Trends in Web characteristics

João Miranda and Daniel Gomes
Foundation for National Scientific Computing
Lisbon, Portugal
{*joao.miranda,daniel.gomes*}*@fccn.pt*

***Abstract*—The Web is permanently changing, with new technologies and publishing behaviors emerging everyday. It is important to track trends on the evolution of the Web to develop efficient tools to process its data. For instance, Web trends influence the design of browsers, crawlers and search engines. This study presents trends on the evolution of the Web derived from the analysis of 3 characterizations performed within an interval of 5 years. The Web portion used as a case study was the Portuguese Web. Several metrics regarding site and content characteristics were analyzed.**

***Keywords*-Web trends; Web measurements; Web characterization**

I. Introduction

In the 90's, only computer experts were able to publish a Web page. In 2009, any common Internet user with few technical skills can become a massive Web publisher, using free tools and services, such as blog platforms, wikis or content management systems. Thus, 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.

Web characterization is a research field that contributes to model its characteristics across time [1]. 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, that can be defined using distinct methodologies. However, it is important to maintain a similar methodology across time to derive evolution trends. For instance, a Web characterization extracted from a university proxy log should not be compared to one gathered from Web crawling. The proxy log characterization reflects a portion composed exclusively by the contents accessed by the proxy users, while a crawl contains a broad scope of the information available on the Web.

This study compares the obtained results from 3 characterizations of a Web portion performed in 2003 [2], 2005 [3] and 2008 [4] to derive evolution trends. The Web portion used as a case study was the Portuguese Web. 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 [5]. 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. Therefore, we believe that the results obtained from the Portuguese Web for the presented metrics reflect the trends of the global Web. Plus, the Portuguese Web was thoroughly characterized in the past using similar methodologies, which enables accurate comparisons. The main contribution of this study is a quantitative analysis of the trends of Web characteristics. Although users may empirically witness some Web evolution phenomena, such as the growth of page sizes, this paper provides an original contribution by statistically measuring evolution trends. Those results enable deriving evolution scenarios that influence the development of applications that process Web data, such as browsers or search engines.

This paper is organized as follows. Section II presents related work, Section III describes the methodology adopted and Section IV presents trends on content and site characteristics. Finally, Section V summarizes the main conclusions and proposes future work.

II. Related work

Web characterization has been a subject of several studies. Pitkow presented a summary of the first efforts to characterize the Web [6]. Najork and Heydon performed a large scale crawl producing several statistics [7]. Boldi et al. presented structural information on the African Web, including structure of pages and most used technologies [8]. 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 [9], [10]. They also published an in-depth study of the Spanish Web [11] 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 [5]. Tolosa et al. presented the characteristics of the Argentinian Web from a crawl of over 10 million pages from 150 000 sites performed in 2006 [12]. Zabicka and Matejka analyzed the Czech Web archive, performing a characterization of the archived contents [13].

Regarding the study of the evolution of the Web across time, the Web Characterization Project analyzed the trends in the size and content of the Web until 2002 [14]. Modesto et al. characterized the evolution of the Brazilian Web between 2000 and 2005 [15]. 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 [16]. Funredes and Union Latine have been studying the presence of languages and cultures on the Web since 1996 [17]. 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 [18].

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 [19]. Noronha et al. presented a system for supporting the archive of Web publications in a digital library [20]. They performed a crawl of selected publications and characterized the obtained collection. The previous characterizations of the Portuguese Web analyzed in this study to derive trends on Web characteristics were published by Gomes et al. [2], [3], [4].

III. 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 *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 (200 response code). The amount of information published is expressed in decimal multiples: 1 KB = 10^3 bytes [21].

The Portuguese Web Archive (PWA) project aims to automatically gather and preserve the information published on the Portuguese Web [22]. The most recent Web characterization results analyzed 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** [4]. There were two previous studies that will be used as baseline to derive evolution trends. The first study presented a thorough characterization of the Portuguese Web derived from a crawl of textual contents performed in 2003 [3], which we will henceforth refer to as **textual03**. The second study presented the most prevalent media types on the Portuguese Web, extracted from a crawl performed in 2005 [2], which we named **allmedia05**. 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.

Table I
2003-2008: comparison of the most common response codes received from Web servers.

