

Evolution of the Chilean Web Structure Composition

Ricardo Baeza-Yates Barbara Poblete

Center for Web Research
Dept. of Computer Science
University of Chile
E-mail: {rbaeza,bpoblete}@dcc.uchile.cl

Abstract

In this paper we present the evolution of the structure of the Chilean Web between 2000 and 2002. Our results show that although the Web grows as expected, also a significant part of it disappears. In addition, some components are much more stable than others. We also compare the expected life cycle of a Web site in the structure with the actual real data.

Year	2000	2001	2002
Pages	730.673	794.218	2.214.253
Sites	10.352	21.207	39.320
Domains	9.102	19.389	35.520

Table 1. TodoCL collections.

1. Introduction

The Web is highly dynamic and little is known about its evolution. There are models that predict when a page will change, but that differs a lot from site to site. There are also generative models for Web growth, but they do not include Web death. In fact, new websites appear and others disappear, but little is known on how this happens. In this paper we present the evolution of the structure composition of the Chilean Web at the site and domain level, based on data gathered from a search engine targeted to this web domain, TodoCL.cl, between years 2000 and 2002.

We define the Chilean Web as all the .cl sites plus all other sites found by crawling that have an IP belonging to a Chilean ISP. The first year the crawl started from an initial sample of sites, but subsequent years started with all .cl domains thanks to NIC Chile. Hence, the number of unconnected sites was low the first year. Also, the last crawl contains more dynamic pages, which in general do not change the Web structure. Table 1 shows the data for these years.

Our results present how the structure evolves, how sites migrate from one component to another component, and where sites appear and disappear. The changes are dramatic, corroborating that perhaps we are trying to study a process that is still in a transient phase, or that cannot be modeled in detail. This is a first step to measure and follow the evolution of part of the Web structure, as well as try to understand the process behind the changes. To the best of our knowledge there are no other studies on Web composition as specific as ours. Most statistical studies deal with global attributes as language or size. We would have liked to separate the Chilean Web in commercial, educational, governmental, etc. sites, but Chile does not use a subdomain level indicating that, so the classification is not trivial.

In section 2 we review the results on the structure of the Web. Section 3 shows the evolution of this structure, and section 4 analyzes the expected changes in the structure with respect to the typical life cycle of a Web site. The last section has some concluding remarks. A preliminary and short version of this paper was presented as a poster in [BYP03].

2. Web Structure

The most complete study of the Web structure [BKM⁺00] focus on page connectivity. One problem with this is that a page is not a logical unit (for example, a page can describe several documents and one document can be stored in several pages.) Hence, we decided to study the structure of how websites were connected, as websites are closer to be real logical units. Not surprisingly, we found in [BYC01] that the structure at the website level was similar to the global Web, and hence we use the same notation of [BKM⁺00]. The components are:

- MAIN, sites that are in the strong connected component of the connectivity graph of sites (that is, we can navigate from any site to any other site in the same component);
- IN, sites that can reach MAIN but cannot be reached from MAIN;
- OUT, sites that can be reached from MAIN, but there is no path to go back to MAIN; and
- other sites that can be reached from IN (T.IN, where T is an abbreviation for tentacles), sites in paths between IN and OUT (TUNNEL), sites that only reach OUT (T.OUT), and unconnected sites (ISLANDS).

In [BYC01] we analyzed the data for 2000 and we extended this notation by dividing the MAIN component into four parts:

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

Figure 1 shows all these components.

We also considered domains in our study, although domains may contain sites that are quite different. For example, web hosting in an ISP provider using a common second-level domain such as co.cl. In table 2 we give the relative size of each component. Notice the size of ISLANDS, which is near 50% of the Chilean Web sites. These sites are usually recent, and the main growth of the Web is in that component. The average update time of pages and sites, and their relation to structure and link ranking techniques was

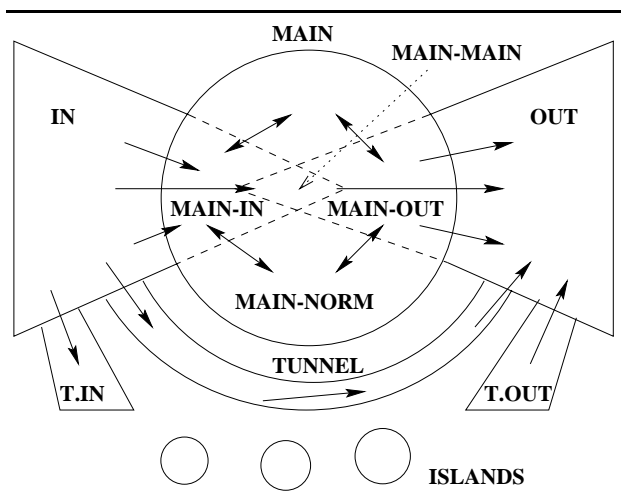


Figure 1. Structure of the Web.

studied in [BYSJC02] for the first two collections (2000 and 2001).

