"CRAWL.PL" Measuring Statistical and Structural Properties of the Polish Web. Technical Report

Carlos Castillo[1], Bartłomiej Starosta[2], Marcin Sydow[2]

[1] University of Rome „La Sapienza",
currently at Yahoo! Research Barcelona, Italy
[2] Polish-Japanese Institute of Information Technology,
ul. Koszykowa 86, 02-008 Warsaw, Poland

Abstract. This document summarizes the results of an experiment made in the Polish-Japanese Institute of Information Technology, Warsaw, Poland during autumn 2005 and winter 2006. The goal of the project was to collect and analyze large portion of Polish Web documents in order to characterize the structure and other properties of the „.pl" domain. Up to the knowledge of the authors, it was the first publicly reported research experiment of this kind over the Polish Web. The following sections include information about downloaded Web pages, Web sites, and their characteristics. We also present various statistics concerning hosts and domains, as well as the link structure. Among the results of the experiment are the first data sets representing graphs of the Polish Web which will be publicly available for other researchers.

Keywords. Web crawling, Web graphs, page rank, Polish Web

1 Introduction

1.1 The Crawl.pl Project

The Goals of the Project

The aim of the "crawl.pl" project is to gather data and compute general statistics concerning a substantial part of the Polish Web, in particular, the ".pl" WWW domain.

Another goal was to create data sets concerning the Polish Web graph which will be freely available for all the interested researchers, since no such data sets over the ".pl" domain are available in the time of writing[1].

Most of the data on which the statistics are based was obtained and processed using the WIRE crawler (see section "the Platform").

In all measurements presented here, references to any particular Web hosts, sites or documents are intentionally avoided. Instead, the analysis focuses on general statistical properties of the Polish Web.

Credits

The project was supported by the grant ST/AI/03/2005 PJIIT founded by the Polish-Japanese Institute of Information Technology in Warsaw, Poland.

Output of the Project

The document reports many measurements of general statistical properties of the Polish Web made during the "crawl.pl" project. Moreover, two data sets concerning the Polish Web graph were prepared: 20M-node document graph and 167K-node host graph. These data sets are freely available for any interested researchers via HTTP. For the details refer to the datasets' URL [2]

To the best of our knowledge the data sets are the first publicly available data sets of this scale concerning the Polish Web graph.

1.2 Related Work

The project described here may be considered as a continuation and completion of the similar projects performed by other researchers on Web collections of other countries. The Chilean Web is analyzed in [3], Korean in [4]. The Greek Web is compared to the Chilean in [6]. A synthesis of this works may be found in [5]. The analysis of quite young African Web is contained in [5].

1 Except for those kept by commercial subjects such as search engines, etc.

An interesting study of relationships between the general statistical properties of the Web and the economical collaboration between countries is in [14].

1.3 Organization of the paper

This report is organized as follows. Chapter 2 describes the organizational and technical aspects of the project and reports some general statistics. Chapter 3 reports various statistics concerning the whole sites rather than particular documents in the ".pl" domain. Chapter 4 concerns the statistics of the document types encountered during the crawl. Chapter 5 is exclusively devoted to visualizing the distributions of various statistics collected during the project. In chapter 6 conclusion and future directions are contained. In the appendix, the high-level harvest log of the crawl is enclosed.

2 General Information about the Crawl

2.1 The Meta-Level Information

Time, Scale and Character of the Crawl

The crawl started at the end of 2005 and continued until the end of January 2006 and was a general crawl. During the crawl, over 20M documents were encountered from over 160K Polish hosts. Detailed numbers are presented in subsection "Size of the Crawl".

For convenience, the contents of most of the Web documents encountered by the crawler were not stored, since it was not necessary in the context of the project goals and would involve massive storage demands. Instead, a fingerprint (an MD5 checksum of the pages' contents) was kept.

However, the complete link structure concerning all the encountered documents and hosts was stored during the crawl and later served as the basis for preparing the publicly available data sets mentioned in subsection "Output of the Project". The scale of collected data is not claimed to be the largest in Poland - for example, the largest (at the time of writing) Polish search engine [1] is indexing a collection that is approx. 4 times larger - but the „crawl.pl" project is the first non-commercial, research-related project of this scale in Poland, up to the knowledge of the authors.

The Platform

The project was running on the desktop PC with 3Ghz CPU, 1GB RAM and 100GB disk under the Linux operating system. Web documents were downloaded using the WIRE crawler created in the Center for Web Research, University of Chile.

Crawl Parameters

Due to the limitations of the machine we were downloading up to 100KB of each Web page and up to 10K URLs per each site. We downloaded up to 5 levels of dynamic pages and up to 15 levels of static pages.

The initial set of "seeds" (starting URLs) contained about 37K URLs[2]. We did an initial, exploratory crawl in which we obtained a set of about 211K host names ending in ".pl", contained approximately 2.8M pages. We removed from this list of hosts a large set of spam hosts containing „link farms", typically pointing to pornographic Web sites in other domains. We kept only 160K of the hosts after this cleaning phase, and downloaded pages only from those hosts. During the actual crawling phase we downloaded about 23M pages.

2.1 General Crawl Statistics

Size of the crawl

The table 2.1 reports the detailed numbers concerning the crawl size. 21M of the downloaded Web pages were unique. This can be compared, for instance, with [15] in which the amount of duplicate pages found was about 22%.

We found that about 60% of the pages were static (see the table 2.1). Dynamic pages were identified because they have a question mark in the URL and/or they end in a filename extension that is known to be used for dynamic Web content languages (`.php`, `.asp`, `.jsp`, etc.).

Table 2.1. Summary of the downloaded pages

Total pages	23,596,078	
Unique	21,622,036	91.63%
Duplicates	1,974,042	8.37%
Static	14,008,730	59.37%
Dynamic	9,587,348	40.63%

In comparison to the portions of some European national Webs examined in [5], the number of 23M of pages examined in the „crawl.pl" project makes it quite a big analyzed collection - more precisely the second largest after the Italian (with 41M examined pages) and before United Kingdom (over 18M) and Spain (with about 16M of pages). Of course, these numbers do not represent the actual sizes of the mentioned Webs, which depend on the limits set for the crawler

2 Thanks are due to NetSprint - a Polish search engine - for help in collecting the initial URL set

(maximum exploration depth and maximum number of pages per host).

Mime Types

The crawler was instructed to download only pages matching the `text/plain` or `text/html` MIME content types. This was included in the crawler's request in the form of an `Accept:` HTTP header. However, a few hosts did not honor this header and respond with several different types of data, even images or video (see table 2.2). The table considers all the responses obtained from servers, including HTTP redirect (second row).

