Relating Web Characteristics

Ricardo Baeza-Yates
Carlos Castillo

Depto. de Ciencias de la Computación
Universidad de Chile
Blanco Encalada 2120
Santiago 6511224, Chile
E-mail: {rbaeza,ccastill}@dcc.uchile.cl *

Abstract

Recent studies have shed light on the structure and dynamics of the Web, as well as Web search behavior. In this paper we present a first step in analyzing how these Web characteristics correlate between them in the context of a closed Internet domain. We discover several relations referent to Web structure and age, Web structure and link based ranking, and Web structure and user search behavior.

1 Introduction

The Web became popular in less than ten years and has grown exponentially to an estimated number of pages of over two billion. Several studies have characterized the Web size, connectivity, dynamics, user search behavior, and languages, to mention a few. However, there is little information about how all those characteristics relate to each other, in particular, which are the main dependencies. In this paper we use the Chilean Web pages to explore those unknown dependencies. Although this is a small subset of the Web, is not a sample of the global Web as in most other studies. In fact, all the pages of a country are much more homogeneous, as they share a culture, are dominated by a single language, and most page visits have a common context. In summary, our subset is very close to a logical collection of pages.

We study the relations between Web connectivity, Web dynamics, page ranking based on link analysis, user search behavior, and other Web measures. As pages are not always logical documents, we consider Web sites as our logical basic units. As a result, we find some known dependencies, but we also discover some unexpected relations, as well as corroborating a few claims. The main discoveries are the relation of the structure of the Web with site age, how link based ranking is related to the structure, the behavior of search users and editors, and that the distribution of search queries is less skewed than Web text data.

The paper is organized as follows. Section 2 presents the scope of our study with its relation to previous work. Section 3 explores the relations of the Web structure with the rest, while section 4 looks at Web dynamics. Section 5 explores user search behavior. The final section discusses some of the results and ongoing work.

*This work was partially supported by Fondecyt project 99-0628 and TodoCL.

2 Scope of the Study

Our study is focused in the Chilean Web, mainly the .cl domain on the first half of this year, when we collected 670 thousand pages, corresponding to approximately 7.500 Web sites. About 93% of the pages are in Spanish, while most of the rest are in English, with an average page size of about 15Kb. The .cl domain currently has about one million pages and more than 10 thousand sites and also grows exponentially, albeit perhaps slower than all the Web. Our data comes from the TodoCL search site [tod00] which specializes on the Chilean Web and is part of a family of vertical search engines built using the Akwan search engine [akw00]. TodoCL also has a directory which is based on the Open Directory Project [ODP99], which at the time of the crawling had about three thousand entries for Chile. A complete characterization of the Chilean Web was presented in [BYC00].

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 Web sites are connected, as Web sites are closer to be real logical units. Not surprisingly, we found that the structure in Chile at the Web site level was similar to the global Web and then we use the same notation of [BKM+00], that is:

(a) MAIN, sites that are in the strong connected component of the connectivity graph of sites;

(b) IN, sites that can reach MAIN but cannot be reached from MAIN;

c) OUT, sites that can be reached from MAIN, but there is no path to go back to MAIN; and

d) other sites that can be reached from IN (t.in), sites in paths between IN and OUT (tunnel), sites that only reach OUT (t.out), and unconnected sites (island).

We extend this notation by dividing 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.

Figure 12 shows the percentage of pages in each component (the left column), while figure 1 shows the structure using number of pages and number of Web sites of each component to represent the area of each part of the diagram.

Our study is driven by this structure, as we want to find correlations between other Web characteristics and its structure. In the sequel we use the diagram based in the number of sites, as that is our logical unit and because the areas of the components are more balanced. Figure 2

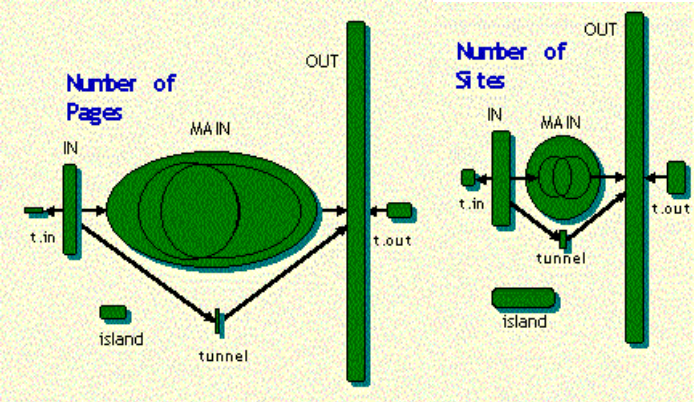

Figure 1: Connectivity structure of the Chilean Web with component areas proportional to number of pages, and number of sites.

shows three examples using text size (no tags), total size (text with tags), and the ratio of both. Each color, from white (minimum) to black (maximum) represents a value using a linear mapping. For example, the leftmost diagram means that MAIN-NORM and MAIN-MAIN are the largest components when taken in account text size without tags. If we include tags, OUT also becomes important (middle diagram). The right diagram shows which component as a whole uses less tags, which is MAIN-NORM.

