Distinctive Features of the Argentinian Web

Gabriel Tolosa
Universidad Nacional de Luján - DCB
Laboratorio de Redes
tolosoft@unlu.edu.ar

Fernando Bordignon
Universidad Nacional de Luján - DCB
Laboratorio de Redes
bordi@unlu.edu.ar

Ricardo Baeza-Yates
Yahoo! Research Latin America
Centro de Investigación de la Web (CIW) - DCC
Universidad de Chile
ricardo@baeza.cl

Carlos Castillo
Yahoo! Research Barcelona
chato@yahoo-inc.com

Abstract

This article presents the most distinguishing features of the Argentinian web as found in a private sample of almost 10 million web pages from 150.000 sites collected in the early 2006. Particularly, we have studied page contents, link structure and technologies used in the construction of the sites.

This study reveals a number of interesting facts: To begin with, there is a significant proportion (97.6%) of ".com.ar" domains with respect to other second level domains. As regards page contents, we have found a predominance of terms related to commercial activity while terms found in site names are mostly related to tourism. As for technologies, 48% of the pages from the sample are static and 52% dynamic, the latter being mostly built using free tools (like PHP). Besides, 76% of the sites are hosted in servers geographically located in Argentina. These two facts show there is an important web-related technological development and communication infrastructure in Argentina.

1. Introduction

The World Wide Web is a wide publication medium used by millions of people for different purposes such as commerce, advertising, education, entertainment and social interaction. As it is constantly growing, the study of its structure and contents provides valuable information to researchers, developers and users.

Generally, the Web is modeled as a directed graph G = (V, E) with a set V of vertices (web pages) and a set E of links between pages [Broder, 2000]. Each edge (denoted by q→p) is an ordered pair (p, q) where q, p ϵ V and represents a hyperlink between pages (nodes) q and p. This situation occurs only with certain pairs of pages. The webgraph encodes a considerable amount of latent information about the underlying structure of the Web which could be used for multiple purposes. As an example, link structure is the basis of the PageRank algorithm [Page, 1998] used by Google to rank search results.

The topology of the web graph has been studied in detail and one of its most interesting features is that it is not random but forms a scale-free network which is characterized by an uneven distribution of nodes and links [Albert, 2002]. In this topology, the number of links in the nodes follows a power law distribution, where $P(x = k) \approx k^{-\beta}$, with $\beta > 0$. This yields the probability that a page x has k links [Kleinberg, 1999] [Barabasi, 1999].

Besides its size and structural complexity, the web presents another challenging issue: contents are created and managed by millions of distributed autonomous users from different organizations (or individuals) who come from different geographical, historical and cultural contexts, resulting in high heterogeneity.

In this situation, the identification of the most important features of the Web, such as its link structure, could provide better insight into the way it operates. For this purpose, researchers have made many efforts on the characterization of the global Web [WCA, 1999; O'Neill, 2003] and small-scale studies of national web domains [Baeza-Yates, 2004; Baeza-Yates, 2005_a; Baeza-Yates, 2005_b; Efthimiadis, 2004; Modesto, 2005]. Their main goal is to gain better understanding of the structure and contents of the Web in order to find behavior patterns and trends which would enable the development of new strategies to make access to resources easier.

Continuing this line of research, this article presents the most distinguishing features of the Argentinian web. The results presented here are part of a larger work which is, to the best of our knowledge, the first study focused on the Argentinian web domain. In this research, we cover the main features reported in similar works and provide

further details on other aspects as well. Finally, an aggregated value estimation of its size has been included. To conclude, an estimation of the size of the studied web is included as a bonus.

This paper is organized as follows: Section 2 presents the methodology used to obtain a sample of the Argentinian web and some aspects of the collection. In section 3 we present the analysis and obtained results. Based on these results, we go on to make a projection of the size of the Argentinian web in section 4. Finally, we provide our conclusions and propose future works.

2. Methodology

We used the WIRE crawler [Castillo, 2005] for the web crawling phase, which was performed during March and April 2006. Only pages under the ".ar" first level domain were collected, starting from an initial seed of about 10,000 URLs from the Argentinian domain obtained from local directories, official governmental pages and the Yahoo! Directory. Although it is widely known that some organizations use the ".com" domain for their web sites, it is not technically simple to obtain the complete list of these cases. In Argentina there are no restrictions on obtaining a domain name within the ".com.ar" subdomain and registrations are free of charge. As this facilitates the acquisition of domain names, we supposed most of the commercial Argentinian web is under the ".com.ar" domain. This causes some new problems, which will be described later on.

