Link Analysis in National Web Domains

Ricardo Baeza-Yates
ICREA Professor
University Pompeu Fabra
ricardo.baeza@upf.edu

Carlos Castillo
Department of Technology
University Pompeu Fabra
carlos.castillo@upf.edu

Abstract

The Web can be seen as a graph in which every page is a node, and every hyper-link between two pages is an edge. This Web graph forms a scale-free network: a graph in which the distribution of the degree of the nodes is very skewed. This graph is also self-similar, in terms that a small part of the graph shares most properties with the entire graph.

This paper compares the characteristics of several national Web domains, by studying the Web graph of large collections obtained using a Web crawler; the comparison unveils striking similarities between the Web graphs of very different countries.

1 Introduction

Large samples from specific communities, such as national domains, have a good balance between diversity and completeness. They include pages inside a common geographical, historical and cultural context that are written by diverse authors in different organizations. National Web domains also have a moderate size that allows good accuracy in the results; because of this, they have attracted the attention of several researchers.

In this paper, we study eight national domains. The collection studied include four collections obtained using WIRE [3]: Brazil (BR domain) [18, 15], Chile (CL domain) [1, 8, 4], Greece (GR domain) [12] and South Korea (KR domain) [7]; three collections obtained from the Laboratory of Web Algorithmics[1]: Indochina (KH, LA, MM, TH and VN domains), Italy (IT domain) and the United Kingdom (UK domain); and one collection obtained using Akwan [10]: Spain (ES domain) [6]. Our 104-million page sample is less than 1% of the indexable Web [13] but presents characteristics that are very similar to those of the full Web.

[1]Laboratory of Web Algorithmics, *Dipartimento di Scienze dell'Informazione*, *Universitá degli studi di Milano*, <http://law.dsi.unimi.it/>.

Table 1 summarizes the characteristics of the collections. The number of unique hosts was measured by the ISC[2]; the last column is the number of pages actually downloaded.

Table 1. Characteristics of the collections.

Collection	Year	Available hosts [mill]	Available hosts (rank)	Pages [mill]
Brazil	2005	3.9	11^{th}	4.7
Chile	2004	0.3	42^{th}	3.3
Greece	2004	0.3	40^{th}	3.7
Indochina	2004	0.5	38^{th}	7.4
Italy	2004	9.3	4^{th}	41.3
South Korea	2004	0.2	47^{th}	8.9
Spain	2004	1.3	25^{th}	16.2
U. K.	2002	4.4	10^{th}	18.5

By observing the number of available hosts and the downloaded pages in each collection, we consider that most of them have a high coverage. The collections of Brazil and the United Kingdom are smaller samples in comparison with the others, but their sizes are large enough to show results that are consistent with the others.

Zipf's law: the graph representing the connections between Web pages has a scale-free topology. Scale-free networks, as opposed to random networks, are characterized by an uneven distribution of links. For a page p, we have $Pr(p \text{ has } k \text{ links}) \propto k^{-\theta}$. We find this distribution on the Web in almost every aspect, and it is the same distribution found by economist Vilfredo Pareto in 1896 for the distribution of wealth in large populations, and by George K. Zipf in 1932 for the frequency of words in texts. This distribution later turned out to be applicable to several domains [19] and was called by Zipf the law of *minimal effort*.

Section 2 studies the Web graph, and section 3 the Host-graph. The last section presents our conclusions.

[2]Internet Systems Consortium's domain survey, <http://www.isc.org/ds/>

2 Web graph

2.1 Degree

The distributions of the indegree and outdegree are shown in Figure 1; both are consistent with a power-law distribution. When examining the distribution of outdegree, we found two different curves: one for smaller outdegrees –less than 20 to 30 out-links– and another one for larger outdegrees. They both show a power-law distribution and we estimated the exponents for both parts separately.

For the in-degree, the average power-law exponent θ we observed was 1.9 ± 0.1; this can be compared with the value of 2.1 observed by other authors [9, 11] in samples of the global Web. For the out-degree, the exponent was 0.6 ± 0.2 for small outdegrees, and 2.8 ± 0.8 for large out-degrees; the latter can be compared with the parameters 2.7 [9] and 2.2 [11] found for samples of the global Web.

2.2 Ranking

One of the main algorithms for link-based ranking of Web pages is PageRank [16]. We calculated the PageRank distribution for several collections and found a power-law in the distribution of the obtained scores, with average exponent 1.86 ± 0.06. In theory, the PageRank exponent should be similar to the indegree exponent [17] (the value they measured for the exponent was 2.1), and this is indeed the case. The distribution of PageRank values can be seen in Figure 2.

