crawl was performed by the PWA between March and May, 2008, using Heritrix 1.12.1 [19], and started from a set of 180 000 seeds under .PT. These seeds were generated from a previous crawl. This crawl had the objective of achieving a high coverage of the Portuguese Web.

The methodology used to define the crawled Web portions can bias the obtained characterizations. Therefore, we expose the differences found between the methodologies used to crawl allmedia08 and the previous crawls textual03 and allmedia05, and discuss their impact on the obtained results. The textual03 was obtained to feed a search engine. Hence, only textual contents were crawled (*html*, *text*, *pdf*, *flash*, *word*, *powerpoint*, *excel*, *tex* and *rtf*) using the Viúva Negra crawler [14]. Notice that, except for plain text, all these formats are able to contain hypertextual features. Thus, we can consider that this crawl was composed, in general, by hypertexts. The crawl contained 3.2 million contents and the file size limit was 2 MB. The allmedia08 was crawled to feed a Web archive using Heritrix and all media types were harvested. Therefore, when comparing results extracted from allmedia08 to textual03, we considered only the subset of textual media types harvested in both crawls. Thus, we named as **textual08** this subset of contents present in allmedia08. The allmedia08 crawl contained over 48 million contents and the file size limit was 10 MB. Both allmedia08 and textual03 were harvested considering the .PT domain as the core of the Portuguese Web and included contents hosted under other domains. However, in textual03 a language detection mechanism was also used as a selection criteria to identify contents hosted outside the .PT domain. The methodology used to crawl allmedia05 was similar to the one used to crawl textual03, except that all media types were included. Thus, the characteristics obtained from allmedia05 and allmedia08 are compared directly.

Table 1 presents the ten most logged response status codes

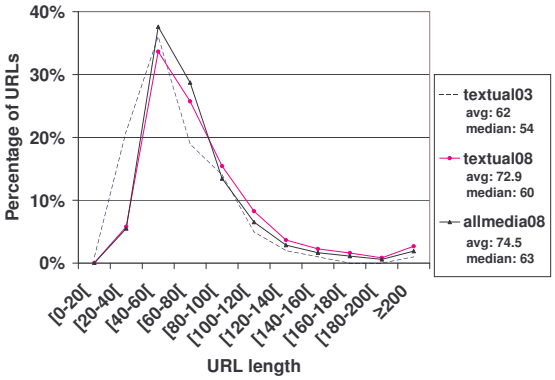

Figure 1: URL length distribution for textual03, textual08 and allmedia08.

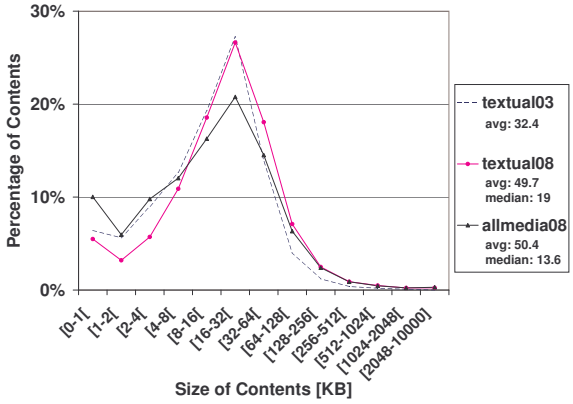

Figure 2: Content size distribution in KB for textual03, textual08 and allmedia08.

in allmedia08 and the values found for those codes in textual03, excluding error codes logged by the crawler. Redirects are inconvenient for Web archives based on the Wayback Machine, since the Wayback Machine does not process them correctly. The total number of logged responses was 3 660 121 in textual03 and 57 148 455 in allmedia08. The observed values differ slightly between crawls but the response code distribution is similar.

We believe that the presented methodological differences did not have a significant impact on the derived trends for Web characteristics.

4. CONTENTS

This Section presents the trends on Web content characteristics that can be used, for instance, to enhance browsers. In allmedia08, the number of contents excluded due to REP was 9.4% of the requests processed. The percentage observed in textual03 was 0.9%. Unlike the crawler used in textual03, which followed only the rules determined by *robots.txt*, Heritrix also takes into account the robots meta tags [7] from the Web pages code. This is a reason for the increase of the REP exclusions when compared to textual03.