Status Code	% textual03	% allmedia08	Trend	Description
200	88.2%	85.2%	-3.4%	OK
302	5.3%	7.2%	+35.9%	Found
404	3.6%	5.1%	+39.5%	Not Found
301	1.1%	1.3%	+18.2%	Moved Permanently
500	0.9%	0.2%	-78.9%	Internal Server Error
403	0.5%	0.2%	-59.6%	Forbidden
401	0.1%	0.2%	+47.4%	Unauthorized
400	0.1%	0.2%	+204.5%	Bad Request
503	0.0%	0.1%	+454.8%	Service Unavailable

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 [23]. 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 size limit of the downloaded contents was 2 MB.

The allmedia08 was crawled to feed a Web archive using Heritrix 1.12.1 [24] 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 content 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.

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

IV. Trends

This Section presents the trends on Web content characteristics that can be used, for instance, to enhance browsers and crawlers. In allmedia08, the number of contents excluded due to Robots Exclusion Protocol (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 [25] from the Web pages code. This is a reason for the increase of the REP exclusions when compared to textual03.

A. HTTP responses

Table I presents a comparison of the response status codes logged in textual03 and allmedia08. 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. Those that suffered the highest relative changes were 400 (Bad Request) and 503 (Service Unavailable) responses. However these response codes are extremely rare, both representing less than 0.2% of all the response codes.

The significant growth of 404 (Not Found) errors suggests that webmasters are becoming less careful on maintaining their pages. On the other hand, the total number of 500 (Internal Server Error) and 503 errors decreased, which suggests that Web servers online offer better availability for their services. A reason for this fact is that, although users still independently maintain their contents, with the widespread usage of Web 2.0 platforms to support Web publishing the servers availability is usually high.

Web usability guidelines suggest that publicly available pages should not link to restricted access contents, such as intranets [26]. The main objective of this guideline is to prevent users frustration when they try to follow links that raised their interest but reference unavailable contents. The total number of forbidden and unauthorized errors decreased from 0.6% to 0,4% (-33%), which indicates that this guideline is being increasingly followed.

There was also a significant growth in the usage of redirects. They are used to prevent linkrot by transparently driving users to alternative locations of contents. On the other hand, there are also heavily used to track users behavior, for instance, in e-commerce sites.

B. URL length

The URL length of contents is a feature that can influence interaction design. It is useful to determine the adequate length for input text boxes that receive URLs (e.g. Internet Archive Wayback Machine [27]), or to determine how many characters of a URL should be presented on a search engine results page. The URL length is also used in search engine ranking algorithms to identify relevant results [28].

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 length of `www.a.com/b.php?f=2` was measured. Thus, this URL presents a length of 19 characters.

In allmedia08, the URLs returning OK responses presented a length ranging from 5 to 2 072 characters. Figure 1 depicts 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 [29], the expected URL length in 2008 would be

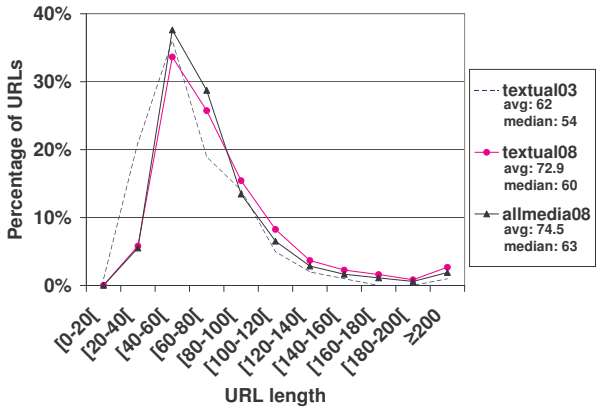

Figure 1. 2003-2008 textual and all media types: comparison of URL length distribution.

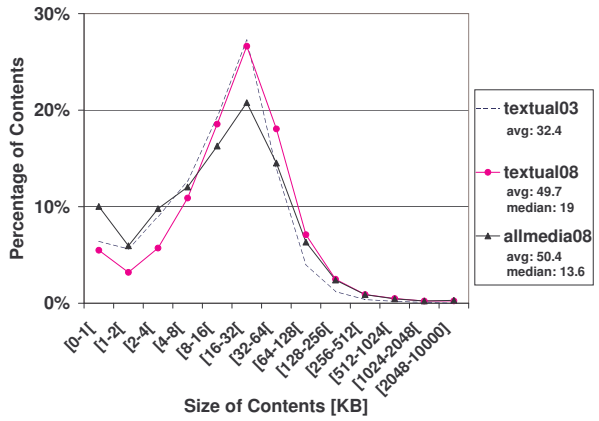

Figure 2. 2003-2008 textual and all media types: comparison of content size distribution in KB.