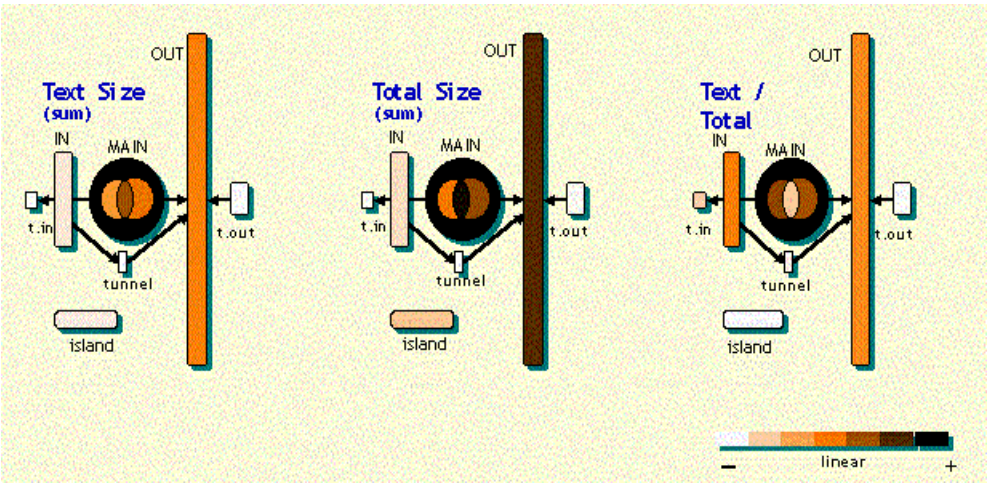

Figure 2: Examples of colored view of the structure with respect to text size without tags, total text size, and their ratio.

We also gathered time information (last-modified information) for each page, to try to correlate dynamic information with other measures. In our data, almost 83% of the pages had a valid last-modified date. Another 2% had a zero value, which in most cases is due to static links to dynamic pages. The other 15% had in most cases no date information. As the Web is young, we use months

or days as time unit. In the case of a Web site, site age is defined as the date of the oldest page, which gives us a lower bound of the site age. Around five thousand Web sites had age larger than 0 (typically, if a page has no date information is due to a problem on the Web server).

Search engines are one of the most visited Web sites and several studies show that most visits are the result of a Web search. The use of this type of tool depends on the user expertise [HS00]. As is customary, TodoCL keeps track of user behavior, and in particular, queries submitted to it. In this study, we used 730 thousand queries in a period of about three months. Those queries had an average length of 2.43 words (this is similar to the AltaVista study [SHMM98]), and 29% of the queries had at least one stopword in it (stopwords are words that are not useful in most cases for a search because they appear in almost all pages). The queries do not have operators because TodoCL uses a menu with three alternatives (search for all the words, some of the words, or a sentence).

An interesting relation between structure and search behavior is due to ranking algorithms based in link analysis. The most well known is PageRank [BP98] which is used in the Google search engine [goo98]. PageRank is static and global in the sense that is precomputed over all pages for all queries. In this sense, can be considered a popularity measure for a page. On the other hand, Kleinberg [Kle98] introduces the concept of Authorities and Hubs, which are computed only on the subset of pages that have the search query. This idea coupled with word based ranking, as is used in most search engines, is presented in [SRNC+00] and used in the TodoBR [tod99] search site.

We adapt link analysis for Web sites in the following way. PageRank models a user surfing the Web in a random fashion, where if you are in a page, with certain probability you get bored and leave the page, or you choose uniformly to follow one of the links on the page that you are (removing self links). Hence, the rank of page p is

$$PR_p = \frac{q}{T} + (1-q)\sum_{i=1}^{k} PR_{r_i}$$

where T is the total number of pages, q is the probability of leaving page p (in the original work $q = 0.15$ is suggested), and r_i are the pages pointed by page p. Figure 3 shows the cumulative page rank distribution in our data, which shows that most pages have a meaningful page rank, with the best pages concentrated in 1% of the total.

Having the rank of a page, we can define the rank of a Web site in many different ways. We can use:

a) The average of the page rank of all site pages (this is not fair with good sites that have too many pages);

b) The maximum page rank of all site pages (this is biased for sites that have only one good page and many bad ones); or

c) The sum of the page rank of all site pages (which is equivalent to having visited one page of the site).