4.1 URL length

The URL length of contents is a feature used in search engine ranking algorithms to identify relevant results [8]. In allmedia08, the URL length was counted as the number of characters excluding the protocol element to replicate the methodology followed in textual03 and enable an accurate comparison for trend analysis. For instance, in the URL `http://www.a.com/b.php?f=2` only the `www.a.com/b.php?f=2` string was considered, thus this URL presents a length of 19 characters.

In allmedia08 valid URLs with lengths ranging from 5 to 2 072 characters were found. Figure 1 presents the URL length distribution for textual03, textual08 and allmedia08. The obtained results show that URL length tends to increase with time. According to the model provided by Gomes [10], the expected URL length for 2008 was 67.6 characters, which represents an error of 7.8%. The obtained results for allmedia08 are similar to those obtained for textual08 and show that the distribution of URL lengths for textual contents is representative of all media types.

4.2 Media types and sizes

New hypertextual formats appear everyday and others evolve to include hypertextual features. On the other hand, some formats disappear. This Section presents trends on content size per media type and prevalence measured by number of contents and amount of data provided.

4.2.1 Average size

Analyzing trends in content sizes is useful, for instance, to estimate the resources required to create Web data repositories.

Table 2 compares the average size of the contents grouped by media type in textual03 and textual08. The content file sizes in textual03 might have been underestimated for media types that are typically large because the file size limit then used was 2 MB. Thus, the average size for textual08 was analyzed considering also this limit to enable an accurate trend analysis. The presented trend in the last column of Table 2 refers to the comparison between textual03 and textual08 with a 2 MB size limit. The obtained results show that except for *powerpoint*, *text/rtf* and *text/tab-separated-values*, the content size for all media types tends to grow. For instance, between 2003 and 2008 the average size for *text/html* contents grew from 21 KB to 30 KB. The average size for certain media types was significantly different when considering different limit sizes. For instance, the average size for *application/pdf* contents considering a limit of 2 MB was 252 KB and grew to 483 KB when this limit was raised to 10 MB. The obtained results show that different content size limits should be imposed during a crawl according to media types.

Figure 2 presents the general distribution of content size across time. The imposed maximum content size limit of 10 MB in allmedia08 resulted in a total of 32 321 truncated contents, which represents just 0.05% of the total downloaded contents. In textual03, 0.5% of the contents achieved the limit size of 2 MB imposed then. Therefore, we assume that these constraints did not bias the obtained results for the general content size distribution.

The distribution for textual contents is similar between textual03 and textual08 with approximately 73% of the contents having sizes between 4 and 64 KB. However, in the later crawl there is a decrease in the number of contents

Media type	Avg Size textual03 <2MB	Avg Size textual08 <2MB	Trend	Avg Size textual08 <10MB
text/html	21 KB	30 KB	45.9%	31 KB
app'n/pdf	207 KB	252 KB	21.6%	483 KB
text/plain	11 KB	44 KB	318.9%	212 KB
app'n/x-shockwave-flash	44 KB	90 KB	104.7%	144 KB
app'n/msword	119 KB	145 KB	22.6%	216 KB
powerpoint	1 055 KB	500 KB	-52.6%	1 134 KB
excel	50 KB	118 KB	135.0%	158 KB
text/rtf	476 KB	143 KB	-70.0%	321 KB
app'n/rtf	121 KB	179 KB	47.7%	206 KB
app'n/x-tex	16 KB	18 KB	9.3%	19 KB
text/tab-separated-values	4 KB	1 KB	-74.8%	1 KB
text/richtext	16 KB	67 KB	313.1%	68 KB

Table 2: Average size of the media types in textual03 and textual08, considering different maximum size constraints.