The analysis of the sample was done following the methodology proposed in [Baeza-Yates, 2005$_c$], studying many aspects of the web such as contents, links and technologies at different levels of granularity (pages, sites and domains). In some cases, we have included additional data which reveal other distinctive features of the domain.

2.1. The WebAR Collection

The software collected a sample of 9,656,218 pages from 149,305 sites which belong to 83,813 different domains. According to official information from NIC Argentina [Vilas, 2006], there exist 1,129,381 registered domain names but only 26% of ".com.ar" domain names (286,635) are active (from an internal NIC study). This occurs, in part, due the fact that, as we mentioned earlier, there are no restrictions on obtaining domain names within ".com.ar" and registrations are free of charge. Table 1 shows the official data of second level domains whereas Table 2 indicates the number of third level domains, fitted to the productivity percentage in ".com.ar".

2nd Level Domain	Number	%
com.ar	1,102,444	97.61
org.ar	14,133	1.25
net.ar	10,112	0.90
gov.ar	2,570	0.23
mil.ar	92	0.01
int.ar	30	0.00
Total	1,129,381	100

Table 1: Distribution of second level domains. Source: NIC Argentina, February, 2006

2nd Level Domain	Active 3rd Level Domains (NIC)	Active 3rd Level Domains (Sample)	Sample %
com.ar	286,635	77,668	27.10
org.ar	14,133	3,846	27.21
net.ar	10,112	817	8.08
gov.ar	2,570	896	34.86
mil.ar	92	21	22.83
int.ar	30	11	36.67
edu.ar (*)	N/A	554	
Total	313,572	83,813	

Table 2 - Sample composition. (*) The "edu.ar" domain is not under the administration of NIC Argentina

3. Analysis and Results

This section presents the analysis and results which show the most distinguishing features of the Argentinian web, selected from a much larger study

3.1. Page Size

We observed an average page size of 10 Kb in the sample, which is considerably smaller compared with the samples of Chile (21 Kb) and Brazil (24 Kb). The distribution of page sizes is highly skewed and can be modeled by a power-law distribution with parameter $\beta = 2.2$ for page sizes greater than 20Kb. Figure 1 shows this distribution and its corresponding fit curve. The cut at 100 corresponds to a configuration parameter in the crawling process.

3.2. Most Frequently Used Terms

We randomly selected a subset of web pages from the sample to analyze its contents. In total, we parsed 396,134 documents to remove HTML tags and extract all the terms from their plain text. Then, we computed term frequency for each page and kept the 40 most frequent on a list. Next, we merged all lists after removing stopwords (both in Spanish and English) and one-character terms.

Finally, we computed DF (Document Frequency), i.e. the number of documents in which each term was present, disregarding TF (Term Frequency) within each document. Table 3 shows the first 10 terms, sorted by DF. In these results we can observe that the first terms are related to commercial activities, typical of sites dedicated to massive sales, auctions, on-line catalogs with redirections to other sites and the like.