We think that the later definition is the best, being more fair, and because also models the probability of visiting sites. This can be formalized as follows. Let $L_{i,j}$ the number of links from Web

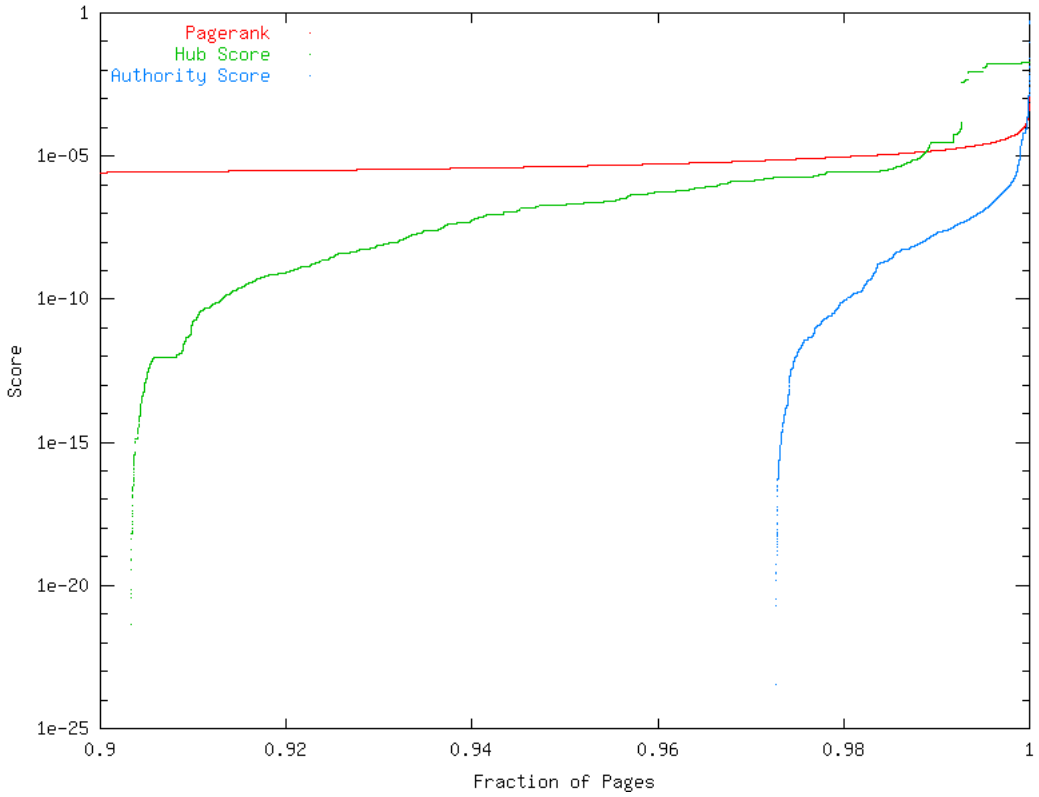

Figure 3: Cumulative distribution of PageRank, hubs, and authorities.

site i to j. We can redefine the rank of Web site w using a random Web site surf, obtaining the following equation:

$$PR_w = \frac{q'}{W} + (1 - q') \sum_{i=1}^{k} L_{w,v_i} PR_{v_i} \ ,$$

where W is the total number of Web sites, q' is the probability of leaving the Web site, and v_i are the sites pointed by w (which could be itself). In this case, as in general we have many pointers from site to site, we weigh each case by $L_{i,j}$.

If we want to simulate the rank based on the sum of page ranks, the equivalent q' should be set to 0.17. If we consider that links from a Web site to itself should not be counted because they are not independent, we set $L_{w,w} = 0$. In this case $q' = 0.4$. Finally, if we want to consider only Web site connectivity, we set $L_{i,j} = 1$ for all i and j, obtaining $q' = 0.37$. This is consistent, because page site connectivity is mainly internal, and then we get bored sooner in a Web site with few or no internal links.

In the case of authorities and hubs, we computed the global authority and hub values per page using the original algorithm. Figure 3 also shows the cumulative distribution of hub and authorities of the pages. We found that only 10% of the pages were meaningful hubs (because about half of the sites do not have links to other sites), while only 3% of the pages had some authority (because

about one third of the sites are not pointed). This means that many directories point to the same pages. The final step of the hub distribution are identical pages which are mirrored in many sites. Hubs and authorities are much more discriminating than PageRank. Then we used the same three definitions as for PageRank to compute the authority and hub value of a Web site. We can define this formally by using again $L_{i,j}$ as for PageRank.

3 Web Structure

We start by correlating the Web site connectivity structure and different Web site characteristics. In each Web site we consider the total number of pages, the total text size (with and without tags), the average page depth, the in-degree (incoming links to a site), and the out-degree (outgoing links of a site). Next, in each component of the structure we compute the average of these measures considering the Web sites in it. Figures 4 and 5 show these relations.

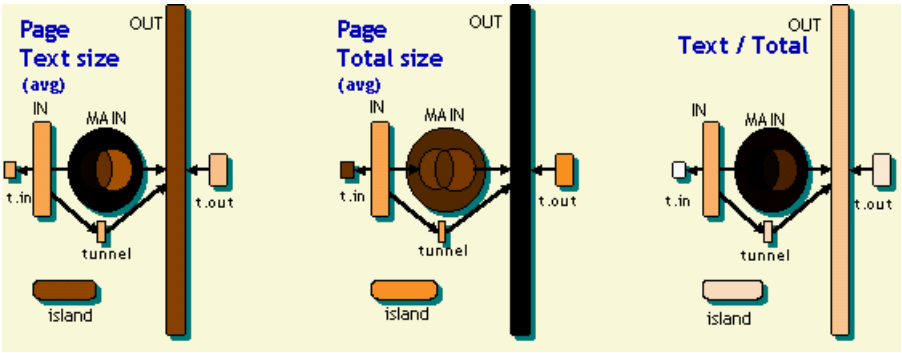

Figure 4: Web Structure vs. Web site characteristics (1).

