Comparing the Characteristics of the Chilean and the Greek Web

Ricardo Baeza-Yates
CWR, Universidad de Chile
rbaeza@dcc.uchile.cl

Carlos Castillo
CWR, Universidad de Chile
ccastill@dcc.uchile.cl

Efthimis N. Efthimiadis
University of Washington
efthimis@u.washington.edu

ABSTRACT

This article summarizes the results of a comparison between the characteristics of two public Web spaces: the pages under the .GR (Greece) domain, and the pages under the .CL (Chile) domain. We show several similarities that contribute to validate more general models for the characteristics of the Web, specially in terms of link structure.

Categories and Subject Descriptors

H.3.5 [**Information Systems**]: On-line Information Systems

Keywords

Web characterization, link analysis

1. MAIN CHARACTERISTICS

The pages were obtained using the WIRE crawler [1] during January 2004. We downloaded pages using a breadth-first scheduler for up to 5 levels for dynamically generated pages, and up to 15 levels for static, HTML pages. We limited the crawler to 20,000 pages per website; and considered only pages under the `.gr` and `.cl` domains.

Table 1 summarizes information about the page collection, as well as some demographic facts that provide the context for this study.

	Greece	Chile
Population [6]	10.9 Million	15.2 Million
Gross Domestic Product [5]	133 US$ bn.	66 US$ bn.
Per-capita GDP, PPP [5]	17,697 US$	10,373 US$
Human development rank [7]	24^{th}	43^{th}
Web servers contacted	28,974	36,647
Pages downloaded	4.0 Million	2.7 Million
Pages with HTTP OK	77.8%	78.3%

Table 1: Summary of characteristics.

Figure 1 shows the depth at which the pages of the collection were found; note that 5 is the limit we set for dynamic pages, as dynamic pages grows exponentially with depth. Figure 2 shows the distribution of page sizes, showing a peak between 10 and 15 Kilobytes.

Figure 3 plots the number of pages per website. This has a very skewed distribution, as few websites account for a large portion of the total web; so we have plotted this in log-log scale.

Submitted for publication.
ACM xxx.xxx.

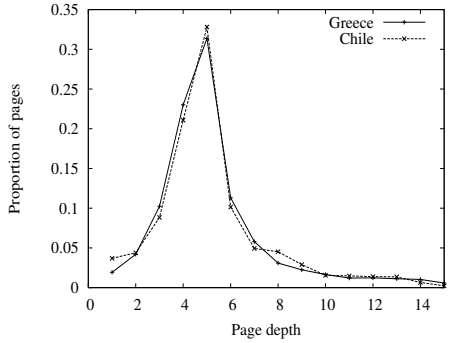

Figure 1: Page depth, 1 is the page at the root of the server.

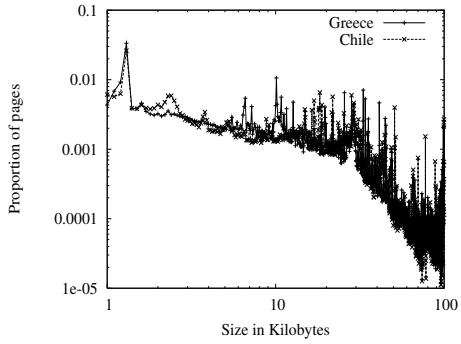

Figure 2: Page size in Kilobytes.

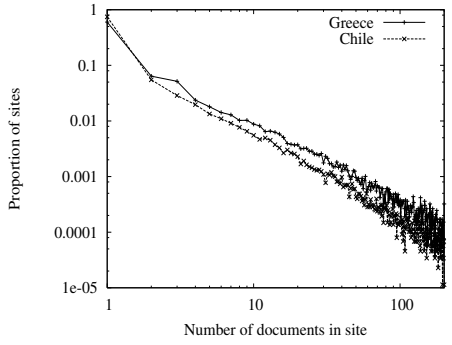

Figure 3: Number of pages per website.

2. LINKS

The web graph is usually characterized as a random graph created by a process of preferential attachment [4], that induces scale-free properties [2]. Figure 4 shows that these two sub-graphs of the web have these characteristics, revealing the existence of self-similarities. The power law parameter depends a lot on the range of data used. Taking degrees of at most 350, we obtain -2.02 and -2.11 for in-degree, and -2.17 and -2.40 for out-degree; for .GR and .CL, respectively. Discarding degrees smaller than 50, the parameters are closer to -2.3 and -2.8 for in-degree and out-degree. This should be compare with the results in [4] that found -2.1 and -2.7, respectively, for 200 million pages in 1999.