Component	Size (%) 2001	Size (%) 2002
MAIN	9.25%	11.98%
IN	5.84%	9.97%
OUT	20.21%	17.15%
TUNNEL	0.22%	0.23%
TENTACLE-IN	3.04%	3.11%
TENTACLE-OUT	1.68%	3.31%
ISLANDS	59.73%	54.21%
MAIN-MAIN	3.43%	4.08%
MAIN-OUT	2.49%	2.77%
MAIN-IN	1.16%	2.24%
MAIN-NORM	2.15%	2.88%

Table 2. Relative size of the number of sites in the components of the Chilean Web.

3. Evolution of the Structure Composition

Table 3 shows the number of sites and domains that have appeared and disappeared from year to year.

In tables 4 and 5 we show the migration of sites among the components. Similarly, table 6 does the same analysis for domains. There are two ways of reading these tables. By columns we have from which component comes the sites/domains in each component. By rows, we see where are today the sites/domains of the components in

Year	Sites			Domains	
	2000	2001	2002	2001	2002
TOTAL	7.497	21.207	39.320	19.389	35.520
NEW	-	15.415	23.937	-	21.397
GONE	-	1.705	5.824	-	5.266

Table 3. Growth and death of sites and domains.

the previous year. The last column and row represent the sites/domains that do not longer exist (GONE) and the new sites/domains (NEW), respectively. Death of a site means that there are no IP addresses associated to it (this might be wrong if the site changes its name) and death of a domain means that there are no sites associated with it (in particular the domain name itself or prefixed by www)¹

Notice that OUT and MAIN are stable components, because about 25% of the sites stay there. It is also interesting to see that MAIN grows from OUT by 20%, and that ISLANDS is the component with largest growth and also death.

Figures 4 and 5 show graphically the migration of sites and domains among the different components, using the same textures to identify from which component in the previous year they came.

4. Analysis of Web Site Migration

Web sites evolve and hence migrate inside the structure. First, a typical Web site should start as part of ISLANDS or IN (depending if they link or not a good Web site). If the site becomes popular and they also link known sites, it migrates to MAIN. If links are not well chosen or updated, they start in or migrate to OUT. Figure 2 shows the expected life path of a website to migrate to MAIN. We also include the migration from MAIN to OUT if the site is not well maintained. Figure 3 shows the real migration (percentage) in the structure. We can notice that there is almost no migration from IN to MAIN or from MAIN to OUT in opposition to what intuition predicts. Also, there are websites that appear directly in MAIN or OUT. This means that a good site seems to be linked from a site in MAIN in less than a year, or that sites obtain links from portals in MAIN (for example, a banner).

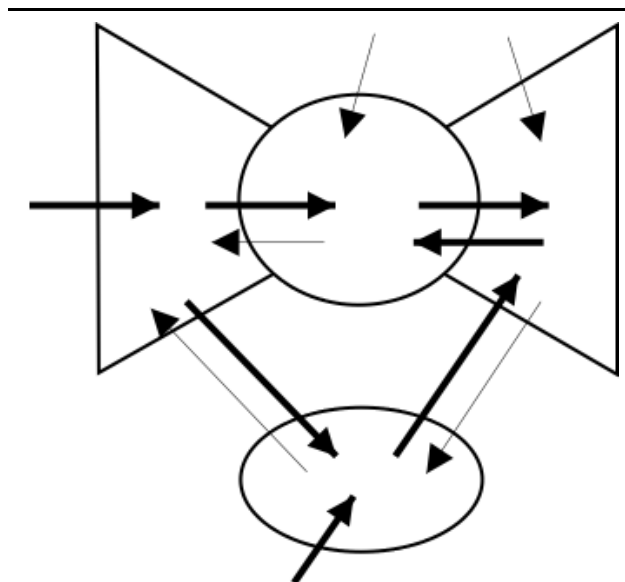


Figure 2. Expected migrations of websites in the structure.

5. Conclusions

The overall number of sites of the Chilean Web is duplicating each year. However, that is the result of more than a 125% increase plus a 25% death. In addition, many sites, sometimes because of ignorance, do not allow crawlers to enter. For example, in 2001, 56% of the domains and 54% of the sites had only one page. However, 25% of them (14% of the total) had an initial Flash page or called a similar kind of program.