We also calculated a static version of the HITS scores [14], counting only external links and calculating the scores in the whole graph, instead of only on a set of pages. The tail of the distribution of authority-score also follows a power law. In the case of hub-score, it is difficult to assert that the data follows a power-law because the frequencies seems to be much more dispersed. The average exponent observed was 3.0 ± 0.5 for hub score, and 1.84 ± 0.01 for authority score.

3 Hostgraph

We studied the hostgraph [11], this is, the graph created by changing all the nodes representing Web pages in the same Web site by a single node representing the Web site. The hostgraph is a graph in which there is a node for each Web site, and two nodes A and B are connected iff there is at least one link on site A pointing to a page in site B. In this section, we consider only the collections from which we have a hostgraph.

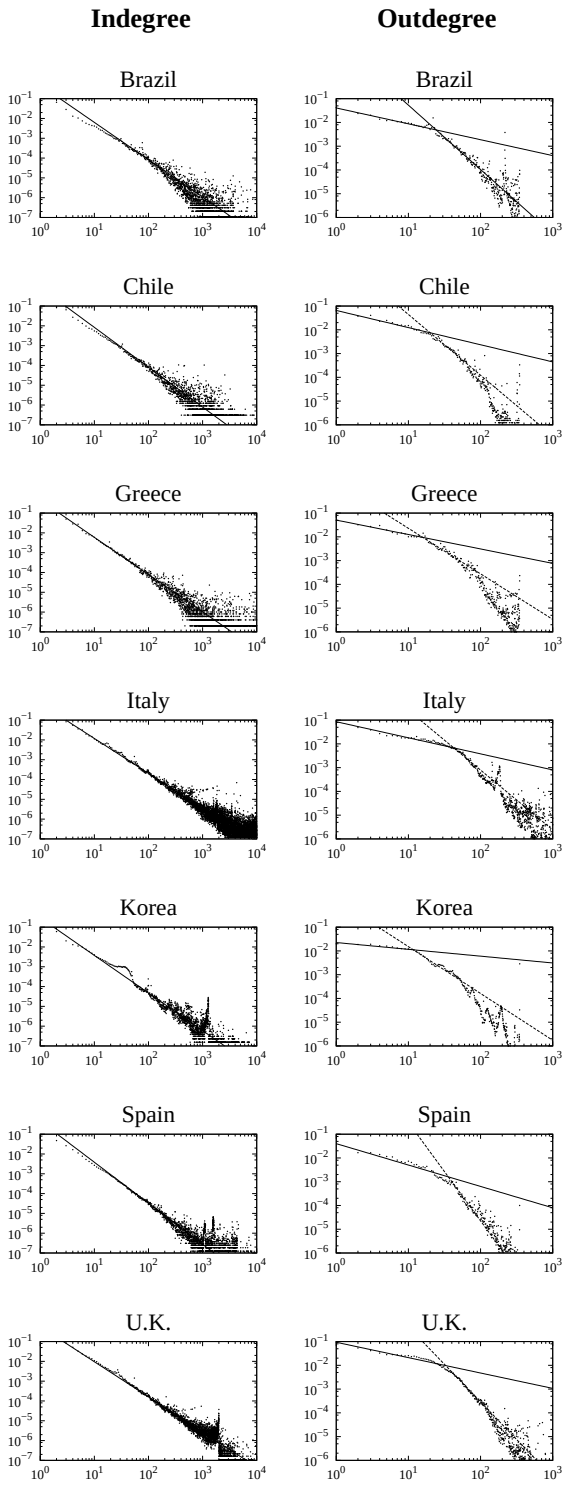

Figure 1. Histograms of the indegree and outdegree of Web pages, including a fit for a power-law distribution.