67.6 characters, which represents a difference 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.

C. Content size

Analyzing trends in content sizes is useful to estimate the storage resources required to create Web data repositories. However, sizes significantly vary according to the media types being stored. Therefore, we present trends for content size in general and per media type.

1) General trends: Figure 2 presents the general distribution of content size across time, including several media types. The imposed maximum content size limit of 10 MB in allmedia08 resulted in a total of 32 321 truncated contents, which represents only 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.

Table II
2003-2008 TEXTUAL MEDIA TYPES: COMPARISON OF THE AVERAGE SIZE, CONSIDERING DIFFERENT MAXIMUM SIZE CONSTRAINTS.

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

The distribution for textual contents is similar between textual03 and textual08 with approximately 73% of the contents having sizes between 4 KB and 64 KB. However, in the latter crawl there was a decrease in the number of contents having sizes below 16 KB and an increase above 32 KB. According to the model provided by Gomes [29], the expected average size in 2008 would be 40.3 KB. This estimation misses the measured average size by 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.

2) Media type trends: Table II 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 used then 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 4^{th} column of Table II 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 media types that suffered higher relative changes were the *text/plain*, *application/x-shockwave-flash*, *excel* and *text/richtext* types.

A comparison between the 3^{rd} and the 5^{th} columns of Table II revealed that the average size for certain media types was significantly affected by the imposition of different size limits. 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.

Table III
2005-2008 ALL MEDIA TYPES: COMPARISON OF THE MOST PREVALENT TYPES, MEASURED BY NUMBER OF DOWNLOADED CONTENTS.

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 IV
2003-2005 TEXTUAL MEDIA TYPES: COMPARISON OF THE MOST PREVALENT TYPES, MEASURED BY NUMBER OF DOWNLOADED CONTENTS.

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%

D. Media type prevalence

New formats appear everyday while others disappear. Although the relative presence of some media types is discreet on the Web, their evolution trends are important to design systems focused on processing specific media types, such as music or scientific publication repositories, fed from information gathered from the Web.

There are hundreds of formats for digital contents that can be potentially published on the Web. However, only some formats are commonly used due to their characteristics, such as size or portability, and it is interesting to identify trends in the evolution of media type prevalence. For instance, mobile phone browsers have limited capacities in comparison to desktop computers and they must include software only to interpret the most commonly used media types.

Table III compares the most prevalent media types, measured by number of contents, in allmedia05 and allmedia08. Though being the most common, there is a slight decrease in the prevalence of *text/html* from 61.2% in allmedia05 to 57.8% in allmedia08. In 2005, the *text/html*, *image/jpeg* and *image/gif* types represented 95.2% of the total number of downloaded contents. This value decreased to 90.1% in 2008, which suggests that media type prevalence tends to be more spread.

Table IV compares the most prevalent media types, measured by number of contents, in textual03 and textual08.

Table V
2005-2008 ALL MEDIA TYPES: COMPARISON OF THE MOST PREVALENT TYPES, MEASURED BY AMOUNT OF DATA.

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%

After 5 years, HTML is still dominant but lost presence to other formats. However, the ranking order is the same. There is a growth of PDF and Flash. The latter is the media type that suffered higher relative change. The Microsoft Office formats (Word, Powerpoint, Excel, RTF) are prevalent among computer desktops. However, their presence is very discreet on the Web and except for Excel, tends to decrease.

Table V compares the prevalence of media types, measured by amount of data, in allmedia05 and allmedia08. In allmedia08, *text/html*, *application/pdf* and *image/jpeg* represent 69.4% of the total size. There is a decrease in *text/html* and *image/jpeg*, and a growth in *application/pdf*. The media type that suffered higher relative change was the *application/octet-stream* type.

The servers visited in the allmedia08 crawl returned 637 distinct media types. However, those commonly used on the Web are 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.

Table III, Table IV and Table V show that HTML is the dominant hypertextual format on the Web. However, formats that were not mainly designed to support hypertexts but were enhanced with hypertextual features, such as PDF or Flash, tend to gain popularity. Controversially, those formats present usability problems and were considered to be unsuitable for online presentation [30], [31].