There is still a lot to do to understand how the composition of the structure changes. For example we can follow specific sites, but in that case have to see if a one-year sampling strategy is enough. We are currently studying the change at the level of pages related to the structure. For example, the largest 20 sites (in pages) in 2002 are all different from the largest sites in 2001.

Acknowledgements

We acknowledge the support of Millennium Nucleus Grant P01-029-F from Mideplan, Chile, and Chilean Fondecyt Project 1020803.

¹ The domain name could be still registered, though.

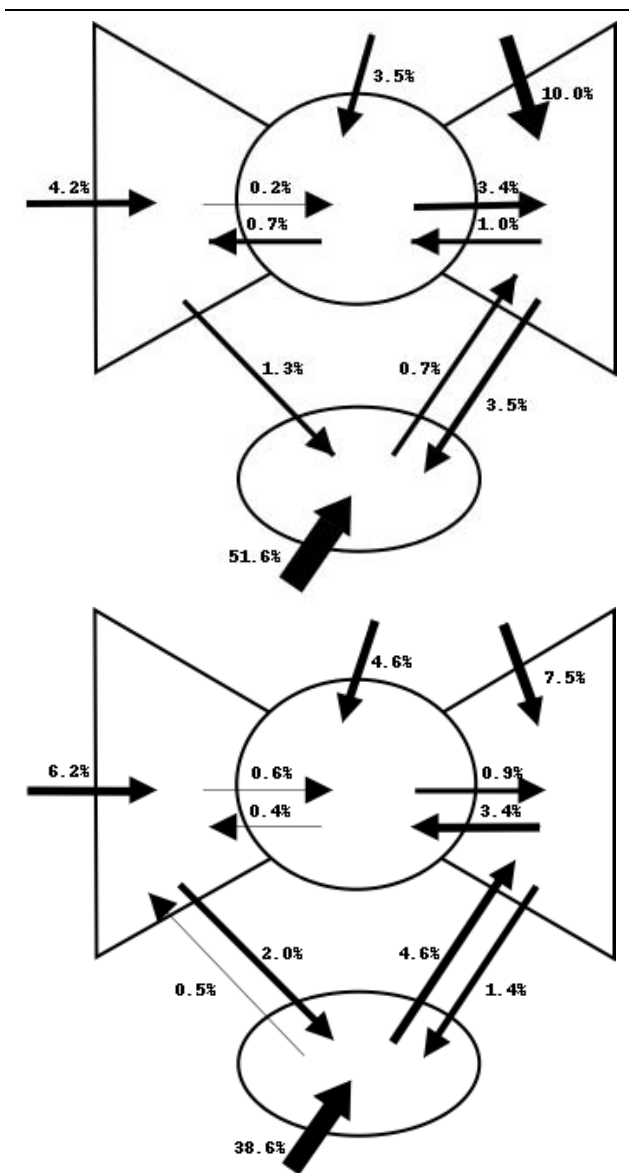


Figure 3. Real migrations of websites in the structure in percentage (top: 2000-2001, bottom: 2001-2002).

References

- [BYC01] Ricardo Baeza-Yates and Carlos Castillo. Relating web characteristics with link analysis. In *String Processing and Information Retrieval*. IEEE Computer Science Press, 2001.
- [BYSJC02] Ricardo Baeza-Yates, Felipe Saint-Jean, and Carlos Castillo. Web dynamics, structure, and link ranking. In *String Processing and Information Retrieval*. Lecture Notes in CS, Springer, 2002.
- [BYP03] Ricardo Baeza-Yates and Barbara Poblete, Evolution of the Web Structure. In *Twelfth World Wide Web Conference*. Budapest, Hungary, 2003.
- [BKM⁺00] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, and A. Tomkins. Graph structure in the Web: Experiments and models. In *9th World Wide Web Conference*, 2000.

2000 \ 2001	MAIN	OUT	IN	ISLANDS	TUNNEL	TIN	TOUT	GONE
MAIN	959	724	140	305	11	61	24	509
OUT	195	1151	39	749	5	96	48	668
IN	39	89	118	279	2	31	25	226
ISLANDS	18	124	14	213	0	14	19	174
TUNNEL	1	1	3	18	0	0	2	3
TIN	5	31	0	18	3	3	2	37
TOUT	3	38	25	131	0	4	12	88
NEW	742	2128	901	10955	27	437	225	-

Table 4. Component changes of sites from 2000 to 2001.