The distribution of out-degree is different, as the in-degree in many cases reflects the popularity of a web page, while the out-degree reflects a design choice of the page maintainer. Also, it is much easier to have a page with many outgoing links than one with many incoming links.

For the graph components, we use the bow-tie structure proposed by Broder et al. [3]; but we considered only links between different websites, collapsing all the pages of a website to a single node of the graph. We show the relative size of components in Figure 5.

Note that that the `MAIN` (the giant strongly connected component) seems to be larger in the Greek web in expense of the `ISLAND` component - this can be an indicator of a better connected Web, although the seeds for the Chilean crawling had more islands.

We also studied the relationship of the collections with other top level domains reflecting cultural and economic relationships; this is summarized in Table 2.

Greece		Chile	
COM	49.2%	COM	58.6%
ORG	17.9%	ORG	15.4%
NET	8.5%	NET	6.4%
Germany	3.7%	Germany	2.6%
United Kingdom	2.6%	United Kingdom	1.4%
EDU	2.6%	EDU	1.3%
TV	1.3%	Mexico	1.2%
Russian Federation	1.3%	Brazil	1.1%
Taiwan	1.1%	Argentina	0.9%
Netherlands	0.9%	Spain	0.9%

Table 2: Most referenced external top-level domains.

3. CONCLUSIONS AND FUTURE WORK

Considering the different culture, language, etc., both domains studied are quite similar. This indicates that small subsets of the Web perhaps resemble better the whole Web than other collections (e.g. the .GOV sample used for the TREC Web track). We are currently preparing the first comprehensive study of the `.GR` domain and are planning to include other country-level domains in a larger comparative study.

4. REFERENCES

[1] BAEZA-YATES, R., AND CASTILLO, C. Balancing volume, quality and freshness in web crawling. In *Soft Computing Systems - Design, Management and Applications* (2002), pp. 565–572.

[2] BARABASI, A. L. *Linked: the new science of networks*. Perseus Publishing, 2002.

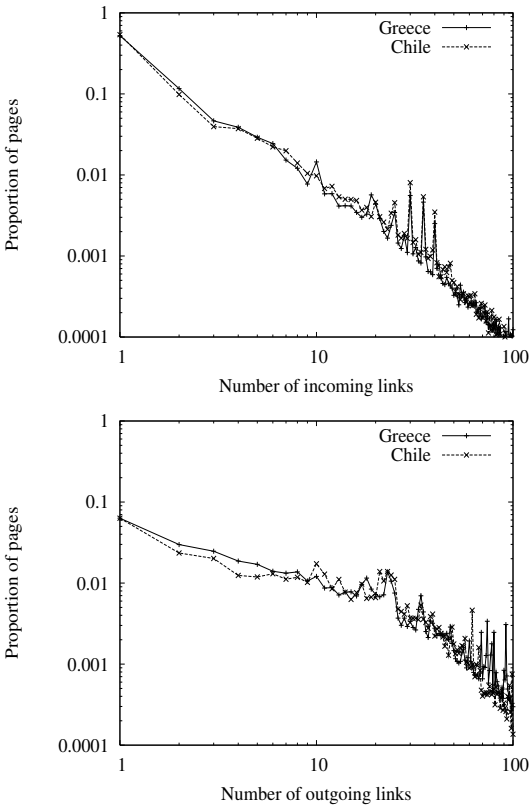

Figure 4: Distribution of in-degree and out-degree.

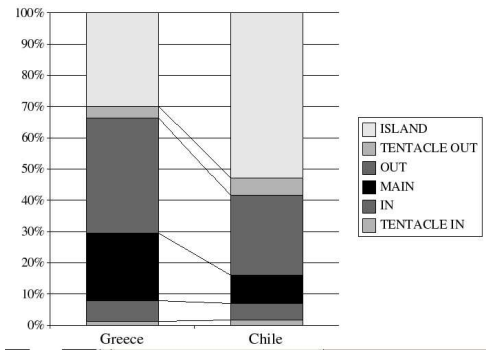

Figure 5: Relative size of graph components.

[3] BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A., AND WIENER, J. Graph structure in the web: Experiments and models. In *9th World Wide Web Conference* (2000), pp. 309–320.

[4] KUMAR, RAGHAVAN, RAJAGOPALAN, SIVAKUMAR, TOMKINS, AND UPFAL. Stochastic models for the web graph. In *FOCS: IEEE Symposium on Foundations of Computer Science (FOCS)* (2000), pp. 57–65.

[5] THE ECONOMIST. Country Profiles, 2002.

[6] UNITED NATIONS. Population Division, 2002.

[7] UNITED NATIONS. Human Development Reports, 2003.