Table 2.2. Mime types

mime type	Documents	Percent
text/html	21,014,667	89.06%
redirect	1,656,999	7.02%
unknown	480,162	2.03%
image	162,616	0.69%
robots/txt	156,094	0.66%
text/plain	81,839	0.35%
application	30,461	0.13%
audio	6,017	0.03%
text/xml	4,697	0.02%
application/flash	1,831	0.01%
text/tex	365	0%
video	270	0%
text/wap	60	0%

Age of Documents

Most documents (about 77%) are less than one year old, and only about 10% of the data has over 2 years, meaning that the Polish Web is still growing and that documents on the Web seems to be fairly well maintained (updated). Only about 2% of documents are older than 4 years (see table 2.3).

Table 2.3. Age in years (concerns only the documents with explicit date found)

Age in Years	Documents	Percent
0	2,786,259	77.01%
1	321,092	8.87%
2	256,891	7.10%
3	100,385	2.77%
4	68,621	1.90%
5	31,762	0.88%
6	27,677	0.76%
7	8,399	0.23%
8	6,756	0.19%
9	4,527	0.13%
10	981	0.03%

HTTP Codes

The table 2.4 shows the distribution of the HTTP response code. The code OK (81.43%) results in a page transfer, as well as all codes from 200-299. The code FOUND means that the requested resource resides temporarily under a different URI (6.02%). About 5% of the requested URLs were NOT FOUND, meaning that they are not found because of a broken link (usually an obsolete link).

Table 2.4. Distribution of HTTP response codes

Http status	HTTP Code	Documents	Percent
Ok	200	19,215,372	81.43%
Found	302	1,421,193	6.02%
Not Found	404	1,278,955	5.42%
Partial	206	501,297	2.12%

Error Timeout	95	298,129	1.26%
Forbidden	403	251,179	1.06%
Moved	301	234,691	0.99%
Error Connect	97	95,807	0.41%
Not Modified	304	97,303	0.41%
Error Disconnected	96	73,611	0.31%
Bad Request	400	34,366	0.15%
Internal Error	500	25,533	0.11%
Error Skipped	92	20,798	0.09%
Not Acceptable	406	14,117	0.06%
Unavailable	503	11,338	0.05%
Error Dns	98	7,936	0.03%
Unauthorized	401	7,420	0.03%
Error Blocked Ip	90	1,945	0.01%
Temporary Redirect	307	2,310	0.01%
Error Protocol	94	195	0%
Invalid	201	3	0%
Invalid	202	104	0%
No Content	204	225	0%
Multiple Choices	300	279	0%
See Other	303	47	0%
Method Not Allowed	405	10	0%
Request Timeout	408	11	0%
Invalid	410	1	0%
Uri Too Long	414	1	0%
Range Error	416	982	0%
Invalid	419	39	0%
Not Implemented	501	3	0%

Bad Gateway	502	865	0%
Gateway Timeout	504	1	0%
Invalid	505	12	0%

The value of about 81% of correct responses and 5% of pages not found is similar to responses in other countries, as described in [5].

2.3 Top-level Domain Interconnections

It is very interesting to examine the shares of the top level national domains of the documents being linked to by the documents in the „.pl" domain.

The table 2.5 contains the list of the 40 most linked top-level domains linked by documents in „.pl".

At least 4 factors seem to determine the presence (or absence) of the particular domains among the top ones:

1. geographical location,
2. commercial relationships,
3. informational relationships,
4. political/social issues.

For example, the influence of the first factor explains the fact that most of the national domains in the top 40 are the European domains, with the German domain (.de) being the leader. Germany is (geographically) the closest western-European country to Poland and is strongly economically connected with Poland.

Similarly, the Russian domain is the top in the table among the eastern and northern-eastern neighbors of Poland. In reality, Lithuania (position 36) and Ukraine (position 39) are geographically closer to Poland, but the Russian Web is much more developed than the Ukrainian or Lithuanian.

Also southern direct geographical neighbors of Poland (Czech and Slovakia) are present in the top 40 domains. The same concerns the northern neighbors (through sea) - Sweden and Denmark.

The second factor - commercial relationships - explains the very high (the first after the „.pl" domain) position of the „.com" domain in the list, despite the fact of distant geographical locations of the subjects from that domain.

Similarly, the third factor - informational relationships - seems to explain the very high position of the „.org" domain.

Finally, the fourth factor - the political/social issues - seems to explain the very meaningful absence of the Belorussian domain, despite the fact that this country shares border with Poland. More precisely, the explanation could be that this country is not democratic yet, and not yet fully open to the economical,

technological and social development and free exchange of information.

Table 2.5. Top 40 mostly linked top-level domains

TOP-LEVEL DOMAIN	Number of external links to web pages found	Percent
PL - Poland	361,783,534	85.52%
COM	25,828,291	6.11%
ORG	10,230,463	2.42%
NET	8,282,139	1.96%
DE - Germany	5,120,552	1.21%
INFO	2,691,847	0.64%
UK	1,047,582	0.25%
BIZ	705,915	0.17%
IT - Italy	620,581	0.15%
CH - Switzerland	428,890	0.10%
BE - Belgium	393,102	0.09%
RU - Russian Federation	381,145	0.09%
NL - Netherlands	372,489	0.09%
AG - Antigua and Barbuda	344,376	0.08%
CZ - Czech Republic	321,152	0.08%
DK - Denmark	315,399	0.07%
SE - Sweden	288,658	0.07%

FR - France	269,250	0.06%
JP - Japan	265,118	0.06%
EDU	229,065	0.05%
US - United States	209,058	0.05%
ES - Spain	200,173	0.05%
FI - Finland	189,123	0.04%
BR - Brazil	168,612	0.04%
CN - China	158,084	0.04%
KR - Korea Republic of	146,387	0.03%
NO - Norway	137,407	0.03%
AT - Austria	123,264	0.03%
MX - Mexico	121,125	0.03%
SK - Slovakia	103,366	0.02%
CA - Canada	95,297	0.02%
RO - Romania	77,804	0.02%
INT	74,263	0.02%
GOV	73,383	0.02%
HU - Hungary	70,637	0.02%
LT - Lithuania	65,921	0.02%
TK - Tokelau	60,495	0.01%
NZ - New Zealand	55,530	0.01%
UA - Ukraine	48,618	0.01%
EE - Estonia	48,099	0.01%

3 Sites

3.1 General Information

Here we present statistics concerning the sites encountered during the crawl.