About 10% of the pages did not have size information and were not included (although we can know the size when bringing the page). The larger Web sites are concentrated in MAIN-NORM and MAIN-IN, which indicate that the other MAIN components are good directories (portals) to other sites, which is corroborated next. The right diagram shows that Web sites in MAIN-OUT, use tags in a proportion larger than other Web sites in MAIN, which implies pages with complicated layout and (or) graphical design (which is reasonable as they seem to be directories).

Depth is related to size and organization of a Web site. Clearly, the Web sites in MAIN are deeper, but notice that the subcomponent MAIN-NORM is the most deeper. This is in contrast with connectivity, because the higher number of in-links are in MAIN-IN and MAIN-MAIN, while the out-degree is concentrated in MAIN-MAIN and MAIN-OUT. The later means that those subcomponents may have "better" directories. Also, as the number of in-links is the same of out-links, as also seen in [BKM+00], in-links are more concentrated than out-links (and reflect the popularity of some sites).

We also computed PageRank, Authorities, and Hubs per site using the three definitions of Web site rank given on the previous section. Figure 6 shows the corresponding diagrams, as well as the total ranking for the component (rightmost column). Looking at the second row, we confirm that the best directories (hubs) are in MAIN-MAIN and MAIN-OUT as pointed out by the out-

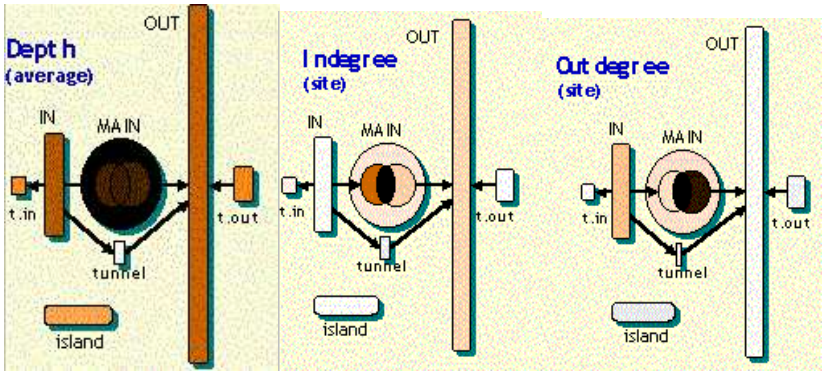

Figure 5: Web Structure vs. Web site characteristics (2).

degree connectivity. On the other hand, the best content (authorities) is concentrated in OUT, MAIN-MAIN and MAIN-NORM, while according PageRank, all the MAIN component has the best content.

4 Web Dynamics

One of the initial motivations of our study was to see if the IN and OUT components were related to Web dynamics or just due to bad Web sites. In fact, Web sites in IN could be considered as new sites which are not linked because of causality reasons. Similarly, OUT sites could be old sites which have not been updated. Figure 7 shows the correlation between the structure and Web site age (oldest, average, and newest page). The average case can be considered as the freshness of a site, while the newest page a measure of update frequency on a site. Figure 8 plots the cumulative distribution of the oldest page in each site for each component of the Web structure versus date in a logarithmic scale (these curves have the same shape as the ones in [BC00] for pages).

These diagrams show that the oldest sites are in MAIN-MAIN, while the sites that are fresher on average are in MAIN-IN and MAIN-MAIN. Finally, the last diagram at the right shows that the update frequency is small in MAIN-MAIN and MAIN-OUT, while sites in IN and OUT are updated less frequently.

Here we obtain some confirmation to what can be expected. The newer sites are in the Island component (and that is why they are not linked, yet). The oldest sites are in MAIN, in particular MAIN-MAIN, so the kernel of the Web comes mostly from the past. What is not obvious, is that on average sites in OUT are also newer than the sites in other components. Finally, IN shows two different parts: there is a group of new sites, but the majority are old sites. Hence, a large fraction of IN are sites that never became popular.

What about the correlation between ranking and age? Figure 9 shows the PageRank of all pages with respect to age. The bottom dots are normal pages, being the lower region, low ranked pages in low ranked sites, which is the most common case from the point of view of a link based ranking. The fact that most of the new or recently modified pages have low rank (the solid red region) shows that Google is biased to old pages. This is bad considering the constant change and fast growth of the Web. This suggests that newer pages should have more weight, in particular if they have