Media type	% contents allmedia05	% contents allmedia08	Trend
text/html	61.2%	57.8%	-5.5%
image/jpeg	22.6%	22.8%	1.2%
image/gif	11.4%	9.4%	-17.4%
app'n/pdf	1.6%	1.9%	18.5%
text/plain	0.7%	1.0%	76.1%
app'n/x-shockwave-flash	0.4%	0.7%	75.3%
app'n/octet-stream	0.1%	0.1%	49.6%
app'n/x-tar	0.1%	0.0%	-33.0%
app'n/x-zip-compressed	0.1%	0.0%	-32.8%
audio/mpeg	0.0%	0.1%	25.1%

Table 3: Top 10 media types by number of downloaded contents in allmedia05 and comparison to allmedia08.

Media type	% contents textual03	% contents textual08	Trend
text/html	95.9702%	93.9178%	-2.1%
app'n/pdf	1.9208%	3.0274%	57.6%
text/plain	1.0229%	1.6207%	58.5%
app'n/x-shockwave-flash	0.5440%	1.1737%	115.8%
app'n/msword	0.4332%	0.1803%	-58.4%
powerpoint	0.0644%	0.0299%	-53.6%
excel	0.0283%	0.0438%	55.0%
text/rtf	0.0069%	0.0010%	-85.2%
app'n/rtf	0.0060%	0.0024%	-59.5%
app'n/x-tex	0.0020%	0.0021%	2.5%
text/tab-separated-values	0.0013%	0.0007%	-45.3%
text/richtext	0.0001%	0.0000%	-40.7%

Table 4: Prevalence of media types in textual03 and textual08.

having sizes below 16 KB and an increase above 32 KB. According to the model provided by Gomes [10], the expected average size for 2008 was 40.3 KB, which represents an error of 23.3%. The obtained results show that, in general, the size of textual contents tends to increase.

The distribution obtained for allmedia08 is more spread across content size values than for textual contents. This proves that content size distribution for textual contents is not representative of the information generally available on the Web.

4.2.2 Prevalence

There are hundreds of formats for digital contents and they all can be potentially published on the Web. However, only some formats are commonly used due to their characteristics, such as size or portability. Hence, it is interesting to identify the trends in the evolution of the mostly used media types. This is useful, for instance, to select software format interpreters for the most common media types to include in mobile phone browsers that have limited capacities in comparison to desktop computers. Regarding media type prevalence, it is interesting to analyze both the number of contents and the amount of data provided due to the large differences between content sizes according to media type.

Table 3 presents the most prevalent media types in allmedia05 and the comparison to allmedia08. Though being the most common, there is a slight decrease in the prevalence of *text/html* type, representing 57.8% of the downloaded contents in allmedia08. In this crawl, the *text/html*, *image/jpeg* and *image/gif* represent 90.1% of the total number of downloaded contents. The presence of audio contents on the Web is very small but it has increased.

Table 4 presents the most downloaded media types, considering only textual contents. After 5 years, HTML is still dominant but lost presence to other formats. There is a growth of PDF and Flash. However, the ranking order is the same. The Microsoft Office formats (Word, Powerpoint, Excel, RTF) are prevalent among computer desktops. However, their presence is insignificant on the Web and except for Excel, tends to decrease.

Table 3 and Table 4 show that HTML is the dominant hypertextual format on the Web. However, although still presenting a discreet presence, the PDF and Flash formats, that were not mainly designed to support hypertexts but were enhanced with hypertextual features, tend to gain popularity. Despite this increase, Nielsen finds this format unsuitable for online presentation, since, though being good for printing, PDF presents usability problems in online interfaces [22]. For the media types present in both tables, the trends are consistent.

Table 5 presents the comparison of the prevalence of media types measured by amount of data between allmedia05 and allmedia08. In allmedia08, *text/html*, *application/pdf* and *image/jpeg* represent 69.4% of the total size. There is a decrease in *html* and *jpeg*, and a growth in *pdf*.

The Web servers visited in the allmedia08 crawl returned 637 distinct media types. However, those commonly used on the Web are generally restricted to a small subset: *html* for hypertext, *jpeg* and *gif* for images, *pdf* for documents, *flash* for animations, *tar* and *zip* for compressed files and *mpeg* for audio.