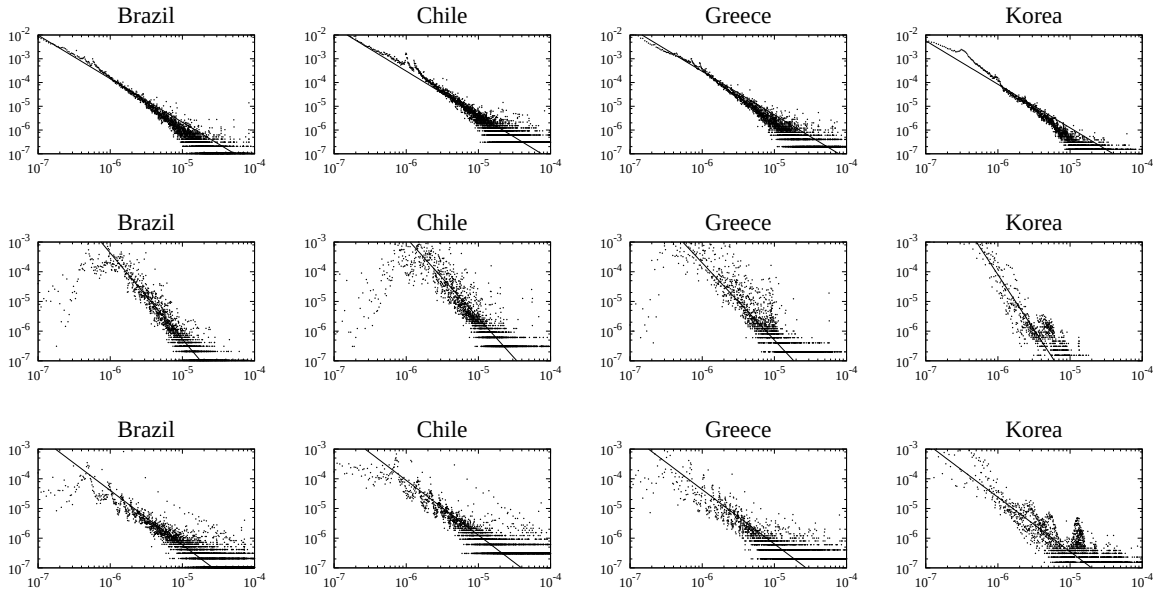

Figure 2. Histograms of the scores using PageRank (top), hubs (middle) and authorities (bottom).

3.1 Degree

The average indegree per Web site (average number of different Web sites inside the same country linking to a given Web site) was 3.5 for Brazil, 1.2 for Chile, 1.6 for Greece, 37.0 for South Korea and 1.5 for Spain. The histograms of indegree is consistent with a Zipfian distribution, with parameter 1.8 ± 0.3.

By manual inspection we observed that in Brazil and specially in South Korea, there is a significant use –and abuse– of DNS wildcarding. DNS wildcarding is a way of configuring DNS servers so they reply with the same IP address no matter which host name is used in a DNS query.

The average outdegree per Web site (average number of different Web sites inside the same country linked by a given Web site) was 2.2 for Brazil, 2.4 for Chile, 4.8 for Greece, 16.5 for South Korea and 11.2 for Spain. The distribution of outdegree also exhibits a power-law with parameter 1.6 ± 0.3.

We also measured the number of internal links, that is, links going to pages inside the same Web site. We normalized this by the number of pages in each Web site, to be able to compare values. We observed a combination of two power-law distributions: one for Web sites with up to 10 internal links per Web page on average, and one for Web sites with more internal links per Web page. For the sites with less than 10 internal links per page on average, the parameter for the power-law was 1.1 ± 0.3, and for sites with more internal links per page on average, 3.0 ± 0.3.

3.2 Web structure

Broder et al. [9] proposed a partition of the Web graph based on the relationship of pages with the larger strongly connected component (SCC) on the graph. The pages in the larger strongly connected component belong to the category MAIN. All the pages reachable from MAIN by following links forwards belong to the category OUT, and by following links backwards to the category IN. The rest of the Web that is weakly connected (disregarding the direction of links) to MAIN is in a component called TENDRILS.

In [2] we showed that this macroscopic structure is similar at the hostgraph level: the hostgraphs we examined are scale-free networks and have a giant strongly connected component. We observed that distribution of the sizes of their strongly connected components is shown in Figure 3.

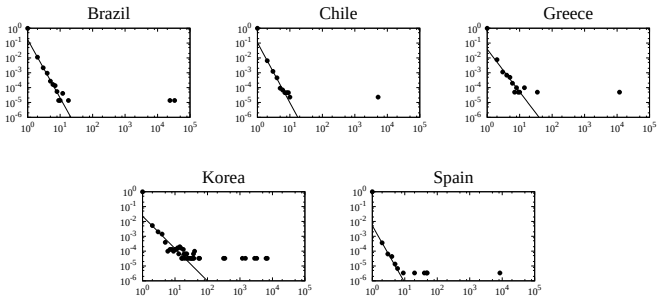

Figure 3. Histograms of the sizes of SCCs.

The parameter for the power-law distribution was 2.7 ± 0.7. In Chile, Greece and Spain, a sole giant SCC appears having at least 2 orders of magnitude more Web sites than the second largest SCC component. In the case of Brazil, there are two giant SCCs. The larger one is a "natural" one, containing Web sites from different domains. The second larger is an "artificial" one, containing only Web sites under a domain that uses DNS wildcarding to create a "link farm" (a strongly connected community of mutual links). In the case of South Korea, we detected at least 5 large link farms.