E. Dynamically generated contents

There are contents published on the Web that are not physically stored on disk. Instead, they are dynamically generated on-the-fly when the server receives a request. Analyzing the presence of dynamically generated contents is interesting to identify technological trends in Web publishing. However, distinguishing dynamically generated from static contents is not straightforward [32]. A possible approach to identify the presence of dynamically generated contents is based on the existence of a question mark in the URL. Nevertheless, this approach provides results on the minimal number of dynamic contents, since there are pages that do not contain any question mark on their URL but are dynamically generated (e.g. `http://site.com/index.php`).

The percentage of URLs containing parameters raised from 47.2% (textual03) to 63.3% (textual08). The obtained results show a clear trend towards the usage of dynamically generated contents for Web publishing. The widespread popularity of free open-source content management systems is a strong reason for this fact. The observed trend towards the widespread of dynamic pages shows that the crawling policies used in the early days of the Web that excluded these pages from crawls to prevent spider traps, should be abandoned [33].

F. Duplication

Despite the hypertextual capacities of the Web to reuse contents without physically duplicating them, contents are not unique. Duplicates occur when the same content is referenced by several distinct URLs and may comprise:

Contents repeated across directories within a site. 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 or content management systems default files are replicated across several sites;

Complete mirrored sites. This is the case of software repositories, for instance, Linux distribution mirrors.

SHA1 (Secure Hash Algorithm 1) is used to compute a short representation of a data sequence [34]. During the crawl of allmedia08, a SHA1 digest was generated for each content and recorded in the crawl log. This digest was used to measure content duplication. Measuring duplication is useful, for instance, to choose adequate storage systems according to their duplicates elimination features [35]. The level of duplication found in textual03 was 15.5% and decreased to 13.1% in textual08.

G. Site size

The number of contents per site influences the definition of crawling strategies [32]. The distribution of contents per site is important to efficiently partition a large data set of URLs across several crawling processes.

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 caused the sites to become larger. However, the distributions obtained for textual08 and allmedia08 are similar, except for a stronger relative presence of sites containing a single content among textual contents.

A comparison between textual03 and textual08 shows that site size tends to increase. The main differences were found on the sites containing only 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. In 2003 it was observed that

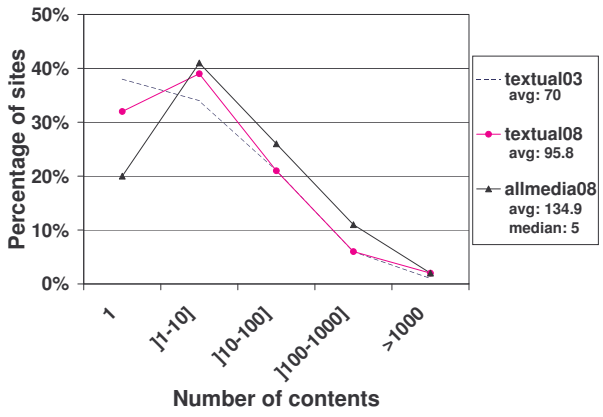

Figure 3. 2003-2008 textual and all media types: comparison of the number of contents downloaded per site.

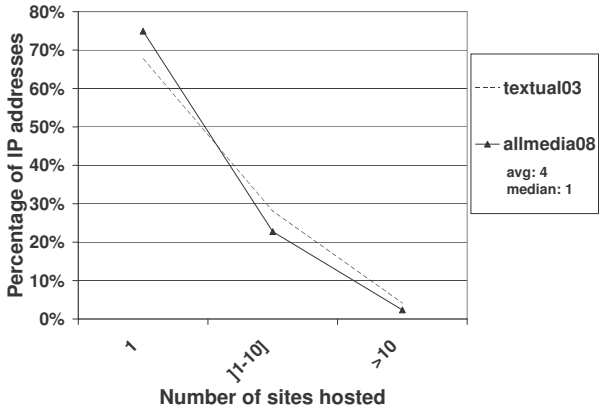

Figure 4. 2003-2008: comparison of the distribution of sites hosted per IP address.

most of the single page sites were under construction or abandoned. The reduction of the number of this type of sites is a positive indicator regarding the quality of the sites available online.

H. Sites hosted per IP address

Virtual hosts enable a single Web server to host several sites [36]. Measuring the distribution of sites across IP addresses is useful to define politeness policies for crawling. For instance, a crawler must respect a courtesy pause between requests to avoid server overload. However, for the servers that host many sites, respecting a courtesy pause to the same site might not be enough to prevent overload.

Figure 4 presents the distribution of sites hosted per IP address in textual03 and allmedia08. The distributions obtained 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.