Site Summary

The table 3.1 reports on various general quantitative statistics concerning the sites. The top rows concern the average sizes of sites in terms of number of documents and occupied storage space (see section 5.2 for distributions of these characteristics). The next two rows concern general link interconnection statistics (average in-degree is equal to average out-degree in any directed graph, so we show this as „average links/page"). The next four rows report on the temporal aspects of the hosts. The average site depth (the last row) is quite low (see also table 3.2). Notice, that the values of most of the statistics presented in this table are affected by the crawler parameter settings (described in subsection "Crawl parameters", chapter 2) and could be potentially slightly higher without these limitations, but the potential difference is not expected to be significant.

Table 3.1. Site summary

Number of sites OK	151,724
Average pages per site	174.4519
Average static pages per site	101.5612
Average dynamic pages per site	72.8908
Average site size in MB	3.6967
Average links/page	11.1392
Average internal links per host	935.1178
Number of sites with valid page age	65,527
Average of age of oldest page in months	19.3451
Average of age of average page in months	17.3518
Average of age of newest page in months	16.4015
Average site max depth	2.4237

In comparison to the other national Web domains, as described in [4] and [5], one may notice, that the Polish Web has smaller average number of pages per site (about 174) than Italy (410), UK (248) and Korea (224), but bigger than Greece (150), Brazil (66) and Chile (58).

Interestingly, larger Webs tend to be more densely connected than smaller ones, i.e. they do not only have more links, but also more links per node. The same happens with „older" scale-free networks. According to Leskovec et al. [16],

scale-free networks become denser when they grow.

The average site size in the Polish Web is nearly 3.7MB, whereas in the Korean Web it is about 1.1MB (excluding multimedia files). This suggests, that Polish pages are bigger than Korean ones. The average maximum depth of a Web site is quite similar in both Webs: 2.4 for the Polish and 2.2 for the Korean. The maximum crawling depth was set in the same way in both crawls.

Depth of Sites

The site depth distribution is presented in table 3.2. Notice that the distribution is not ideally monotonic (rows 2 and 3).

Table 3.2. Site max depth

Max depth	Sites	Percent
1	64,347	42.4100%
2	24,719	16.2900%
3	26,500	17.4700%
4	22,282	14.6900%
5	9,212	6.0700%
6	1,706	1.1200%
7	939	0.6200%
8	521	0.3400%
9	354	0.2300%
10	212	0.1400%
11	127	0.0800%
12	98	0.0600%
13	77	0.0500%
14	52	0.0300%
15+	578	0.3800%

As we see, about 42% of sites include only one page (the host's home page). According to [5], it is quite a large number compared to the UK (24%) or Indochina, Italy and Greece (about 30%), but is similar to Brazil, Chile and Portugal and less than in Spain (60%).

3.2 Graph Structure

This section deals with the structure of the host graph (the terms „host" and „site" are used interchangeably here, despite the fact they do not necessarily mean the same, in general).

The Web graph has proven to be a very valuable abstract of the Web itself. One of the first studies of the Web graph are [11,12 ,8]. In the latter two, the so called „bow-tie" structure of the Web graph was discovered and analyzed, which is based on (directed) connectivity characteristics of the Web graph such as sizes of the strongly connected components (SCC) and relative sizes of the largest SCC and other categories based on directed connectivity.

The following subsections present connectivity-based characteristics for the Polish host graph collected during the crawl.pl project.

General Statistics

The table 3.3 presents some general statistics concerning connectivity of the hosts that we were able to contact. Notice that for almost 16K of sites (almost 10% of all examined sites) we could not download any page (mostly because the Web sites did not exist, were not responding, or consisted of „redirects" to other Web sites).

The number of sites without in-links is almost twice smaller than the number of sites without out-links. Among sites without in-links are usually the newly created sites which did not attract much interest yet and thus are not linked yet.

Notice the giant largest SCC component - 50.5% of all the sites with valid pages (or 45.72% of all sites). This means that for a majority of valid sites there exists a directed path to and from the most of other sites. The relative size of the largest SCC is similar to those observed in [12,8], but much bigger than that of the Korean Web [4].

On the other hand, almost half of the examined sites constitute a single, isolated nodes (islands). In general, the sizes of SCC usually follow a heavy-tailed distribution, which is the case in the Polish Web, too (see table 3.4).

Table 3.3. General statistics

Total number of site names known	167,604
Sites with at least one page ok	151,724
Sites without in links (but at least one page ok)	26,441
Sites without out links (but at least one page ok)	43,797

Size of largest SCC	76,633
Number of SCCs with one site only (singletons)	71,366

Sizes of Strongly Connected Components

The table 3.4 presents the sizes of strongly connected components.

Table 3.4. SCC sizes

SCC size	Number of SCC components	Percent
1	71,366	98.94%
2	534	0.74%
3	91	0.13%
4	46	0.06%
5	25	0.03%
6	16	0.02%
7	8	0.01%
8	2	0%
9	5	0.01%
10	4	0.01%
11	2	0%
12	1	0%
13	2	0%
14	3	0%
17	2	0%
20	2	0%
21	1	0%
24	1	0%
25	1	0%
31	1	0%
45	1	0%
62	1	0%
64	3	0%
65	1	0%
85	1	0%
86	1	0%

99	1	0%
102	1	0%
209	1	0%
298	1	0%
302	1	0%
76,633	1	0%

Macroscopic Graph Structure

In this section we report the absolute and relative sizes of connectivity-based components defined as follows:

- MAIN - the largest strongly connected component of the site graph
- IN - the sites (not in MAIN) from which there is a directed path to MAIN
- OUT - the sites (not in MAIN) to which there is a directed path from MAIN
- TUNNEL - the sites out of MAIN lying on directed paths from IN to OUT
- TIN, TOUT - the sites out of IN, OUT, TUNNEL and MAIN that could be reached from IN (could reach OUT)
- ISLAND - isolated groups of sites

The MAIN component can be further divided into several subcomponents (after [17]), as follows:

- MAIN_MAIN - the sites that can be directly reached from IN and can directly reach OUT
- MAIN_IN - the sites that out of MAIN_MAIN but that can be directly reached from IN
- MAIN_OUT - the sites out of MAIN_MAIN but can directly reach OUT

The results of measurements are contained in the table 3.5.

Table 3.5. Graph component sizes

Component name	Number of sites	Percent
MAIN_NORM	17,771	11.71%
MAIN_MAIN	18,262	12.04%
MAIN_IN	9,294	6.13%
MAIN_OUT	31,306	20.63%

IN	22,749	14.99%
OUT	34,133	22.50%
TIN	5,366	3.54%
TOUT	1,558	1.03%
TUNNEL	787	0.52%
ISLAND	10,498	6.92%

3.3 The Block Structure

In this section we present an interesting picture of the adjacency matrix of the whole site graph crawled during the project.

On the figure 3.1, the X and Y axes correspond to the id numbers of the sites. On this figure, there is a point of coordinates x and y if and only if there is a link from site x to the site y (or more precisely, there is a link from any document on the site x to any document on the site y).