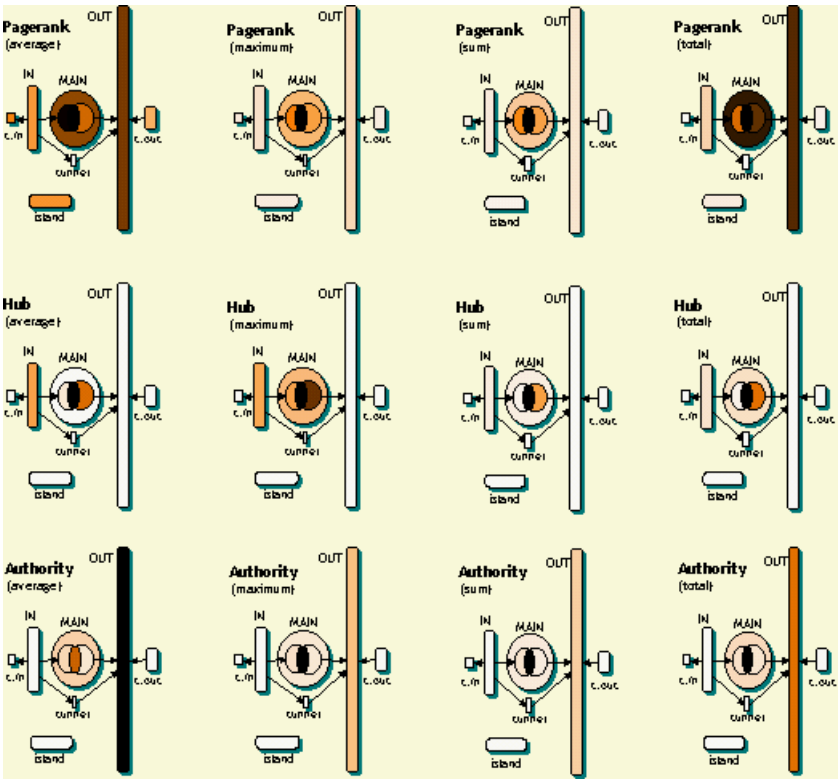

Figure 6: Web Structure vs. Web site rank.

incoming links. However, in that case, not always we can know if the links were put before or after the page changed. Following this line of thought, as also links are not usually modified, old links will give better rank to pages that may have old or even invalid information.

5 Web Search Behavior

The Web collection has approximately two million different words. It is well known that the size of the vocabulary follows a sub-linear model (Heaps' law [BYRN99]) with exponent around .5 for English text data. In our Web collection the exponent goes up to .63, which is consistent with the fact that we have two languages, many more mistakes, and other sequences that are not words in any natural language. On the other hand, most of the words in the queries did appear in the collection.

One interesting issue not related to Web structure, is how queries (which can be seen as small documents) differ from Web pages (document collection). Figure 10 show the word frequency distribution in the text collection and the queries as well as the document frequency distribution of the words in the collection.

It is well known that word frequency can be modeled by a generalized Zipf distribution, where the frequency of the i-th word is proportional to $i^{-\theta}$ (for example, see [BYRN99, ch. 6]). Our

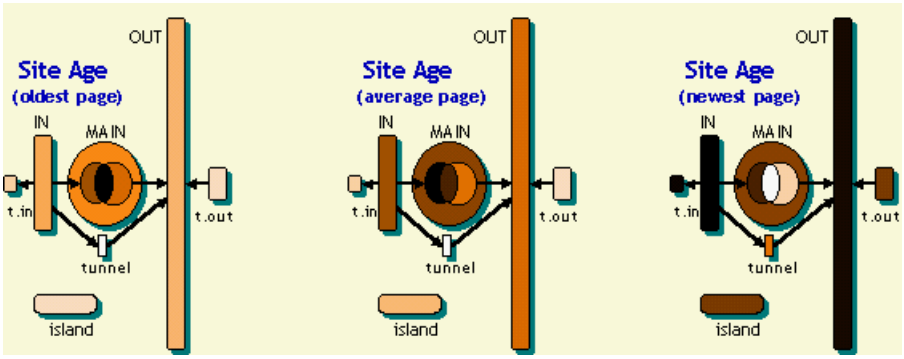

Figure 7: Web Structure vs. Web site age.

intuition was that query words were more biased than the words in the collection, because there are very popular terms such as MP3. Using least squares on the central part of our data (that is, eliminating most frequent words and the right tail) we obtained the following parameters for them: $\theta = 1.59$ for the collection (term or document frequency) and $\theta = 1.24$ for the queries. Two of the models are also plotted in figure 10 (the third model is a line parallel to the bottom one). That is, our intuition was wrong, and most queries are less skewed than words in the collection. Another unexpected result is that the document and term distribution in the collection are almost parallel and they only meet at the right end, with very infrequent words (instead of approaching to each other slowly).

How the queries relate to the collection? Figure 11 shows the normalized frequency of the words in the queries using the frequency order of the words in the collection. The fact that queries are less skewed is corroborated by the green dots over the red line, which are more frequent on the right. On the other hand, stopwords in the queries appear below the red line (green region at the left).

To relate search behavior to the Web structure, we used the information of which pages were visited after a search. Figure 12 shows the fraction of sites in each component visited after a search with respect to Web structure as well as which pages are chosen by ODP editors to build the directory. We can clearly see that searching users choose pages very differently from ODP editors. One reason could be that the behaviors are similar, but the choices for good pages are not. Notice that because the ODP links are inside TodoCL, and TodoCL belongs to MAIN, there cannot be OPD pages in the IN component. To solve this problem, we excluded TodoCL from MAIN in this analysis.