V. Conclusions and Future Work

This study presented Web evolution trends by comparing characterizations of the Portuguese Web performed in 2003, 2005 and 2008. Although in some cases the obtained trends may be peculiar to the Portuguese Web, we believe that, in general, they represent the global Web because a national Web is composed by a broad set of sites with distinct publication scopes. The analysis of the evolution of the Web has impact on software development in several fields, such as Web design or Web search.

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 increased significantly. The most prevalent media types tend to define the general feature distributions but each media type presents peculiar trends. For instance, the general trend is that content size tends to increase. However, the obtained results showed that the content size for some media types is decreasing. This shows that, contrary to common belief, sizes do not grow for all media types. The use of dynamically generated contents is wide-spreading and it already represents more than half of the pages available online. The number of contents hosted per site tends to increase. The usage of virtual hosts to support several sites on the same server maintained stable.

The obtained results suggest that sites tend to be more available, although the presence of broken links on their pages is growing. Software to process Web data must be carefully design according to the peculiar characteristics of the media types it is going to address. Media type formats such as HTML, GIF or JPEG became standards on the Web. However, the growing prevalence of formats not originally designed to be accessed online, such as PDF, might raise usability problems. The growing prevalence of proprietary formats, such as Flash, raises barriers to Web data access and preservation.

Future work will involve analyzing the evolution of metrics related to Web quality. The emergence of new technologies and widespread of broadband connections are strong contributors to the continuous popularity growth of the Web. However, these factors do not imply that the Web user experience is getting significantly better. It would be interesting to measure the evolution of Web usability and accessibility across time.

Acknowledgments

This work was co-funded by POSC/EU.

References

[1] W3C, “Web characterization activity statement,” http://www.w3.org/WCA/Activity.html, 1999.

[2] 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*, ser. LNCS, no. 4172/2006. Springer-Verlag, September 2006, pp. 196–207. [Online]. Available: http://dx.doi.org/10.1007/11863878_17

[3] D. Gomes and M. J. Silva, "Characterizing a national community web," *ACM Transactions on Internet Technology*, vol. 5, no. 3, pp. 508–531, 2005.

[4] J. Miranda and D. Gomes, "An Updated Portrait of the Portuguese Web," in *14th Portuguese Conference on Artificial Intelligence (EPIA 2009)*, Aveiro, Portugal, October 2009. [Online]. Available: http://arquivo-web.fccn.pt/sobre-o-arquivo/an-updated-portrait-of-the-portuguese-web

[5] R. Baeza-Yates, C. Castillo, and E. Efthimiadis, "Characterization of national web domains," *ACM Transactions on Internet Technology*, vol. 7, no. 2, 2007. [Online]. Available: http://www.chato.cl/research/

[6] J. E. Pitkow, "Summary of WWW characterizations," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 551–558, 1998. [Online]. Available: http://citeseer.nj.nec.com/article/james98summary.html

[7] M. Najork and A. Heydon, "On high-performance web crawling," Compaq Systems Research Center, SRC Research Report, 2001. [Online]. Available: http://www.hpl.hp.com/techreports/Compaq-DEC/SRC-RR-173.html

[8] 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. [Online]. Available: http://www2002.org/CDROM/poster/164/

[9] R. Baeza-Yates, C. Castillo, and E. N. Efthimiadis, "Comparing the Characteristics of the Chilean and the Greek Web," 2004. [Online]. Available: http://www.chato.cl/papers/baeza04_comparing_chilean_web_greek_web.pdf

[10] R. Baeza-Yates, F. Lalanne, C. Castillo, and G. Dupret, "Comparing the characteristics of the Korean and the Chilean Web," Korea-Chile IT Cooperation Center ITCC, Technical report, 2004. [Online]. Available: http://www.chato.cl/papers/baeza_04_comparing_chilean_web_korean_web.pdf

[11] R. Baeza-Yates, C. Castillo, and V. López, "Characteristics of the web of Spain," *Cybermetrics - International Journal of Scientometrics, Informetrics and Bibliometrics*, vol. 9, no. 1, 2005. [Online]. Available: http://www.cindoc.csic.es/cybermetrics/articles/v9i1p3.html

[12] G. Tolosa, F. Bordignon, R. Baeza-Yates, and C. Castillo, "Characterization of the Argentinian Web," *Cybermetrics*, vol. 11, no. 1, pp. 3+, July 2007. [Online]. Available: http://www.cindoc.csic.es/cybermetrics/articles/v11i1p3.html

[13] P. Zabicka and L. Matejka, "Czech Web archive analysis," *New Rev. Hypermedia Multimedia*, vol. 13, no. 1, pp. 27–37, 2007.