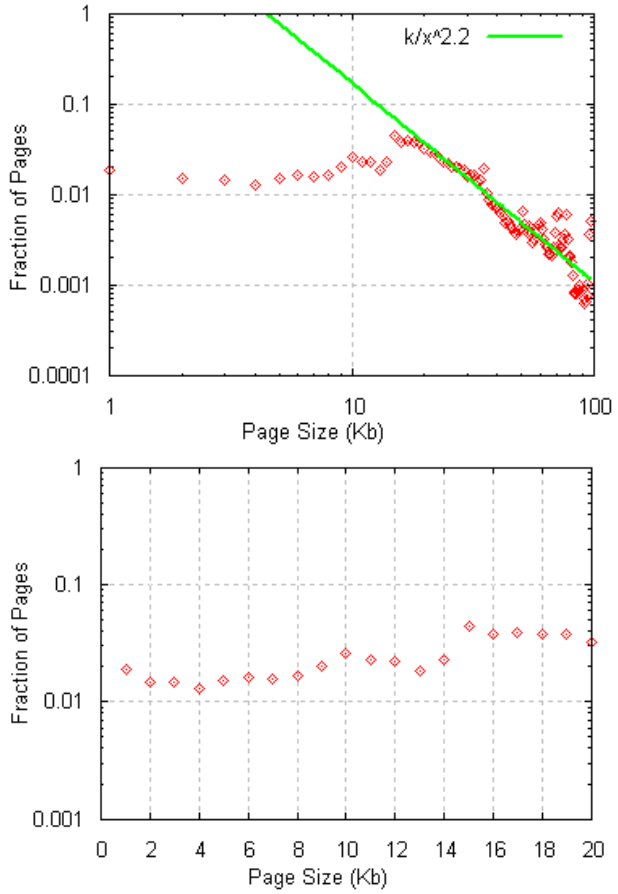

Figure 1 – Distribution of page sizes with fit curve (top). Zoomed-in portion of the graph for sizes less than 20 Kb (bottom)

3.3. Terms in Site Names

As a complement to the analysis of terms in page contents, we studied the most frequently used terms for site and domain names as found in URLs. Neither the labels of the first and second level domains nor the term "www", commonly used as the main web server name, were taken into account for the analysis. As an example, in the site name "www.tyr.unlu.edu.ar" we kept only the "tyr.unlu" substring. Besides, we considered the dot (".") and the hyphen ("-") as term separators. Table 4 lists the first 10 most frequent terms in site names.

In contrast to the contents analysis where most terms were related to massive sale activities, within sites names there are many terms concerning tourism (shown in italics). An interesting finding is that several sites have a domain name formed by one term concatenated to the word "Argentina". For instance, hoteleinrgentina, hotelesargentina. Within the top 100 most used terms, the word Argentina appears 12,468 times (4.6%).

Rank	Term	Number of Documents	%
1	precio	67,966	17.16
2	compra	67,456	17.03
3	inicio	60,362	15.24
4	artículos	59,831	15.10
5	venta	58,930	14.88
6	argentina	56,944	14.37
7	cuotas	50,047	12.63
8	tarjeta	49,926	12.60
9	comprar	46,824	11.82
10	pagofacil	46,729	11.80

Table 3 – First 10 most frequent terms used in document contents

Rank	Term	Sites
1	*campings*	51,318
2	*sbviajes*	21,922
3	*argentina*	7,384
4	*tango*	7,256
5	*europa*	6,835
6	*brasil*	6,472
7	*aereos*	6,215
8	*paquetes*	6,194
9	noticias	6,174
10	ofertas	6,103

Table 4 – First 10 most frequent terms used in site and domain names

3.4. Sites and Pages per Domain

We found 149,305 sites which correspond to 83,813 third level domains. A large number of sites belong to only one domain, that is, to a unique organization, and there is a high proportion of domains with only one site as well. This seems rather striking for the following reasons: first, there exist few organizations with internal divisions and hence administratively separated web sites (for instance, fi.uba.ar and fceyn.uba.ar). Second, those organizations with only one administrative unit maintain the internal structure of the site using other strategies such as divisions in physical directory structure. This information is shown in Table 5. In addition, we found 99 domains in the sample with more than 50 sites each and only 9 which do not belong to ".com.ar" (Table 6).

Then, we grouped the web pages by second level domain as the use of these domains is a common practice in Argentina in about 95% of the cases. However, there are some exceptions like the Universidad de Buenos Aires (uba.ar) and some government organizations (educ.ar, nic.ar, nacion.ar), but they account for a small fraction. Table 7 summarizes this information together with the number of downloaded pages in each domain.

Here, it is important to remember that registration rules in Argentina are not restrictive enough for commercial domains (".com.ar") and they are totally free of charge. Some years ago there were no limitations on the registration and anyone could register any available domain. Now, there is a restriction which requires a valid DNS server in a web hosting provider to resolve domain names, partly preventing massive registrations. This fact is the main reason for the existence of more than 1,000,000 registered domains where only 26% are active. In the case of other second level domain names (".edu.ar", ".gov.ar"), there are strict registration rules whereupon such domains are only assigned to qualified organizations although registration is still free of charge.