Figure 12 shows that for the users, the Web structure is different than, say the collection itself or ODP editors (which have very restrictive policies). This means that the search engine is discriminating the sites, guiding the users to good resources. In fact, if the proportions were the same for the search engine and the directory, would mean that the search engine is biased to popular and old sites. This type of diagram can be used to evaluate a ranking algorithm and obtain a distribution accordingly to a certain goal. For example, more uniform or biased towards newer sites.

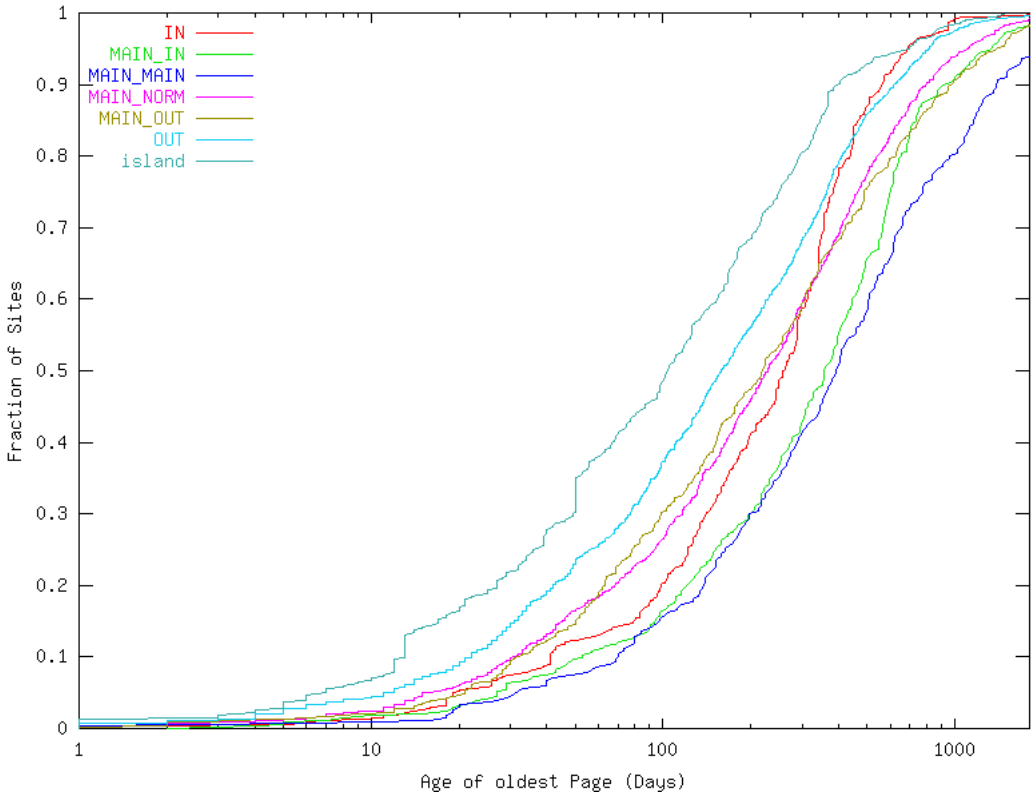

Figure 8: Cumulative distribution of sites in different components vs. age.

6 Concluding Remarks

In this paper we have attempted a first study to correlate different Web characteristics. One first criticism might be the data size. Although one million pages is small nowadays, is big enough for a statistical study. In addition, we have the advantage that we can crawl .cl almost completely (over the 95% of the Web sites), which is not the case in larger studies, and is not biased to "popular" or "better" pages. That is, as the coverage is larger, the results are in some sense more complete.

A corollary of studying a subgraph of the Web and finding many similar results as larger parts of the Web, either at the page or site level, is that corroborates the auto-similarity of the Web. That is, this one experimental evidence of auto-similarity at the connectivity level, at the organization level (Web sites), and even in a geographical dimension. Additionally, as Chile is a developing country, we can include an economical dimension (assuming that all these dimensions are independent). This leaves the door open to studies comparing Webs of countries in the developed and developing world.

Perhaps the most interesting relations affecting the final user is the dependencies between ranking of pages and dates due to specific ranking algorithms that are not fair in the time dimension; as well as the dependency of people choices, as is the bias of Web editors to older sites (possibly because are easier to find due to the previous dependency), which affect the information of good