Initially, the id numbers of the sites where roughly the same as the ordering numbers of the crawled sites. These numbers were, in some way, random so that related sites did not necessarily have neighboring numbers.

Because of this, before drawing the picture, we renumbered the sites as follows. We sorted the sites lexicographically using their URLs, treating the reversed components of the URL as the sort keys. The sequential numbers of URLs sorted in this way served as the new id numbers of the sites. Similar approach was present in [10] and makes the URLs grouped by the top level domain, domain and hostname which makes the closely related URLs (in the terms of their name structure) being the neighbors of each other in the ordering.

After the renumbering, we revealed an apparent block structure of the site adjacency matrix in the Polish site-graph (analogously to the results presented in [10]). Importantly, the block structure was not visible before we renumbered the sites as described above.

The picture of the adjacency matrix renumbered in the way described here seems to be a quite powerful tool for Web graph analysis since one can easily locate on the picture the most dominating blocks of the matrix.

Interestingly, the largest blocks in the matrix are not the leading Polish portals but the most active Polish blog sites. Blog sites are known to be extremely intensively interlinked. The dominating blog-blocks on the picture are (from bottom-left to upper-right):

1. blog.pl (id numbers around 20,000),
2. eblog.pl (id numbers around 50,000),

3. mylog.pl (id numbers around 90,000).

This simple, but powerful visualization technique may be helpful in supporting the detection of spamming sites by quick identification of the largest link-farms (see [9]). This may be very valuable for analysts working for search engines.

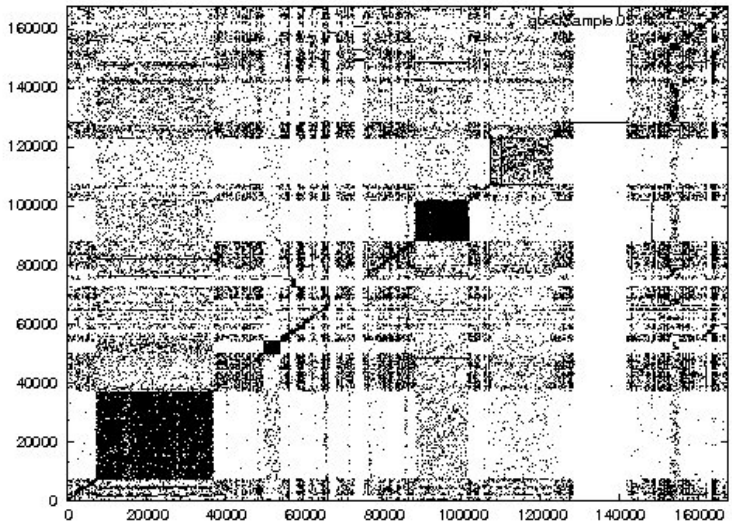

Figure 3.1. The block structure revealed in the site-graph adjacency matrix after appropriate site renumbering (see the text). The most heavily inter-linked sites are major Polish blog sites (from bottom-left to upper right): blog.pl, eblog.pl, mylog.pl (see the text for more details).

4 Documents

4.1 Multimedia

About 1G of image files and about 100K of video files were found, so the relation between them is 1/10M. It is worth noting that in the Korean Web the relation between video and image files in 2004 was 1/500 (see [4]). Within 8M of pages in the Korean Web (December 2004) there are 57M images and 23M of the Polish Web documents include 1G of images. So the ratio of images in the Korean Web is 57/8 ~ 7, whereas in the Polish Web it is much higher - 1000/23 ~ 43.5.

Image Files

The most popular image format in Polish Web is GIF, as one should expect. We found more than 900M of such images (or rather links to them), what constitutes about 90% of all image files (see table 4.1). The second common format is JPG with nearly 10% share which makes nearly 100M files. What seems surprising, the Portable Network Graphics format (PNG) takes only 1.6% with about 16M files, although it was designed to be the main Internet image file format and was conceived as a replacement for GIF.

Table 4.1. Links to image files

File name extension	Number of links found	Percent
Gif	910,804,348	88.57%
Jpg	96,239,368	9.36%
Png	16,524,717	1.61%
Ico	4,382,058	0.43%
Bmp	363,421	0.04%
Img	5,166	0%
Tiff	3,963	0%
Wmf	2,433	0%
Pbm	526	0%

The distribution of image formats is similar to other national Webs analyzed so far (see [5]), except for the African Web, where GIF files take 31% share and JPG take 68% (see [7]).

Video Files

There is no obvious leader among video formats as we may observe by images. The most common type of video files is WMV with 44% share which makes about 48K files. A similar amount is shared by two other formats, by half: MPG (25%) and AVI (23%) (see table 4.2). Note, that the distribution of video formats in the Korean Web, as described in [4], is quite different: here the WMV format dominates all the others with more than 81% share, and MPG takes next 14% leaving only about 3.4% for AVI.

Table 4.2. Links to video files

File name extension	Number of links found	Percent
wmv	48,473	44.18%
mpg	28,239	25.74%
avi	25,141	22.91%
mov	7,871	7.17%
qt	5	0%

Audio Files

Among music files the MP3 is the most common format with 450K of files which make 60% of all the audio files. The other worth noting file types are PLS (18%), RAM (8%) and MID (8%) (see table 4.3). In this case the situation is completely different than that of the Korean Web, because there the MP3 format takes only about 10% and the dominating role play REAL, ASF and WMA. It could be predicted, that the number of mp3 files - and, in consequence, its predominance over the other formats - will grow still more in the close future due to the growing popularity of portable mp3 players.

Table 4.3. Links to audio files

File name extension	Number of links found	Percent
mp3	452,533	60.28%
pls	136,528	18.19%
ram	61,910	8.25%
mid	59,631	7.94%
asf	16,197	2.16%
wav	12,240	1.63%
wma	11,241	1.50%
au	419	0.06%

4.2 Documents

We divided various document formats into two categories: textual documents

(HTML and pure text) and other (not HTML). Here, we treat XML documents as textual ones.

Textual Documents

We found about 435M links to HTML documents and 157K links to text documents (see table 4.4). This means, that each page includes the average of 435/23 = 19 links to HTML documents.

Table 4.4. Links to textual documents

File name extension	Number of links found	Percent
html	435,386,988	99.96%
txt	157,868	0.04%

Non-html documents

The table 4.5 summarizes occurrences of various other types of documents. As it is easily seen, the dominating XML and PDF documents are dominating, with almost 40% share each.