4.3 Dynamically generated contents

Media type	% data allmedia05	% data allmedia08	Trend
text/html	42.9%	35.4%	-17.3%
image/jpeg	21.0%	16.1%	-23.3%
app'n/pdf	14.8%	17.9%	20.4%
app'n/x-tar	3.6%	1.2%	-65.9%
image/gif	3.0%	1.6%	-46.4%
text/plain	2.1%	4.2%	98.8%
audio/mpeg	1.6%	2.7%	65.6%
app'n/x-shockwave-flash	1.2%	2.1%	78.2%
app'n/x-zip-compressed	1.1%	1.0%	-13.1%
app'n/octet-stream	1.0%	2.3%	125.6%

Table 5: Top 10 media types measured by amount of data in allmedia05 and comparison to allmedia08.

There are contents that do not exist physically on disk but that are dynamically generated on-the-fly when the Web server receives a request. Distinguishing dynamically generated from static contents is not straightforward [6]. Analyzing the presence of dynamically generated contents is interesting to identify technological trends in Web publishing. In allmedia08, the analysis of URLs to identify the presence of dynamically generated contents followed two approaches: embedded parameters and extension analysis. The former is based on the existence of a question mark in the URL. The latter is based on the analysis of known extensions for dynamically generated content technology (*php*, *asp*, *jsp*, *cfm*, *cgi*).

The percentage of URLs containing parameters raised from 47.2% in textual03 to 63.3% in textual08. The number of URLs containing embedded parameters in allmedia08 was 44.4% of the total number of contents.

We have not found previous results obtained through extension analysis for the Portuguese Web. However, Baeza et al. showed significant differences between national webs regarding the technologies used to publish dynamically generated contents [1]. Therefore, the obtained results for this metric might be peculiar to each national Web according to the influence that technologies have on each market.

The obtained results show a clear trend towards the usage of dynamically generated contents for Web publishing, using especially PHP technology. The widespread popularity of open-source free content management systems is a strong reason for this fact.

There are ranking algorithms that favor static contents. Since the usage of dynamically generated contents is increasing, this criterion loses relevance.

4.4 Duplication

Despite the hypertextual capacities of the Web to reference and reuse contents without performing physical duplication, the contents available on the Web are not unique. Duplicates occur when the same content is referenced by several distinct URLs and may comprise:

Contents repeated in different directories of a site. This happens, for instance, when contents are copied, rather than moved, and the original location is not deleted;

Contents physically duplicated in different sites. This happens, for instance, when images are replicated from site to site, rather than referencing the original location, or with uncustomized default files automatically generated by content management systems;

Complete mirrored sites. This is the case of software repositories as happens, for instance, with linux distribution mirrors.

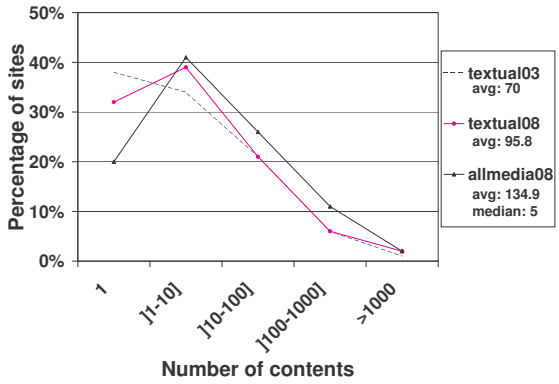

Figure 3: Distribution of the number of contents per site for textual03, textual08 and allmedia08.

During the crawl of allmedia08, a SHA1 digest was generated for each content and recorded in the crawl log. SHA1 (Secure Hash Algorithm 1) is used to compute a short representation for an input data sequence [16]. This digest was used to measure content duplication. Measuring duplication is useful, for instance, to help choosing adequate storage systems according to their duplicates elimination features. In allmedia08, approximately 48.7 million downloaded contents were crawled for 40 million different digests, which means that 17.7% of the downloaded contents were duplicates, representing 15.2% of the total amount of data downloaded. However, the level of duplication within textual08 decreases to 13.1%, which suggests that non-textual contents are more prone to be duplicated. The level of duplication found in textual03 was 15.5%.