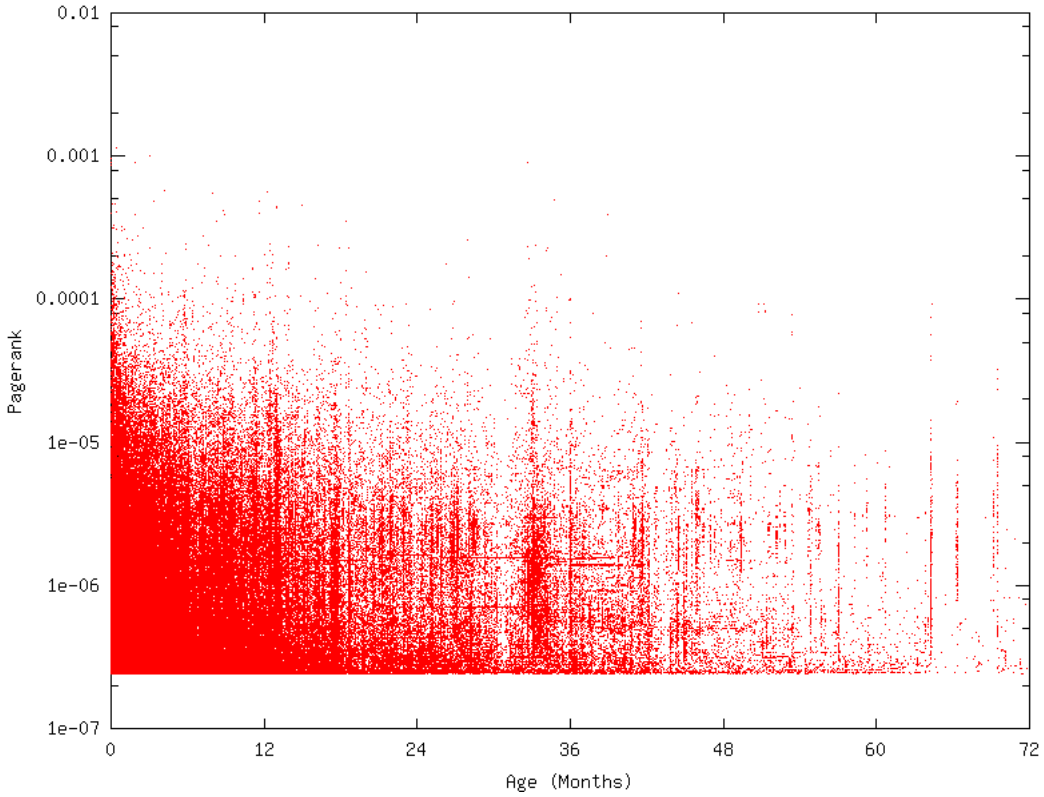

Figure 9: PageRank vs. Age.

directories (hubs). Both dependencies can have a large impact in electronic commerce as they benefit older sites.

We had more data that we did not include for lack of space, which will be included in the final version. For example, the cumulative distribution of PageRank, hub, and authorities value of pages per component, as well as tables summarizing all the numerical results.

References

[akw00] Akwan: Main page. `http://www.akwan.com`, 2000.

[BC00] B. Brewington and G. Cybenko. How dynamic is the web? In *9th Int. WWW Conference*, Amsterdam, Holand, May 2000.

[BKM+00] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: Experiments and models. In *9th Int. WWW Conference*, Amsterdam, Holand, May 2000.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In *7th WWW Conference*, Brisbane, Australia, April 1998.