	Number	%
Sites found in the sample	149,305	
Thirth level domains	83,813	
Thirth level domains with more than one site	2,389	2.85
Thirth level domains with only one site	81,424	97.15

Table 5 – Third level domains

3.5. Link Structure

3.5.1. Strongly Connected Components

The distribution of Strongly Connected Components (SCC) of the Hostgraph was studied too. An SCC is a directed subgraph in which all nodes can reach the others (inside the subgraph) following the direction of the links. We studied the SCCs in our sample of the Argentinian web and found a giant component (Table 8). The distribution of the sizes follows a power-law with parameter $\beta = 2.74$ in its central portion (Figure 2).

Domain	Sites
gba.gov.ar	139
utn.edu.ar	132
unlp.edu.ar	118
esc.edu.ar	109
unlu.edu.ar	95

Domain	Sites
unc.edu.ar	80
fi.uba.ar	59
mendoza.gov.ar	59
fcen.fcen.uba.ar	52
uns.edu.ar	47

Table 6 – Non-commercial domains with the largest number of sites

2nd Level Domains	Sites	% of Sites	Downloaded Pages
com.ar	140,533	94.1248	9,077,243
org.ar	4,155	2.7829	276,393
edu.ar	1,784	1.1949	114,965
gov.ar	1,516	1.0154	145,719
net.ar	976	0.6537	14,543
uba.ar	239	0.1601	29,637
mil.ar	60	0.0402	2,359
educ.ar	12	0.0080	2,927
int.ar	11	0.0074	188
retina.ar	6	0.0040	742
mecon.ar	6	0.0040	143
nic.ar	3	0.0020	61
sld.ar	1	0.0007	1
promocion.ar	1	0.0007	3
nacion.ar	1	0.0007	16
gobiernoelectronico.ar	1	0.0007	76
TOTAL	**149,305**	**100**	**9,664,921**

Table 7 – Number of downloaded documents by second level domain

Size of SCC	Number of components	Size of SCC	Number of components
1	66.021	12	2
2	432	14	1
3	81	16	2
4	164	20	1
5	18	21	1
6	9	22	1
7	8	23	1
8	1	29	1
9	4	38	1
10	2	44	1
11	2	80.968	1

Table 8 – Sizes of strongly connected components

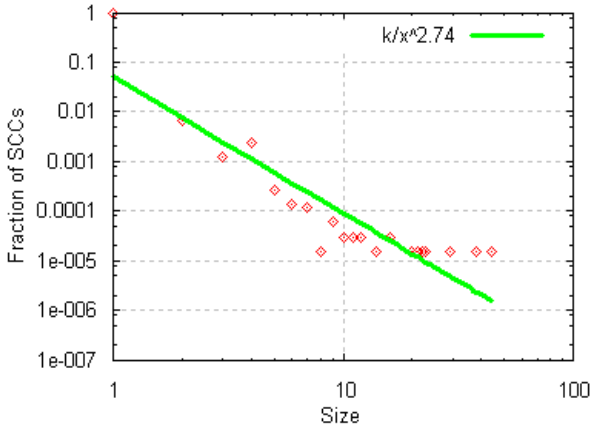

Figure 2 – Distribution of the SCCs sizes and its fit curve

3.5.2. Macroscopic Structure

In an important study of the web on a global scale applying a graph model, Broder et al. [Broder, 2000] proposed a structure known as "bow-tie", which shows the relationships between each web page and the biggest Strongly Connected Component (SCC). This model enables the location of each page in one of six regions: MAIN, includes the biggest SCC (remember that its internal nodes can reach the others by following links); IN, comprises all the nodes which can reach those inside MAIN but cannot be reached the other way round; OUT is the subset of nodes reached from any node in MAIN but which cannot reach them; ISLANDS corresponds to disconnected nodes, i.e. those which do not have any links to the mentioned components; TENTACLES (nodes which are reached only from portions of IN or OUT) and TUNNELS (nodes from IN that reach OUT directly).

