Charting the Greek Web

Efthimis N. Efthimiadis and Carlos Castillo
The Information School, University of Washington and Universidad de Chile {efthimis@u.washington.edu}

♦Introduction

This research reports on the first systematic way of charting the Greek web. During May 2004 the Greek web space was crawled using the WIRE crawler.

We show several similarities that contribute to validate more general models for the characteristics of the Web, especially in terms of link structure.

We downloaded 4 million 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 100Kb per page; and considered only pages under the .gr domain.

We report on the Greek web graph and the distribution of in-degree and out-degree links. For the graph components, we use the bow-tie structure proposed by Broder et al.; but we considered only links between different websites, collapsing all the pages of a website to a single node of the graph.

The MAIN, that is the giant strongly connected, component seems to be large in the Greek web in expense of the ISLAND component - this can be an indicator of a better connected Web.

We also report on the relationship of the Greek web with other top level domains reflecting cultural and economic relationships.

Collection Statistics

Site summary

Number of sites ok	29,191
Number of sites with valid page age	22,090
Average internal links	1,093.67
Average pages per site	146.90
Average static pages per site	85.42
Average dynamic pages per site	61.48
Average of age of oldest page in months	16.86
Average of age of average page in months	12.23
Average of age of newest page in months	9.74
Average in-degree	5.37
Average out-degree	5.37
Average site max depth	3.65
Average site size in MB	2.61

Webpages summary

Total Web pages	4,051,326	
Unique	3,781,912	93.35%
Duplicates	269,414	6.65%
Static	2,524,270	62.31%
Dynamic	1,527,056	37.69%

Languages

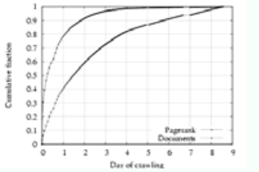
We took a sample of 6,331 pages, and analyzed their contents to compare their word lists against a series of lists of stop words in several languages on the Greek Web. We found about 63% of the pages in Greek, and 34% in English. Other languages appeared with much less frequency.

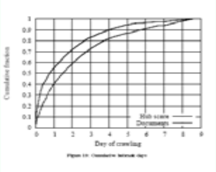
Documents sampled	6,331
Documents with more than 50 words	5,350
Documents that were not identified	2,141

Document language	Number of documents	Percent
greek	2,660	63.48%
english	1,436	34.27%
german	32	0.76%
french	24	0.57%
italian	9	0.21%
norwegian	9	0.21%
spanish	8	0.19%
dutch	5	0.12%
portugues	3	0.07%
turkish	3	0.07%
danish	1	0.02%
catala	0	0%
swedish	0	0%

♦Crawling

The graphs below show that in just 4 days 80% of the documents and more than 95% of the important pages (by Pagerank) from the Greek Web were crawled.

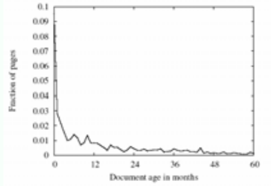




Web Page Characteristics

Step name	HTTP Code	Documents	Percent
Ok	200	[illegible]	[illegible]
Found	302	[illegible]	[illegible]
Not Found	404	[illegible]	[illegible]
Error Dns	[illegible]	[illegible]	[illegible]
Moved	301	[illegible]	[illegible]
Error Timeout	[illegible]	[illegible]	[illegible]
Internal Error	500	[illegible]	[illegible]
Bad Request	400	[illegible]	[illegible]
Not Acceptable	406	[illegible]	[illegible]
Partial	206	[illegible]	[illegible]
Forbidden	403	[illegible]	[illegible]
Error Disconnected	[illegible]	[illegible]	[illegible]
See Other	303	[illegible]	[illegible]
Not Modified	304	[illegible]	[illegible]
Unauthorized	401	[illegible]	[illegible]
Error Skipped	[illegible]	[illegible]	[illegible]
Error Protocol	[illegible]	[illegible]	[illegible]
Multiple Choices	300	[illegible]	[illegible]
Bad Gateway	502	[illegible]	[illegible]
Error Blocked Ip	[illegible]	[illegible]	0%
No Content	204	[illegible]	0%
Method Not Allowed	405	[illegible]	0%
Range Error	416	[illegible]	0%
Not Implemented	501	[illegible]	0%
Unavailable	503	[illegible]	0%

Document Age

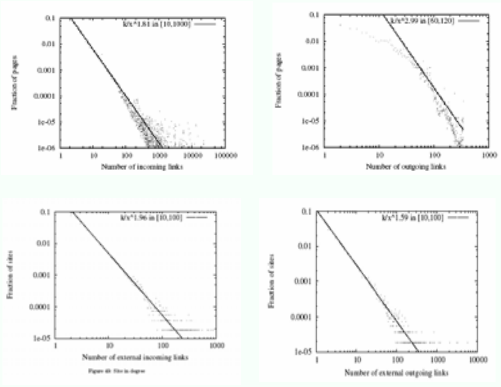


Age in Years	Documents	Percent
0	1,070,229	58.31%
1	284,801	15.52%
2	172,899	9.42%
3	146,247	7.97%
4	68,233	3.72%
5	47,779	2.60%
6	19,972	1.09%
7	10,242	0.56%
8	8,195	0.45%
9	2,219	0.12%
10	2,080	0.11%