Table 4.5. Other types of non-html documents

File name extension	Number of links found	Percent
Xml	1,263,341	38.42%
Pdf	1,190,982	36.22%
Doc	589,243	17.92%
Xls	93,814	2.85%
Rtf	55,661	1.69%
Ppt	46,617	1.42%
Mso	17,898	0.54%
Ps	11,938	0.36%

Asc	9,784	0.30%
Dvi	4,316	0.13%
Rdf	1,837	0.06%
Tex	1,562	0.05%
Log	693	0.02%
Sgml	219	0.01%

4.3 Software and Programming Languages

Source Files

The C/C++ language source and header files dominate over other programming languages sources with almost 70% share. What may seem strange, is the small number of Java sources - we found only 312 of them. There are also more than 1.5K of Unix/Linux shell scripts (.sh). The table 4.6 includes the details.

Table 4.6. Links to programming languages sources

File name extension	Percent
c/cpp	38.20%
h	29.06%
sh	13.25%
in	10.01%
cc	6.81%
java	2.60%
ada	0.07%

Software-related files

We investigated the amount of various well known file types which are strongly related to operating systems. We found over 258K of Windows executables, about 1.5K ISO images (CD or DVD). There are also almost 8K of Linux software packages (.rpm and .deb).

Table 4.7. Links to software-related files

File name extension	Number of links found	Percent
exe	258,345	94.83%
rpm	5,988	2.20%
patch	3,344	1.23%
deb	1,848	0.68%
iso	1,582	0.58%
diff	994	0.36%
pdb	331	0.12%

Embedded and script languages

The most common language accompanying Web pages is CSS with 23M of links. It means that each Web page has, on average, a 1 link to a CSS file. But, of course, it does not mean that each page actually must have such a link - some pages may have more than one CSS link. The second popular language is Shockwave Flash with almost 25% share (see table 4.8).

Table 4.8. Links to embedded and script source files

File name extension	Number of links found	Percent
css	23,874,594	65.35%
swf	8,557,253	23.42%
pl	3,955,898	10.82%
js	114,132	0.31%
py	30,046	0.08%

4.4 Other file types

Compressed Files

The most popular compression format in Polish Web is zip with nearly 60% share. The second one is its GNU/Unix counterpart - gz - with about 24% share (see table 4.9).

Table 4.9. Compressed files

File name extension	Number of links found	Percent
zip	440,254	57.86%
gz	181,767	23.89%
rar	99,791	13.11%
bz2	15,982	2.10%
tar	15,837	2.08%
lhz	3,153	0.41%
z	2,449	0.32%

4.5 Summary

In general, the statistics presented in the tables above show that the software related to GNU/Linux (like „rpm" or „deb" packages) is much less present that the software related to the proprietary operating systems (like „.exe").

The table 4.10 summarizes shares of the top 50 file types (according to extension). As one may see, links to dynamic Web pages written using PHP or ASP are equally frequent as links to static ones (html or htm).

Table 4.10. Top 50 file extensions

File name extension	Number of links found	Percent
gif	910,804,348	45.93%
php	405,622,679	20.45%
html	380,646,057	19.19%
jpg	95,288,200	4.80%
htm	51,888,454	2.62%
asp	32,212,757	1.62%
css	23,874,594	1.20%
png	16,524,717	0.83%
swf	8,556,804	0.43%
phtml	8,494,058	0.43%
aspx	6,787,442	0.34%
php3	4,613,916	0.23%

cgi	4,533,518	0.23%
ico	4,382,058	0.22%
pl	3,955,898	0.20%
pxf	2,869,280	0.14%
jsp	2,079,184	0.10%
xhtml	1,517,092	0.08%
dhtml	1,335,385	0.07%
dll	1,311,868	0.07%
xml	1,263,341	0.06%
pdf	1,190,982	0.06%
shtml	1,021,512	0.05%
jpeg	946,259	0.05%
cfm	887,264	0.04%
tomcat	794,322	0.04%
pxml	624,698	0.03%
jhtml	604,484	0.03%
doc	589,243	0.03%
mp3	452,533	0.02%
zip	440,254	0.02%
ng	405,289	0.02%
bmp	363,415	0.02%
action	331,174	0.02%
do	325,871	0.02%
php5	315,319	0.02%
tpl	304,778	0.02%
exe	258,345	0.01%
nsf	201,581	0.01%
gz	181,767	0.01%
pww	170,929	0.01%
php4	168,240	0.01%
txt	157,868	0.01%

mspx	143,971	0.01%
tvp	143,113	0.01%
rd	139,990	0.01%
js	114,132	0.01%
spec	103,358	0.01%
rar	99,791	0.01%
spt	96,308	0%

5 Power-Law Distributions

In this chapter, we present some visualizations of various statistics that seem to be governed by (or are close to) the power law - a distribution which is very frequently encountered in Web mining.

In many cases concerning the measurements of real Web-related characteristics, only the tails of empirical distributions follow the „model" distributions known from theory, with the „heads" being irregular.

However, in many examples presented here, even the „heads"[3] of he empirically observed distributions are astonishingly regular which means that the corresponding statistical processes that govern the measured characteristics are particularly strong.

5.1 Power Law

We say that a real, positive random variable has the power-law distribution, iff its density function is as follows:

$$f(k) = \frac{c}{k^g}$$

where k is a positive real number and c is a constant proportionality factor. The parameter g is called the exponent of the distribution. Notice that the density function of any power-law distribution with the exponent g, visualized on a plot

3 Notice that many figures concern only the top 80 hosts out of the all 160 thousand. In these cases, we mean the top 80 hosts with regard to the particular measured quantity (i.e. the „top 80" usually means different sets for different figures)

with logarithmic axes looks as a straight line with negative slope g.

5.2 Sites

The out-degree and the in-degree of the host graph are both power-law distributed (figures 5.1 and 5.2), as is a well known fact reported in previous publications on this subject (e.g. [11,8 ,13]).