This structure was extended in [Baeza-Yates, 2001] where further detail was provided on the MAIN component by splitting it into four subregions: MAIN-MAIN is the set of nodes which are reached directly from IN or can directly reach nodes in OUT, MAIN-IN contains nodes which are reached from IN but do not belong to MAIN-MAIN, MAIN-OUT contains nodes which reach those in OUT but do not belong to MAIN-MAIN, MAIN-NORM, comprises the remaining nodes.

The size of the MAIN region (54.23%) reveals that the Argentinian web is generally well connected, especially if we compare it with other countries such as Chile, where this portion accounts for 21.76%, or Brazil, with 25.27% of the nodes. Sites inside the OUT component (28.15%) represent a smaller portion with respect to Brazil (45.33%) but is similar to Chile's (26.12%). If we consider that two of the reasons that generate the migration of a node to the OUT component are its age and the lack of update we can see that in Argentina this proportion is not greater than those of other countries.

On the other hand, sites located in the IN and ISLAND components are only accessible from their home pages as they can either be new pages or be badly connected. In this case, proportions are similar to Chile's as regards the IN component (6.65%) but not the ISLANDS (46.16%). Figures in Brazil are 12.95% and 12.35% respectively. The low proportion of elements in the ISLANDS component reinforces our idea of good connectivity in the web of Argentina.

3.6. Static vs. Dynamic Pages

For this study we divided downloaded documents into two groups, trying to identify successfully those considered "dynamic pages". These are HTML pages which are not stored on a web server's hard disk but instead created on the fly by a program and later sent to a web client. In order to detect dynamic pages we used two simple criteria. Firstly, we checked whether a document's extension belonged to any of the most common scripting languages (like PHP, ASP, CGI, etc). Secondly, we selected all URLs including the "?" character, which generally introduces script parameters when the GET method of the HTTP protocol is used. Although some pages may have been misclassified using these criteria, we believe they represent minor exceptions which do not affect the final results.

After the analysis of dynamic and static documents, we found a surprising parity (Table 9) with a slight superiority of dynamic pages (52%). This proportion is somewhat high compared to figures in Spain and Chile, where dynamic pages account for 22% and 38% respectively. These values suggest that there is an important infrastructure for web development in Argentina to support the business logic of many organizations. Another possible explanation lies in the fact that both the Argentinian and Chilean web are newer than the Spanish web, which is why dynamic technologies are so common.

Finally, we analyzed the distribution of links to documents whose extensions are used to build dynamic pages (Figure 3). Here, there is an important proportion of the hypertext pre-processing language PHP with 52% followed by Perl with 39%. Both tools are free of charge for development and use. In the web of Spain there is a 46% of PHP use followed by ASP (44%), while in Chile there PHP represents 78% and ASP only 16%. In Brazil, figures are more than 70% and 20% respectively. In these three countries the use of Perl language is low.

	Documents	%
TOTAL	12,276,090	100.00
Dynamic	6,383,050	52.00
Static	5,893,040	48.00

Table 9 – Distribution of static and dynamic documents

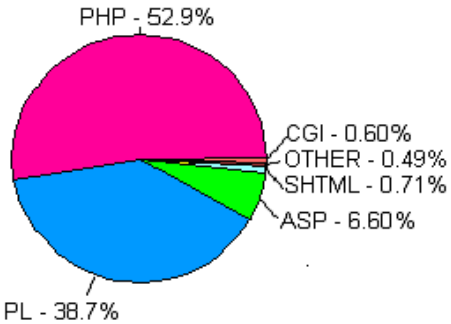

Figure 3 – Distribution of links to dynamic pages

3.7. Relationship between Site Names and Network Addresses

From the total number of sites, we randomly selected a sample of 23,965 (16.05%) sites to analyze the relationship between site names and the network addresses (IP addresses) of the web servers that hosted those sites. Results are presented in Table 10. As shown in the above table, there are few IP addresses (only 16, which represent 0.48% of the addresses in the sample) in which 10,670 sites (44.52% out of the total) are hosted.

3.8. Distribution of Sites by Country

Using the same sample as in the previous section, we evaluated the geographic location of each IP address to find the country each web site from the ".ar" domain was hosted in. To establish the relationship between IP addresses and countries we used the GeoIPCountryWhois database by the Maxmind (http://www.maxmind.com/) company (Table 11).