[14] OCLC, "Web characterization," 2003. [Online]. Available: http://wcp.oclc.org/

[15] 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*, São Leopoldo, RS, July 2005, pp. 2005–2017. [Online]. Available: http://bibliotecadigital.sbc.org.br/download.php?paper=178

[16] 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*, vol. 9, no. 4, April 2003. [Online]. Available: http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

[17] Funredes and U. Latine, "Langues et cultures sur la Toile," 2007. [Online]. Available: http://dtil.unilat.org/LI/2007/index_fr.htm

[18] 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. [Online]. Available: http://iwaw.net/08/IWAW2008-Lasfargues.pdf

[19] M. J. Nicolau, J. Macedo, and A. Costa, "Caracterização da informação WWW na RCCN," Universidade do Minho, Tech. Rep., 1997. [Online]. Available: http://marco.uminho.pt/~macedo/netcensus/

[20] N. Noronha, J. P. Campos, D. Gomes, M. J. Silva, and J. Borbinha, "A deposit for digital collections," in *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, ECDL*, ser. LNCS, P. Constantopoulos and I. T. Sølvberg, Eds., vol. 2163. Springer, 2001, pp. 200–212. [Online]. Available: http://xldb.di.fc.ul.pt/referencias/paper_ecdl01.pdf

[21] IEEE, "IEEE Trial-Use Standard for Prefixes for Binary Multiples," *IEEE Std 1541-2002*, pp. 0_1–4, 2003.

[22] 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. [Online]. Available: http://iwaw.net/08/IWAW2008-Gomes.pdf

[23] D. Gomes and M. J. Silva, "The Viúva Negra crawler: an experience report," *Softw. Pract. Exper.*, vol. 38, no. 2, pp. 161–188, 2008.

[24] 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. [Online]. Available: http://www.iwaw.net/04/proceedings.php?f=Mohr

[25] M. C. Drott, "Indexing aids at corporate websites: the use of robots.txt and meta tags," *Inf. Process. Manage.*, vol. 38, no. 2, pp. 209–219, 2002.

[26] J. Nielsen, "Fighting Linkrot," 1998. [Online]. Available: http://www.useit.com/alertbox/980614.html

[27] "Internet Archive Wayback Machine." [Online]. Available: http://web.archive.org/collections/web.html

[28] R. Fagin, R. Kumar, K. Mccurley, J. Novak, D. Sivakumar, J. Tomlin, and D. Williamson, "Searching the workplace web," 2003. [Online]. Available: http://citeseer.ist.psu.edu/fagin03searching.html

[29] D. Gomes, "Web Modelling for Web Warehouse Design," Ph.D. dissertation, University of Lisbon, March 2007. [Online]. Available: http://xldb.fc.ul.pt/daniel/docs/papers/thesisDcgomes.pdf

[30] J. Nielsen, "Flash: 99% Bad," October 2000. [Online]. Available: http://www.useit.com/alertbox/20001029.html

[31] ——, "Pdf: Unfit for Human Consumption," 2008. [Online]. Available: http://www.useit.com/alertbox/20030714.html

[32] C. Castillo, "Effective web crawling," Ph.D. dissertation, University of Chile, November 2004. [Online]. Available: http://www.dcc.uchile.cl/~ccastill/crawling_thesis/effective_web_crawling.pdf

[33] J. Cho, H. Garcia-Molina, and L. Page, "Efficient crawling through URL ordering," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 161–172, 1998. [Online]. Available: http://citeseer.ist.psu.edu/article/cho98efficient.html

[34] National Institute of Standards and Technology, "FIPS 180-2, Secure Hash Standard, Federal Information Processing Standard (FIPS), Publication 180-2," Department of Commerce, Tech. Rep., August 2002. [Online]. Available: http://csrc.nist.gov/publications/fips/fips180-2/fips180-2withchangenotice.pdf

[35] D. Gomes, A. L. Santos, and M. J. Silva, "Webstore: A manager for incremental storage of contents," Department of Informatics, University of Lisbon, DI/FCUL TR 04–15, November 2004. [Online]. Available: http://www.di.fc.ul.pt/tech-reports/04-15.pdf

[36] The Apache Software Foundation, "Apache Virtual Host documentation." [Online]. Available: http://httpd.apache.org/docs/1.3/vhosts/