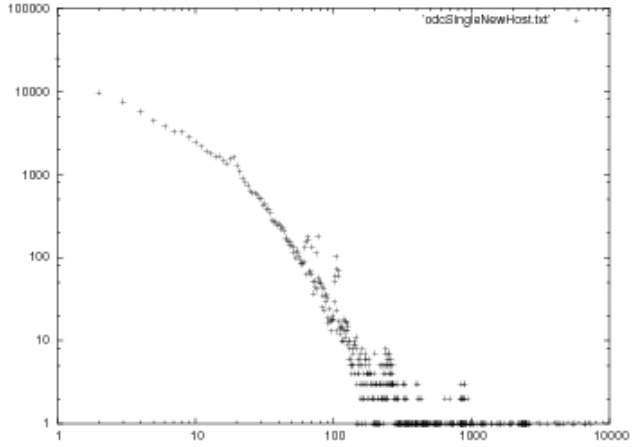

Figure 5.1. Out-degree distribution of the crawled host graph. Multiple links between sites were treated as single. X-axis: value of out-degree, Y-axis: number of nodes having that out-degree

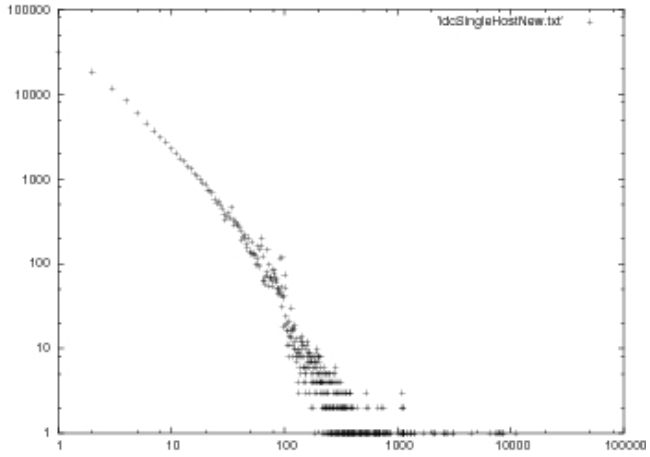

Figure 5.2. In-degree distribution of the crawled host graph. Multiple links between sites were treated as single. X-axis: value of in-degree, Y-axis: number of nodes having that in-degree

The distribution of the raw content length (in Bytes) on the top 80 hosts is depicted on the figure 5.3. Except for the top 10 positions (which are irregularly distributed), the power-law is observable.

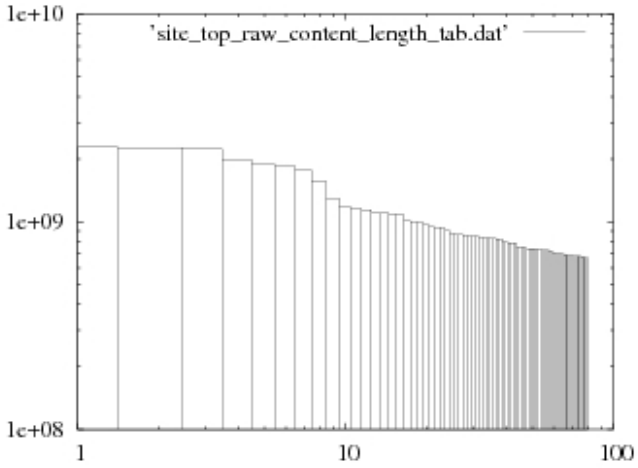

Figure 5.3. Raw content length in Bytes (Y-axis) on the 80 top hosts. Power-law in the range 10-80

The figure 5.4 presents the distribution of the number of internal links (i.e. links that are between documents on the same site) on the top 80 sites. Except for the first 3 positions it is very close to power-law.

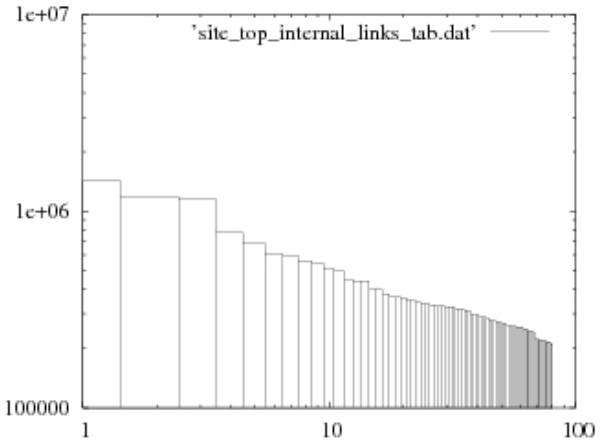

Figure 5.4. Distribution of the number of internal links (Y-axis) on the top 80 sites.

5.3 Page Rank

In this section, we present a set of figures visualising the distributions of PageRank-related quantities.

By siteRank we mean PageRank computed on the graph of hosts (instead on docs). The figure 5.5 presents the distribution of siteRank on the top 80 hosts. It is not ideally power-law distributed but is very close to this distribution.

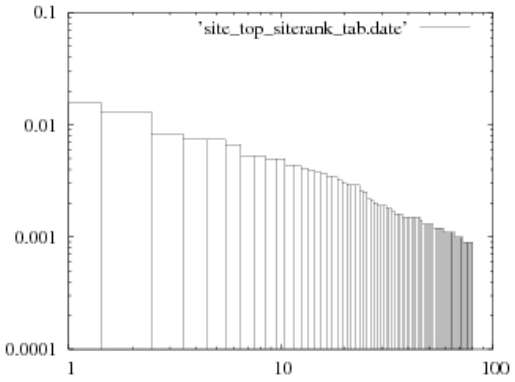

Figure 5.5. SiteRank (Y-axis) on the 80 top hosts. Almost power-law

The figure 5.6 presents the distribution of the PageRank of particular documents accumulated across the whole sites, for the top 80 sites. As in the previous case (fig. 5.5) it does not seem to be ideally power-law distributed but is very close to this.

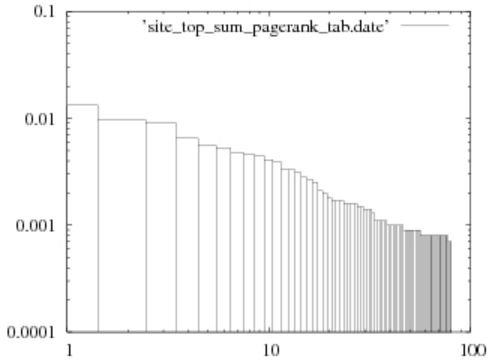

Figure 5.6. PageRank of documents accumulated across the whole sites (Y-axis) for the top 80 hosts

We end this PageRank-related section with an impressive plot of PageRank distribution on the whole host graph (fig. 5.7). Here, the meaning of axes is slightly different than on the other figures. The Y-axis corresponds to the value of PageRank and these values are sorted from left to right for over 160,000 hosts examined during the crawl.

The strong regularity of this distribution (especially for positions higher than 500) is readily observable.

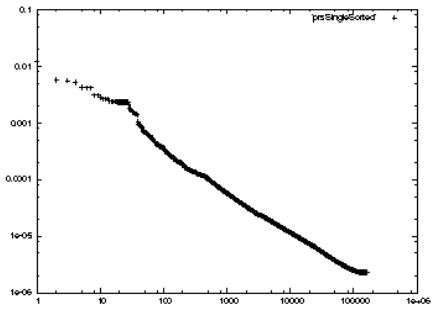

Figure 5.7. HostRank (Y-axis) of all the hosts, sorted inversely from maximal (left) to minimal (right). Very strong regularity above 500th position.