Category	Addesses	%	Sites	%
100 or more	16	0.48	10,670	44.52
from 50 to 99	12	0.36	778	3.25
from 10 to 49	342	10.29	6,394	26.68
from 1 to 9	2,953	88.87	6,123	25.55
Total	3,323	100.00	23,965	100.00

Table 10 – Relationship between site names and network addresses

Country	Sites	%
Argentina	18,177	75.87
United States	4,700	19.62
Canada	351	1.47
Brazil	224	0.94
Colombia	150	0.63
Spain	89	0.37
France	84	0.35
United Kingdom	60	0.25
Israel	48	0.20
Lithuania	39	0.16
Chile	6	0.03
Germany	5	0.02
Other	24	0.10
TOTAL	23,957	100.00

Table 11 – Distribution of sites by country

From the data presented in Table 11 we conclude that almost 76% of the sites are hosted in web servers located in Argentina. We believe this is another indicator of the technological development of this country. On the other hand, there are significant economic limitations for Argentinians to hire hosting services abroad owing to the unfavorable rate of exchange between pesos and dollars or euros.

4. Projection of the Size of the Web of Argentina

From the combination of all the data collected by the WIRE crawler module and the data provided by NIC Argentina regarding domain name registrations and "active" sites under the ".com.ar" domain, we made a projection of the number of sites, total size and number of pages in the Argentinian web domain. This projection was divided by second level domain according to the data from Section 2.1. Furthermore, we computed the average number of documents per site and the average size of a site (Table 12a y 12b).

Domains	Number of third level domains (NIC)	Number of third level domains (sample)	%
com.ar	286,635	77,668	27.10
org.ar	14,133	3,846	27.21
net.ar	10,112	817	8.08
gov.ar	2,570	896	34.86
mil.ar	92	21	22.83
int.ar	30	11	36.67
edu.ar	Not available	554	
TOTAL	**313,572**	**83,813**	

Table 12a – Site statistics divided by second level domain

Domains	Number of sites (sample)	Average number of documents per site	Average size of a site (in MB)
com.ar	140,533	134	1.820
org.ar	4,155	102	1.645
net.ar	976	34	0.218
gov.ar	1,534	163	2.179
mil.ar	60	52	0.538
int.ar	11	23	0.172
edu.ar	2,036	127	1.293
TOTAL	**149.305**		

Table 12b – Site statistics (cont.) divided by second level domain

Projections were made using the information above and preserving proportions, yielding the results presented in Table 13. For this analysis, it should be noted that the crawler was configured to limit the maximum number of pages per site and maximum crawling depth, so many pages remained uncollected.

For the calculations, we have assumed that the remaining uncollected sites exhibit similar behavior (as regards the number of pages and size). This is probably an optimistic projection since the set of uncollected sites

appears to be smaller and less visible than the collected sample, at least with the strategy and initial seed used in our crawler.

	Projection		
Domains	**Sites**	**Documents**	**Size (MB)**
com.ar	518,639	69,476,127	943,811
org.ar	15,268	1,553,961	25,119
net.ar	12,080	414,666	2,633
gov.ar	4,400	717,009	9,587
mil.ar	263	13,603	141
int.ar	30	676	5
edu.ar	2,036	258,859	2,632
TOTAL	**552,717**	**72,434,902**	**983,929**

Table 13 – Projection of the size of the Argentinian web

5. Conclusions

This work presents the most distinguishing features of the Argentinian web as found in a private sample of 9,656,518 html pages collected from 149,305 sites which amount to 83,813 third level domains. In this research, we covered features concerning page contents, link structure and technologies, comparable to indicators in similar works, and provided other revealing details.

There exists a high proportion of sites under the ".com.ar" second level domain, even if only active sites were considered (26%), as reported by NIC Argentina. This is probably due to the fact that domain names are free of charge, which is why we believe registration policies should be carefully reviewed.

The distribution of page sizes is highly skewed. As regards vocabulary, we found that top terms were related to commercial activities, typical of sites dedicated to massive sales, auctions and on-line catalogs. This may cause loss of precision in some kinds of searches because these sites are generally well ranked in search engines. However, the words most commonly used in site names, extracted from the URLs, are related to tourism, an ever growing activity in Argentina. From the analysis of link structure and page connectivity we could determine that the studied web is generally well connected. An indicator of this fact is that the MAIN component of the macroscopic structure accounts for 54.23% of sites while there is a low proportion of ISLANDS (9.21%).