♦Links

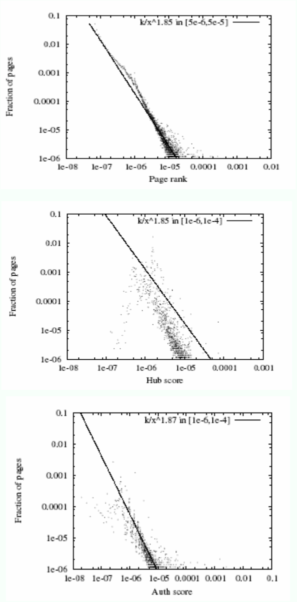
♦Degree

Distribution of in- and out-degree.



The distribution of in-degree is much more skewed than the distribution of out-degree, as shown in the figure above. Having a Web page with a large in-degree is much more difficult than having a page with a large out-degree.

Distribution of Pagerank, global Hubs and Authority Scores



♦ Most linked Web sites

The most linked web sites on the Greek Web are listed below. There is a very strong presence of university and government related web sites in the top places.

Top sites by in_degree	in_degree
users.otenet.gr	900
www.culture.gr	856
www.ypepth.gr	794
www.forthnet.gr	713
www.in.gr	712
www.ntua.gr	703
www.otenet.gr	559
www.auth.gr	545
www.minenv.gr	484
www.uoa.gr	481

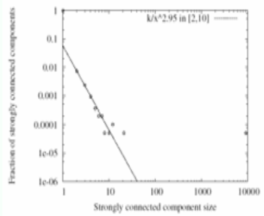
Top sites by out_degree	out_degree
www.visto.gr	5,745
www.dimostavrou.gr	3,314
www.pantelon.gr	3,087
users.otenet.gr	2,094
homepages.pathfinder.gr	1,228
www.evresi.gr	1,181
www.goo.gr	1,098
www.kairatos.com.gr	883
www.money.gr	847
www.bsa.gr	847

♦ Strongly Connected Components

Total number of site names known	54,748
Sites with at least one page ok	29,191
Sites without in links (but at least one page ok)	7,414
Sites without out links (but at least one page ok)	15,386
Size of largest SCC	9,191
SCC-id of largest SCC	33,423
Number of SCCs with one site only (singletons)	19,354

The distribution of the sizes of the strongly connected components (SCC) on the graph of Web sites is given below. A giant strongly connected component appears, as observed by Broder et al. (2000). This is a typical signature of a scale-free network. The distribution of SCC sizes is presented in the figure below, here we are considering only Web sites with links to other Web sites.

SCC size	Number of SCC components	Percent
1	19,354	98.84%
2	143	0.73%
3	46	0.23%
4	18	0.09%
5	7	0.04%
6	4	0.02%
7	4	0.02%
8	1	0.01%
10	1	0.01%
12	2	0.01%
21	1	0.01%
9,191	1	0.01%

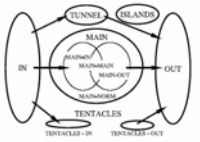


♦ Web Links Structure

For the analysis we use an extension of the MAIN component (Broder et. al) introduced by (Baeza-Yates & Castillo) for analyzing Web structure. This divides the MAIN component into four parts:

(a) MAIN-MAIN, which are sites that can be reached directly from the IN component and can reach directly the OUT component;
(b) MAIN-IN, which are sites that can be reached directly from the IN component but are not in MAIN-MAIN;
(c) MAIN-OUT, which are sites that can reach directly the OUT component, but are not in MAIN-MAIN;
(d) MAIN-NORM, which are sites not belonging to the previously defined subcomponents.

Note that the Web sites in the ISLANDS component are found only by directly accessing the home page of those Web sites. This is possible because we had 70% of the registered domains under .gr at the time of the study. The distribution of Web sites into components is shown in the figure below. This structure evolves over time.



Component name	Number of sites	Percent
MAIN_NORM	4,676	16.02%
MAIN_MAIN	1,056	3.62%
MAIN_IN	533	1.83%
MAIN_OUT	2,926	10.02%
IN	1,677	5.74%
OUT	12,064	41.33%
TIN	184	0.63%
TOUT	675	2.31%
TUNNEL	32	0.11%
ISLAND	5,368	18.39%

♦ References

Broder, A., 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 9th Conference on WWW, pg: 309-320, Amsterdam, Netherlands, May 2000.

Baeza-Yates, R.and C. Castillo. Relating Web characteristics with link based Web page ranking. In Proceedings of String Processing & Information Retrieval, pages 21-32, Laguna San Rafael, Chile, November 2001. IEEE Cs. Press.