5.4 File Extensions

Finally, we present an interesting graph (fig. 5.8) of the distribution of the volume of the 80 most popular file extensions encountered during the crawl. We can see almost ideal power-law distribution here. The fact of power-law distribution of this quantity does not seem to have been reported in any previous work.

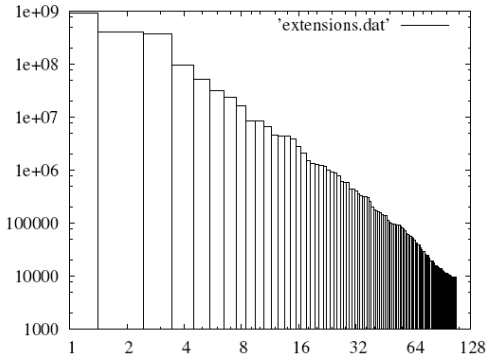

Figure 5.8. Distribution of the volume (number of encountered files) for the 80 most popular file extensions encountered during the crawl. Almost ideal power-law - this fact does not seem to have been reported in any previous work

6 Conclusions and Further Work

6.1 Conclusions

In this paper, we have reported measurements of various statistical properties of the Polish Web and related them to the similar reports concerning other countries or regions.

The „crawl.pl" project seems to be the first research project of this kind in Poland, up to the knowledge of the authors. Its current status and results are in a preliminary stage but the project is intended to be continued.

Despite the preliminary stage of the work we report some observations that might be interesting for other researchers or analysts and that have not been reported before. In addition, the project gave much experience to the authors, and this experience will be invaluable in making the future continuation of similar experiments more perfect.

An important outcome of the project are the first datasets concerning the Polish Web graph, which will be publicly available for researchers.

6.2 Further Work

It would be very valuable to record consecutive snapshots of the Polish Web in regular time intervals (e.g. each 3 months) in order to grasp the temporal aspects of the evolution of the Polish Web.

Also, in future crawls, it would be very valuable to provide enough storage devices to record the contents of documents, rather than only the link structure.

Another direction of development is to increase the scale of the crawls to make the analyses and datasets more representative.

7 Appendix. The Harvest Log Extracts.

Table 7.1. Summary table

Batch	Begin	Active	Number	Docume	Bytes
1	2005/12	167,60	167,604	0	65,650,41
2	2005/12	157,29	157,385	146,429	2,475,252,
3	2005/12	62,970	300,000	183,327	4,154,373,
4	2005/12	51,381	300,000	214,611	876,620,7
5	2005/12	48,882	300,000	208,162	738,111,3
6	2005/12	45,632	300,000	204,115	647,289,3
7	2005/12	44,148	300,000	202,194	785,448,2
8	2005/12	42,369	300,000	180,953	136,039,4
9	2005/12	40,275	300,000	174,179	39,351,27
10	2005/12	37,314	300,000	185,465	664,744,6
11	2005/12	32,934	300,000	174,127	167,187,4
12	2005/12	23,575	300,000	199,074	1,434,906,
13	2005/12	24,095	300,000	197,603	1,045,270,
14	2005/12	22,847	300,000	229,492	2,212,036,

15	2005/12	22,776	300,000	207,912	1,539,974,
16	2005/12	22,469	300,000	206,620	1,459,116,
17	2005/12	22,284	300,000	228,291	2,267,132,
18	2005/12	21,930	300,000	200,267	1,165,328,
19	2005/12	21,334	300,000	226,466	2,235,645,
20	2005/12	21,490	300,000	234,026	2,417,583,
21	2005/12	21,630	300,000	195,553	1,089,897,
22	2005/12	21,475	300,000	217,018	1,936,427,
23	2005/12	21,460	300,000	230,770	2,331,747,
24	2005/12	21,352	300,000	199,262	1,145,550,
25	2005/12	21,337	300,000	221,516	2,113,128,
26	2005/12	21,051	300,000	233,074	2,349,272,
27	2005/12	128,59	300,000	141,346	3,654,637,
28	2005/12	21,088	300,000	226,597	2,276,719,
29	2005/12	20,522	300,000	227,831	2,356,504,
30	2005/12	20,144	300,000	202,909	1,326,274,
31	2005/12	19,742	300,000	202,897	1,336,429,
32	2005/12	19,515	300,000	226,527	2,368,132,

33	2005/12	19,159	300,000	216,306	1,925,417,
34	2005/12	18,289	300,000	215,141	2,035,218,
35	2005/12	17,351	300,000	204,495	1,718,407,
36	2005/12	13,766	300,000	202,294	1,693,001,
37	2005/12	11,657	300,000	197,563	1,404,723,
38	2005/12	11,823	300,000	213,319	2,135,496,
39	2005/12	12,075	300,000	211,742	1,889,343,
40	2005/12	12,195	300,000	213,992	1,992,910,
41	2005/12	12,254	300,000	191,079	1,246,054,
42	2005/12	12,308	300,000	200,263	1,611,113,
43	2005/12	12,459	300,000	208,727	1,877,902,
44	2005/12	12,543	300,000	197,759	1,502,432,
45	2005/12	12,578	300,000	215,076	2,115,086,
46	2005/12	12,647	300,000	208,974	1,954,915,
47	2005/12	104,92	300,000	152,522	176,218,8
48	2005/12	12,466	300,000	215,388	2,181,180,
49	2005/12	12,562	300,000	203,435	1,685,745,
50	2005/12	12,543	300,000	206,948	1,901,978,

51	2005/12	12,475	300,000	213,198	2,080,471,
52	2005/12	12,612	300,000	216,600	2,250,642,
53	2006/01	12,565	300,000	209,763	2,059,524,
54	2006/01	12,623	300,000	219,055	2,322,300,
55	2006/01	12,622	300,000	222,886	2,475,503,
56	2006/01	12,536	300,000	210,051	2,049,666,
57	2006/01	12,555	300,000	190,660	1,394,439,
58	2006/01	12,358	300,000	204,071	1,980,593,
59	2006/01	12,296	300,000	188,903	1,169,735,
60	2006/01	12,263	300,000	195,891	1,688,662,
61	2006/01	12,217	300,000	206,470	2,020,595,
62	2006/01	12,245	300,000	190,887	1,258,716,
63	2006/01	12,033	300,000	198,093	1,808,265,
64	2006/01	11,518	300,000	179,561	1,014,909,
65	2006/01	8,606	300,000	185,326	975,744,7
66	2006/01	99,732	300,000	156,411	4,211,090,
67	2006/01	7,615	300,000	185,391	775,464,6
68	2006/01	7,557	300,000	175,047	451,400,9