From a technological perspective, we found that, from the total number of downloaded pages, 48% are static and the rest, dynamic. The latter are built, in a great proportion, using free programming tools such as PHP (53%) and Perl (39%). Besides, almost 76% of the sites are hosted in servers located in Argentina and 68% of the network addresses (IP addresses) under ".ar" belong to address blocks assigned to Argentina. Such findings reveal that there is an important web-related technological development and communication infrastructure in Argentina.

Finally, from the projection of the total size of the Argentinian web, we can infer that if the set of uncollected sites actually exhibits behavior similar to the studied sample, then more than half a million sites containing 70 million documents are accessible, which would amount to 1 TB of information. It would be particularly interesting to get the full list of sites under the ".ar" domain in order to collect a larger sample and make new estimates, especially for sites with lower visibility (i.e. poorly connected sites and ISLANDS).

We believe it will be interesting to continue research in order to trace maps reflecting the evolution and dynamics of this information space which will enable the study of its behavior and evolution over time. Besides, there is an emerging need for the development of local information services exploiting many interesting features with the main goal of improving user experience, especially in content based applications. One such example could be a tool to filter answers coming from commercial sites in search engines' result lists.

6. Acknowledgements

We would like to thank Eng. Jorge Vilas from NIC Argentina for the information about registered domains and the technical staff at RETINA for their invaluable help.

7. References

[Albert, 2002] R. Albert, and A.-L. Barabasi. "Statistical mechanics of complex networks". Review of Modern Physics, Vol. 74, pp. 47-94, 2002.

[Baeza-Yates, 2001] R. Baeza-Yates, and C. Castillo. "Relating Web characteristics with link based Web page ranking". In Proceedings of String Processing and Information Retrieval (SPIRE), IEEE Cs.Press, pp, 21-32, Laguna San Rafael, Chile, 2001.

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

[Baeza-Yates, 2005_a] R. Baeza-Yates, and C. Castillo. "Características de la Web Chilena 2004". Technical Report, Center for Web Research, University of Chile, 2005.

[Baeza-Yates, 2005$_b$] R. Baeza-Yates, C. Castillo, and V. Lopez. "Characteristics of the Web of Spain". Cybermetrics, Vol. 9, No. 1, 2005.

[Baeza-Yates, 2005$_c$] R. Baeza-Yates, C. Castillo, and E. Efthimiadis. "Characterization of national Web domains". To appear in ACM Transactions on Internet Technology, 2007.

[Barabasi, 1999] A.L. Barabasi, and A. Albert. "Emergence of Scaling in Random Networks". Science, (286), pp. 509-512, 1999.

[Broder, 2000] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. "Graph Structure in the Web". In Proceedings of the WWW9 Conference pp. 309-320, 2000.

[Castillo, 2005] C. Castillo, and R. Baeza-Yates. "WIRE: an Open Source Web Information Retrieval Environment". Workshop on Open Source Web Information Retrieval (OSWIR), 2005.

[Efthimiadis, 2004] 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.

[Kleinberg, 1999] Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. "The Web as a Graph: Measurements, Models and Methods". In Proceedings of the International Conference on Combinatorics and Computing, 1999.

[Modesto, 2005] M. Modesto, A. Pereira, N. Ziviani, C. Castillo, and R. Baeza-Yates. "Un Novo Retrato da Werb Brasileira". In Proceedings of SEMISH, São Leopoldo, Brazil, 2005.

[O'Neill, 2003] E. O'Neill, B. Lavoie, and R. Bennett. "Trends in the Evolution of the Public Web 1998 - 2002". D-Lib Magazine, Vol. 9, No. 4, 2003.

[Page, 1998] L. Page, S. Brin, R. Montwani, and T. Winograd. "The Pagerank Citation Ranking: Bringing Order to the Web". Technical Report, Stanford Digital Library Technologies Project, 1998

[Vilas, 2006] J. Vilas. Solicitud de datos para investigación sobre "Caracterización de la web Argentina". Comunicación Personal, February 6, 2006.

[WCA, 1999] Web Characterization Activity. http://www.w3.org/WCA/