5. SITES

This Section presents the properties analyzed regarding sites, such as number of contents per site, site size and site distribution per IP address.

5.1 Site size

The number of contents per site influences the crawler's data partitioning of the queues.

Figure 3 presents the distribution of contents crawled per site for textual03, textual08 and allmedia08. The inclusion of more media types in allmedia08 than in textual08 causes the sites to become larger. However, the distributions obtained for textual08 and allmedia08 are similar, except for a stronger presence of sites containing a single content among textual contents. One reason for this fact is that the single content of these sites is typically an HTML page.

A comparison between textual03 and textual08 shows that site size tends to increase with time. The main differences were found on the sites containing just 1 content, in which it decreased, and on the sites containing between 1 and 10 contents, in which it increased. The values obtained for sites larger than 10 contents remained similar.

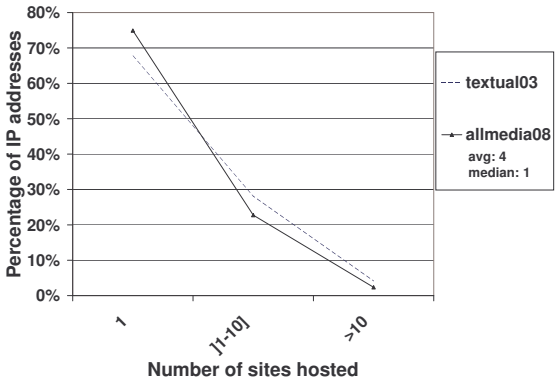

Figure 4: Distribution of sites hosted per IP address in textual03 and allmedia08.

5.2 Sites hosted per IP address

Virtual hosts enable a single Web server to host several sites. Measuring the distribution of sites across IP addresses is useful to define politeness policies for crawling. For instance, the crawler may be set to respect a courtesy pause between requests to the same IP address or to the same site, in order to avoid server overload.

Figure 4 presents the distribution of sites hosted per IP address in textual03 and allmedia08. The distributions obtained for textual03 and allmedia08 are similar. However, there is a slight increase in the number of IP addresses that host only one site against the remaining categories. The obtained results show that, in general, crawling courtesy pauses based on site name are adequate because most servers host a single site.

6. CONCLUSIONS AND FUTURE WORK

This study presented evolution trends by comparing an updated Web characterization with previous ones, for content and site characteristics. Several adjustments had to be made to avoid deriving trends biased by methodological differences. The Web portion used as case study to perform this analysis was the Portuguese Web. Although in some cases the obtained results might peculiar to this national Web, such as dominance of technology used to dynamically generate contents, we believe that, in general, they represent the global Web.

The absolute values for content characteristics tend to increase at different paces. After 5 years, the URL length increased slightly but the average content size presented significant differences. The most prevalent media types tend to define the general distributions but each media type presents peculiar characteristics and trends. For instance, the general trend is that content size tends to increase. However, the obtained results showed that the size for some media type contents is decreasing. This clearly shows that, contrary to common belief, sizes do not grow for all media types. This study also validates the extent to which contents have increased or decreased.

Web's evolution is not meeting the evolution of mobile Web and the Web of countries in development, since they do not provide the growing bandwidth contents demand every day.

The number of contents hosted per site tends to increase but there is a significant percentage of sites that provide a large number of unsuccessful responses. The usage of virtual hosts to support several sites on the same server maintained stable.

Future work will involve analyzing several crawls performed across longer periods of time for the Portuguese Web Archive which will enable to derive trends more accurately.

7. REFERENCES

[1] R. Baeza-Yates, C. Castillo, and E. Efthimiadis. Characterization of national web domains. *ACM Transactions on Internet Technology*, 7(2), 2007.

[2] R. Baeza-Yates, C. Castillo, and E. N. Efthimiadis. Comparing the Characteristics of the Chilean and the Greek Web, 2004.

