How Are Web Characteristics Evolving?

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 in permanent evolution. There are new technologies and Web publishing behaviors emerging everyday. This study presents trends on the evolution of the Web, derived from the comparison of two characterizations of a web portion performed within a 5 year interval. The Portuguese Web was used as a case study. Several metrics regarding content and site characteristics were analyzed.

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

Experimentation, Measurement

Keywords

Web trends, Web characterization, Web measurements

1. INTRODUCTION

The Web is prone to suffer significant changes on its characteristics with a short notice, affecting, for instance, the media types most commonly used for publication. It is important to keep track of trends on the evolution of the Web to develop efficient tools for processing its data. 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.

This study presents a new characterization of a Web portion derived in 2008, presenting measurements for metrics that were not studied in previous works and that can be used as baseline for future trend analysis. It also compares the obtained results with previous studies 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., there are characteristics shared across countries and valid for the global Web, such as URL length or HTTP responses distributions [1]. Thus, we believe that the measurements obtained for the Portuguese Web reflect the trends of the global Web.

2. METHODOLOGY

The new Web characterization results presented in this study were extracted from a crawl performed in 2008 by the Portuguese Web Archive [4], including all media types. We named this crawl **allmedia08** and used two previous studies as baseline to derive trends. The first study presented a thorough characterization of the Portuguese Web derived from a crawl of 3.2 million textual contents performed in 2003 to feed a search engine [5], which we named **textual03**. The second study presented the most prevalent media types on the Portuguese Web, based on a crawl from 2005 [3], which we named **allmedia05**. When comparing results from allmedia08 with textual03, we considered only the subset of textual media types harvested in both crawls. We named as **textual08** this subset of contents from allmedia08. The characteristics obtained from allmedia05 and allmedia08 were compared directly.

3. CONTENTS

The number of contents downloaded was 48 718 404 in a total amount of 2.5 TB of data. The URL length of contents is a feature used in search engine ranking algorithms to identify relevant results [2]. After 5 years, the median URL length increased from 54 to 60 characters and the average from 62 to 72.9.

Analyzing trends on content sizes is useful to estimate the resources required to create Web data repositories. Size distribution for textual contents is similar between textual03 and textual08. After 5 years, the average size for *text/html* pages grew from 21 KB to 30 KB and except for *powerpoint*, *text/rtf* and *text/tab-separated-values*, the content size for all media types tends to grow.

New hypertextual formats appear everyday and others evolve to include hypertextual features. Identifying trends on the evolution of the most used media types is useful, for instance, to select software format interpreters to include in mobile phone browsers that have fewer resources than desktop computers. When comparing allmedia05 to allmedia08, the obtained results show a slight decrease in the prevalence of *text/html* (61.2% to 57.8%). Although still presenting a relatively discreet presence, the PDF and Flash formats tend to gain popularity. PDF prevalence increased from 1.6% to 1.9% and Flash prevalence increased from 0.4% to 0.7%. The

HT'09, June 29–July 1, 2009, Torino, Italy.
ACM 978-1-60558-486-7/09/06.

Media type	% contents textual03	% contents textual08	Trend
text/html	95.9702%	93.9178%	↓
app'n/pdf	1.9208%	3.0274%	↑
text/plain	1.0229%	1.6207%	↑
app'n/x-shockwave-flash	0.5440%	1.1737%	↑
app'n/msword	0.4332%	0.1803%	↓
powerpoint	0.0644%	0.0299%	↓
excel	0.0283%	0.0438%	↑
text/rtf	0.0069%	0.0010%	↓
app'n/rtf	0.0060%	0.0024%	↓
app'n/x-tex	0.0020%	0.0021%	↑
text/tab-separated-val's	0.0013%	0.0007%	↓
text/richtext	0.0001%	0.0000%	↓

Table 1: Prevalence of media types and trend in textual03 and textual08.