2001 \ 2002	MAIN	OUT	IN	ISLANDS	TUNNEL	TIN	TOUT	GONE
MAIN	1214	339	158	42	1	17	8	183
OUT	901	1683	188	532	15	128	43	796
IN	233	98	292	196	1	22	16	382
ISLANDS	422	1351	786	5182	23	365	299	4240
TUNNEL	11	15	3	4	1	2	0	12
TIN	78	215	25	128	2	66	5	127
TOUT	52	79	41	59	0	18	24	84
NEW	1801	2965	2430	15173	50	608	910	-

Table 5. Component changes of sites from 2001 to 2002.

2001 \ 2002	MAIN	OUT	IN	ISLANDS	TUNNEL	TIN	TOUT	GONE
MAIN	918	218	79	35	0	4	4	141
OUT	892	1424	167	466	14	97	35	560
IN	206	79	288	182	2	19	9	326
ISLANDS	487	1276	970	4967	25	320	242	4074
TUNNEL	4	1	3	1	0	0	0	4
TIN	88	226	22	134	0	59	8	102
TOUT	35	22	39	35	0	2	19	59
NEW	1376	2176	2644	14171	27	419	584	-

Table 6. Component changes in domains from 2001 to 2002.

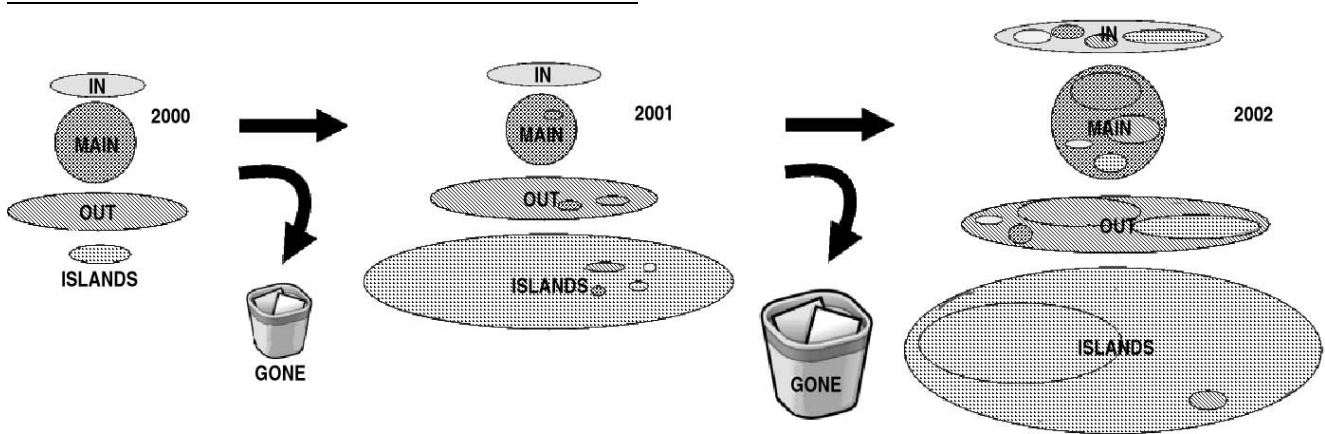


Figure 4. Flow of sites among components. The same texture in each component indicates the origin of old sites.

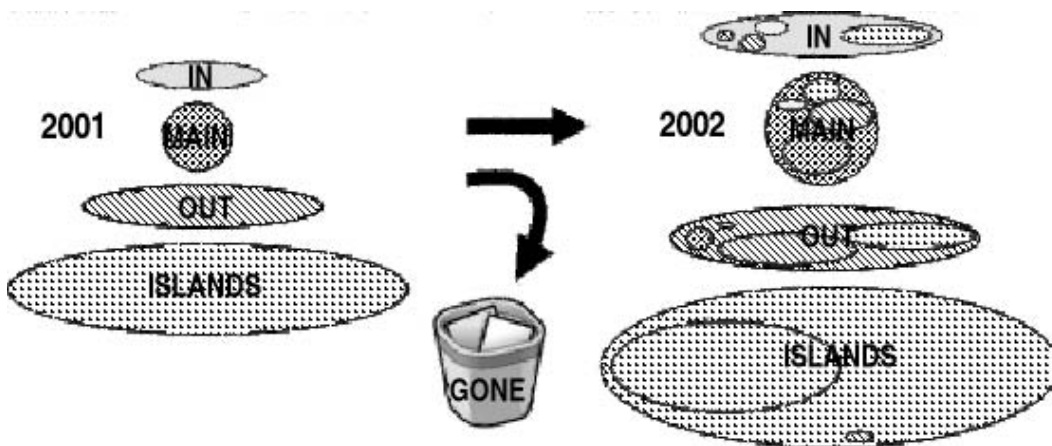


Figure 5. Flow of domains among components.