69	2006/01	7,649	300,000	198,415	1,281,454,
70	2006/01	7,648	300,000	191,410	1,043,304,
71	2006/01	7,704	300,000	187,937	711,385,8
72	2006/01	7,584	300,000	184,585	712,313,3
73	2006/01	7,744	300,000	203,067	1,351,115,
74	2006/01	7,746	300,000	193,929	1,035,501,
75	2006/01	7,711	300,000	187,211	842,553,8
76	2006/01	7,701	300,000	192,426	1,105,713,
77	2006/01	7,714	300,000	200,610	1,134,986,
78	2006/01	7,699	300,000	188,390	700,031,2
79	2006/01	7,736	300,000	198,900	1,188,992,
80	2006/01	7,620	300,000	207,233	1,527,825,
81	2006/01	7,562	300,000	172,551	108,962,2
82	2006/01	7,364	300,000	174,530	482,126,0
83	2006/01	6,968	300,000	207,128	1,246,895,
84	2006/01	4,327	300,000	199,912	99,381,35
85	2006/01	141,29	300,000	135,429	3,068,899,
86	2006/01	69,965	300,000	211,456	792,226,0

87	2006/01	116,40	300,000	161,014	4,142,381,
88	2006/01	43,854	300,000	177,318	435,631,9
89	2006/01	42,854	300,000	177,284	287,398,0
90	2006/01	42,070	300,000	181,734	478,578,8
91	2006/01	40,177	300,000	180,628	344,066,7
92	2006/01	36,602	300,000	155,949	81,608,05
93	2006/01	23,940	300,000	187,563	1,226,310,
94	2006/01	22,304	300,000	219,595	2,320,425,
95	2006/01	21,964	300,000	208,207	2,274,692,
96	2006/01	21,937	300,000	216,747	2,323,833,
97	2006/01	21,310	300,000	181,716	1,031,329,
98	2006/01	9,932	300,000	167,988	434,526,7
99	2006/01	5,651	300,000	1	9,544
100	2006/01	4,683	300,000	1	7,884
101	2006/01	4,054	300,000	0	0
102	2006/01	5,708	300,000	157,972	13,223,66
103	2006/01	120,96	300,000	137,508	2,727,175,
104	2006/01	2,656	300,000	195,627	2,546,782,

105	2006/01	3,012	300,000	210,696	3,414,631,
106	2006/01	2,860	300,000	186,500	1,722,390,
107	2006/01	2,812	300,000	173,409	1,142,994,
108	2006/01	3,418	300,000	183,121	668,973,7
109	2006/01	3,048	300,000	122,910	3,039,452,
110	2006/01	3,072	300,000	172,955	75,289,54
111	2006/01	2,302	263,596	168,539	4,261,861,
112	2006/01	1,735	218,677	135,477	3,664,487,
113	2006/01	1,365	189,595	96,366	2,721,832,
114	2006/01	1,178	168,960	96,149	2,732,151,
115	2006/01	1,051	153,616	74,903	2,274,763,
116	2006/01	963	143,785	85,082	2,402,439,
117	2006/01	868	129,007	76,646	2,114,823,
118	2006/01	599	109,837	58,487	1,662,140,
119	2006/01	273	162,604	64,648	1,844,998,
120	2006/01	217	146,798	52,465	1,390,294,
121	2006/01	180	121,718	30,320	833,215,2
122	2006/01	141	102,004	20,749	555,194,7

123	2006/01	123	97,711	19,466	509,971,2
124	2006/01	108	91,228	13,028	315,992,1
125	2006/01	101	87,375	9,769	227,865,0
126	2006/01	97	85,136	4,249	96,517,65
127	2006/01	86	71,883	8,957	213,656,7
128	2006/01	92	83,703	4,196	99,511,70
129	2006/01	82	70,988	7,036	163,974,0
130	2006/01	89	79,823	13,151	297,649,6
131	2006/01	87	79,078	18,947	438,090,6
132	2006/01	85	75,868	23,822	549,322,7

References

1. *http://www.netsprint.pl/serwis/*.
2. *http://www.users.pjwstk.edu.pl/~msyd/PolishWebDatasets.html*.
3. Baeza-Yates R., Castillo C., (Nov 2000), *Characterizing the chilean web*, Chilean Computer Science Congress, Santiago, Chile.
4. Baeza-Yates R., Lalanne F., (2004), *Characterization of the Korean Web,* Technical report.
5. Baeza-Yates R., Castillo C., Efthimiadis E., (2006), *Characterization of national web domains,* ACM TOIT.
6. Baeza-Yates R., Castillo C., Efthimiadis E., (2004), *Comparing the characteristics of the chilean and the greek web*, Technical report.

7. Boldi P., Codenotti B., Santini M., Vigna S., (2002), *Structural properties of the african web,* In: Proceedings of the 11th International WWW Conference(11), Honolulu, Hawaii, USA.
8. Broder, Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., Wiener J., (2000), *Graph structure in the web*, In: Proceedings of the 9th WWW Conference.
9. Gyongyi Z., Garcia-Molina H., (2005), *Web spam taxonomy*, In: First International Workshop on Adversarial Information Retrieval on the Web.
10. Kamvar S., Haveliwala T., Manning C., Golub G., (2003), *Exploiting the block struture of the web for computing pagerank,* In: Stanford University Technical Report.
11. Kleinberg J., Kumar R., Raghavan P., Rajagopalan S., Tomkins A., (1999), *The web as a graph: measurements, models and methods,* In: Proceedings of the 5th Annual International Computing and Combinatorics Conference.
12. Kumar R., Raghavan P., Rajagopalan S., Sivakumar D., Tomkins A., Upfal E., (2000), *The Web as a graph*. In: Proc. 19th ACM SIGACT-SIGMOD-AIGART Symp. Principlesof Database Systems, PODS, pages 1-10. ACM Press, 15-17.
13. Huberman B., Adamic L., (2000), *Power law distribution of the world wide web,* Technical comment. Science, 287.
14. Bayeza-Yates R., Castillo C. (2006), *Relationship between links and trade,* In Proceedings of the 15th World Wide Web Conference (posters), Edinburgh, Scotland, May 2006., pages 927-928.
15. Fetterly D., Manasse M., Najork M., (2004) *On the evolution of clusters of near-duplicate web pages,* Journal of Web Engineering, 2(4):228-246.
16. Leskovec J., Kleinberg J., Falutsos C., (2005), *Graphs over time: densification laws, shrinking diameters and possible explanations*. In: KDD'05: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 177-187, New York, NY, USA.
17. Saint-Jean F., Baeza-Yates R., Castillo C., (2003), *Web dynamics, structure, page quality,* In: Proceedings of the 12th International WWW Conference, Workshop on Algorithms and Models for the Web Graph.