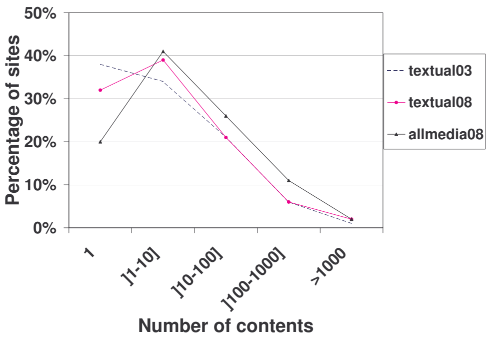

Figure 1: Distribution of the number of contents per site for textual03, textual08 and allmedia08.

same trend stands when comparing textual03 and textual08, as can be seen in Table 1.

Measuring the duplication of contents influences the choice of storage systems according to their duplicates elimination features. During the crawl of allmedia08, a SHA1 digest was generated for each content. Approximately 48.7 million contents were crawled for 40 million distinct digests. This means that 17.7% of the contents were duplicates, representing 15.2% of the total amount of data downloaded. The obtained results show that most contents are unique (92.8%) and that most duplicated contents occur twice (5.1%). The *text/html* type is responsible for 38.1% of the duplicates. The CSS and JavaScript contents are commonly duplicated instead of being reused, presenting 57.9% and 38.1% of duplicates caused by duplication within the same site, which inhibits the advantages of sharing files from these media types across pages [6].

4. SITES

A site was considered valid if it returned a 200 response code to at least one request. In allmedia08, the total number of sites visited was 484 398 and 74.6% of them were valid.

Figure 1 presents the distribution of contents per site for textual03, textual08 and allmedia08. A site has on average 134.9 contents, with a median of 5 contents. Sites are typically small, 87% presented less than 100 contents. After 5 years, the average size increased from 70 to 95.8 contents.

The percentage of successful responses returned by a site is an indicator of its quality. A site that presents a large percentage of broken links mines the trust of its users. On average, in allmedia08 each site returned OK responses (200 status code) to 75% of the requests. If every Successful (200 to 206) and Redirection (300 to 307) response codes are also considered as successful responses, this number increases to 82.4%. One may argue that larger sites are harder to maintain and should present a higher rate of broken links. However, the correlation factor found between site size and OK responses was 0.06, and 0.04 between site size and Successful and Redirection responses. This shows that there is no relation between site size and successful response percentage.

Measuring the distribution of sites across IP addresses is useful to define politeness policies for crawling: a crawler may be set to respect a courtesy pause between requests to the same IP address to avoid server overload. Regarding the distribution of sites hosted per IP address, the obtained results show that, on average, each IP address hosts 4 sites (median of 1). Only 2% of the IP addresses host more than 10 sites. The distributions for textual03 and allmedia08 are similar. However, there is a slight increase in the number of IP addresses that host only 1 site. The obtained results show that, in general, crawling courtesy pauses based on site name are adequate because most servers host a single site.

5. CONCLUSIONS

Content characteristics tend to evolve at different paces. After 5 years, the URL length increased slightly but the average content size presented significant differences. Most prevalent media types tend to determine the general trends but each type presents peculiar characteristics. For instance, the general trend is that content size tends to increase. However, the obtained results showed that size for some media contents is decreasing. A surprising result was that duplication is prevalent among certain types, such as CSS and JavaScript, which contradicts Web design best practices.

The number of contents hosted per site tends to increase but sites provide a large number of unsuccessful responses.

The crawl log of allmedia08 and the extended version of this study are available at `http://arquivo.pt/resources`.

6. 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. Fagin, R. Kumar, K. Mccurley, J. Novak, D. Sivakumar, J. Tomlin, and D. Williamson. Searching the workplace Web, 2003.

[3] 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.

[4] 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.

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

[6] S. Koyani, R. Bailey, and J. Nall. *Research-Based Web Design & Usability Guidelines*. Department of Health and Human Services, 2006.