Regarding the Web structure, the distribution between sites in general gives the component called OUT a large share. If we do not consider sites that are weakly connected to MAIN, IN has on average 8% of the sites, MAIN 28%, OUT 58% and TENDRILS 6%. The sites that are disconnected from MAIN are 40% on average, but contribute less than 10% of the pages.

4 Conclusions

Even when the collections were obtained from countries with different economical, historical and geographical contexts, and speaking different languages we observed that the results across different collections are always consistent when the observed characteristic exhibits a power-law in one collection. In this class we include the distribution of degrees, link-based scores, internal links, etc.

Besides links, we are working in a detailed account of the characteristics of the contents and technologies used in several collections [5].

Acknowledgments: We worked with Vicente López in the study of the Spanish Web, with Efthimis N. Efthimiadis in the study of the Greek Web, with Felipe Ortiz, Bárbara Poblete and Felipe Saint-Jean in the studies of the Chilean Web and with Felipe Lalanne in the study of the Korean Web. We also thank the Laboratory of Web Algorithmics for making their Web collections available for research.

References

[1] R. Baeza-Yates and C. Castillo. Caracterizando la Web Chilena. In *Encuentro chileno de ciencias de la computación*, Punta Arenas, Chile, 2000. Sociedad Chilena de Ciencias de la Computaci´on.

[2] R. Baeza-Yates and C. Castillo. Relating Web characteristics with link based Web page ranking. In *Proceedings of String Processing and Information Retrieval SPIRE*, pages 21–32, Laguna San Rafael, Chile, 2001. IEEE CS Press.

[3] R. Baeza-Yates and C. Castillo. Balancing volume, quality and freshness in Web crawling. In *Soft Computing Systems - Design, Management and Applications*, pages 565–572, Santiago, Chile, 2002. IOS Press Amsterdam.

[4] R. Baeza-Yates and C. Castillo. Caracter´ısticas de la Web Chilena 2004. Technical report, Center for Web Research, University of Chile, 2005.

[5] R. Baeza-Yates and C. Castillo. Characterization of national Web domains. Technical report, Universitat Pompeu Fabra, July 2005.

[6] R. Baeza-Yates, C. Castillo, and V. L´opez. Caracter´ısticas de la Web de España. Technical report, Universitat Pompeu Fabra, 2005.

[7] R. Baeza-Yates and F. Lalanne. Characteristics of the Korean Web. Technical report, Korea–Chile IT Cooperation Center ITCC, 2004.

[8] R. Baeza-Yates and B. Poblete. Evolution of the Chilean Web structure composition. In *Proceedings of Latin American Web Conference*, pages 11–13, Santiago, Chile, 2003. IEEE CS Press.

[9] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web: Experiments and models. In *Proceedings of the Ninth Conference on World Wide Web*, pages 309–320, Amsterdam, Netherlands, May 2000. ACM Press.

[10] A. S. da Silva, E. A. Veloso, P. B. Golgher, , A. H. F. Laender, and N. Ziviani. CoBWeb - A crawler for the Brazilian Web. In *Proceedings of String Processing and Information Retrieval (SPIRE)*, pages 184–191, Cancun, Mxico, 1999. IEEE CS Press.

[11] S. Dill, R. Kumar, K. S. Mccurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Trans. Inter. Tech.*, 2(3):205–223, 2002.

[12] E. Efthimiadis and C. Castillo. Charting the Greek Web. In *Proceedings of the Conference of the American Society for Information Science and Technology (ASIST)*, Providence, Rhode Island, USA, November 2004. American Society for Information Science and Technology.

[13] A. Gulli and A. Signorini. The indexable Web is more than 11.5 billion pages. In *Poster proceedings of the 14th international conference on World Wide Web*, pages 902–903, Chiba, Japan, 2005. ACM Press.

[14] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[15] M. Modesto, ´a. Pereira, N. Ziviani, C. Castillo, and R. Baeza-Yates. Un novo retrato da Web Brasileira. In *Proceedings of SEMISH*, São Leopoldo, Brazil, 2005.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[17] G. Pandurangan, P. Raghavan, and E. Upfal. Using Pagerank to characterize Web structure. In *Proceedings of the 8th Annual International Computing and Combinatorics Conference (COCOON)*, volume 2387 of *Lecture Notes in Computer Science*, pages 330–390, Singapore, August 2002. Springer.

[18] E. A. Veloso, E. de Moura, P. Golgher, A. da Silva, R. Almeida, A. Laender, R. B. Neto, and N. Ziviani. Um retrato da Web Brasileira. In *Proceedings of Simposio Brasileiro de Computacao*, Curitiba, Brasil, 2000.

[19] G. K. Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley, Cambridge, MA, USA, 1949.