[3] R. Baeza-Yates, C. Castillo, and V. López. Characteristics of the web of Spain. *Cybermetrics - International Journal of Scientometrics, Informetrics and Bibliometrics*, 9(1), 2005.

[4] R. Baeza-Yates, F. Lalanne, C. Castillo, and G. Dupret. Comparing the characteristics of the Korean and the Chilean Web. Technical report, Korea-Chile IT Cooperation Center ITCC, 2004.

[5] 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.

[6] C. Castillo. *Effective Web Crawling*. PhD thesis, University of Chile, November 2004.

[7] M. C. Drott. Indexing aids at corporate websites: the use of robots.txt and meta tags. *Inf. Process. Manage.*, 38(2):209–219, 2002.

[8] R. Fagin, R. Kumar, K. Mccurley, J. Novak, D. Sivakumar, J. Tomlin, and D. Williamson. Searching the workplace web, 2003.

[9] Funredes and U. Latine. Langues et cultures sur la toile. `http://dtil.unilat.org/LI/2007/index_fr.htm`, 2007.

[10] D. Gomes. *Web Modelling for Web Warehouse Design*. Phd thesis, University of Lisbon, November 2006.

[11] D. Gomes, S. Freitas, and M. J. Silva. Design and selection criteria for a national web archive. In *ECDL 2006 - 10th European Conference on Research and Advanced Technology for Digital Libraries*, number 4172/2006 in LNCS, pages 196–207. Springer-Verlag, September 2006.

[12] D. Gomes, A. Nogueira, J. Miranda, and M. Costa. Introducing the portuguese web archive initiative. In *8th International Web Archiving Workshop (IWAW08)*, Aarhus, Denmark, September 2008.

[13] D. Gomes and M. J. Silva. Characterizing a national community web. *ACM Transactions on Internet Technology*, 5(3):508–531, 2005.

[14] D. Gomes and M. J. Silva. The viúva negra crawler: an experience report. *Softw. Pract. Exper.*, 38(2):161–188, 2008.

[15] IEEE. IEEE trial-use standard for prefixes for binary multiples. *IEEE Std 1541-2002*, pages 0_1–4, 2003.

[16] N. institute˜of˜standards˜and technology. Fips 180-2, secure hash standard, federal information processing

standard (fips), publication 180-2. Technical report, DEPARTMENT OF COMMERCE, August 2002.

[17] F. Lasfargues, C. Oury, and B. Wendland. Legal deposit of the french web: harvesting strategies for a national domain. In *8th International Web Archiving Workshop (IWAW08)*, Aarhus, Denmark, September 2008.

[18] M. Modesto, Álvaro R. Pereira Jr., N. Ziviani, C. Castillo, and R. Baeza-Yales. Um novo retrato da web brasileira. In *XXXII SEMISH - Anais do Seminário Integrado de Software e Hardware*, pages 2005–2017, São Leopoldo, RS, July 2005.

[19] G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic. Introduction to heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04)*, Bath, UK, September 2004.

[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] J. Nielsen. Pdf: Unfit for Human Consumption. http://www.useit.com/alertbox/20030714.html, 2008.

[23] N. Noronha, J. P. Campos, D. Gomes, M. J. Silva, and J. Borbinha. A deposit for digital collections. In P. Constantopoulos and I. T. Sølvberg, editors, *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, ECDL*, volume 2163 of *LNCS*, pages 200–212. Springer, 2001.

[24] OCLC. Web characterization. `http://wcp.oclc.org/`, 2003.

[25] E. T. O'Neill, B. F. Lavoie, and R. Bennett. How "world wide" is the web?: 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] G. Tolosa, F. Bordignon, R. Baeza-Yates, and C. Castillo. Characterization of the argentinian web. *Cybermetrics*, 11(1):3+, July 2007.

[28] W3C. Web characterization activity statement. `http://www.w3.org/WCA/Activity.html`, 1999.

[29] P. Zabicka. Czech Web archive analysis. *New Review of Hypermedia and Multimedia*, 13(1):27–37, 2007.