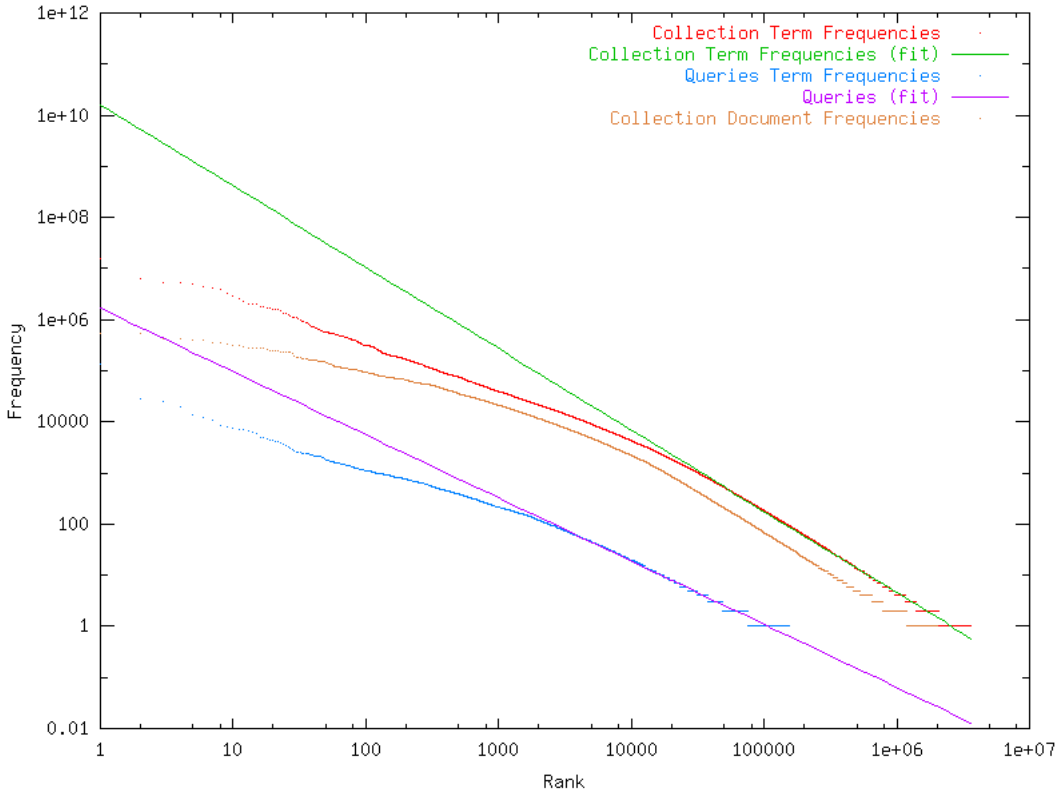

Figure 10: Term and document frequency in the collection and term frequency in the queries.

[BYC00] R. Baeza-Yates and C. Castillo. Characterizing the Chilean web (in spanish). In *Chilean Computer Science Congress*, Santiago, Chile, Nov 2000. Available in www.todocl.cl/stats.phtml.

[BYRN99] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley & ACM Press, Harlow, UK, 1999.

[goo98] Google: Main page. `http://www.google.com`, 1998.

[HS00] C. Holscher and G. Strube. Web search behavior of internet experts and newbies. In *9th Int. WWW Conference*, Amsterdam, Holand, May 2000.

[Kle98] Jon Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. of the 9th ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, San Francisco, USA, Jan 1998.

[ODP99] Open directory project: Main page. `http://odp.org`, 1999.

[SHMM98] C. Silverstein, M. Henzinger, J. Marais, and M. Moricz. Analysis of a very large alta vista query log. Technical Report 1998-014, Compaq Systems Research Center, Palo Alto, CA, USA, 1998.

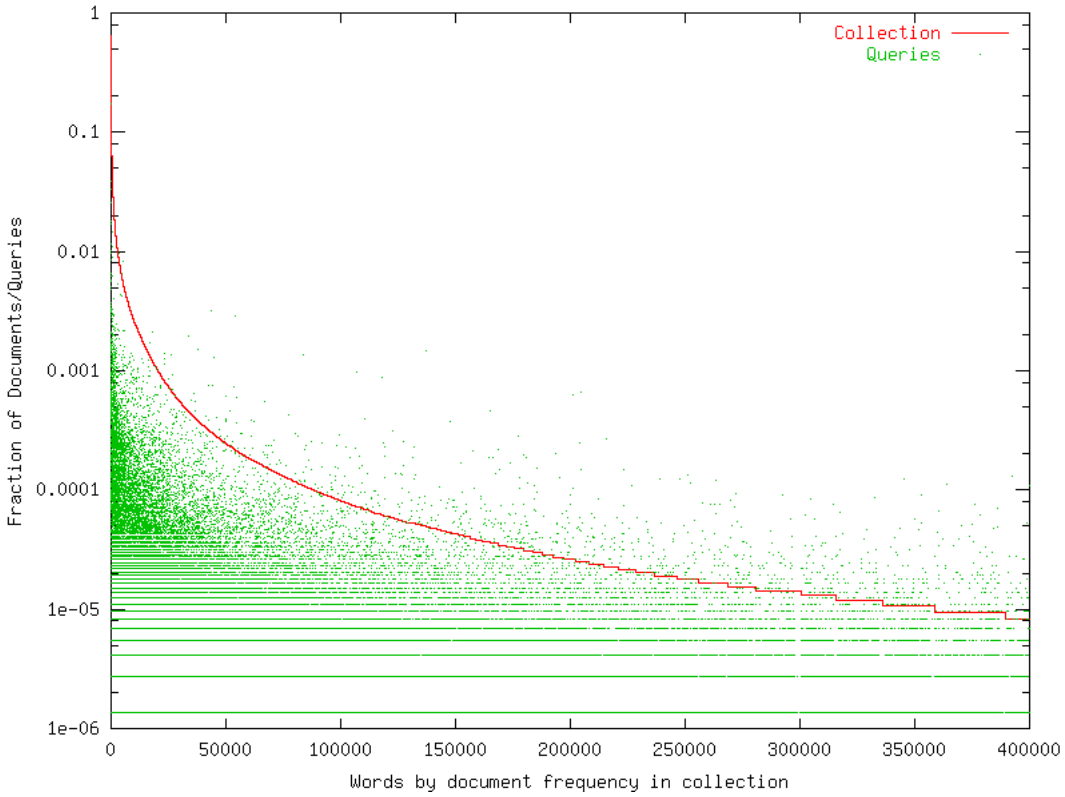

Figure 11: Relation between the words in the collection and the queries.

[SRNC+00] Ilmério Silva, Berthier Ribeiro-Neto, Pável Calado, Edleno Moura, and Nívio Ziviani. Link-based and content-based evidential information in a belief network model. In *Proc of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 96–103, Athens, Greece, July 2000. Best student paper.

[tod99] Todobr: Main page. `http://www.todobr.com.br`, 1999.

[tod00] Todocl: Main page. `http://www.todocl.com`, 2000.

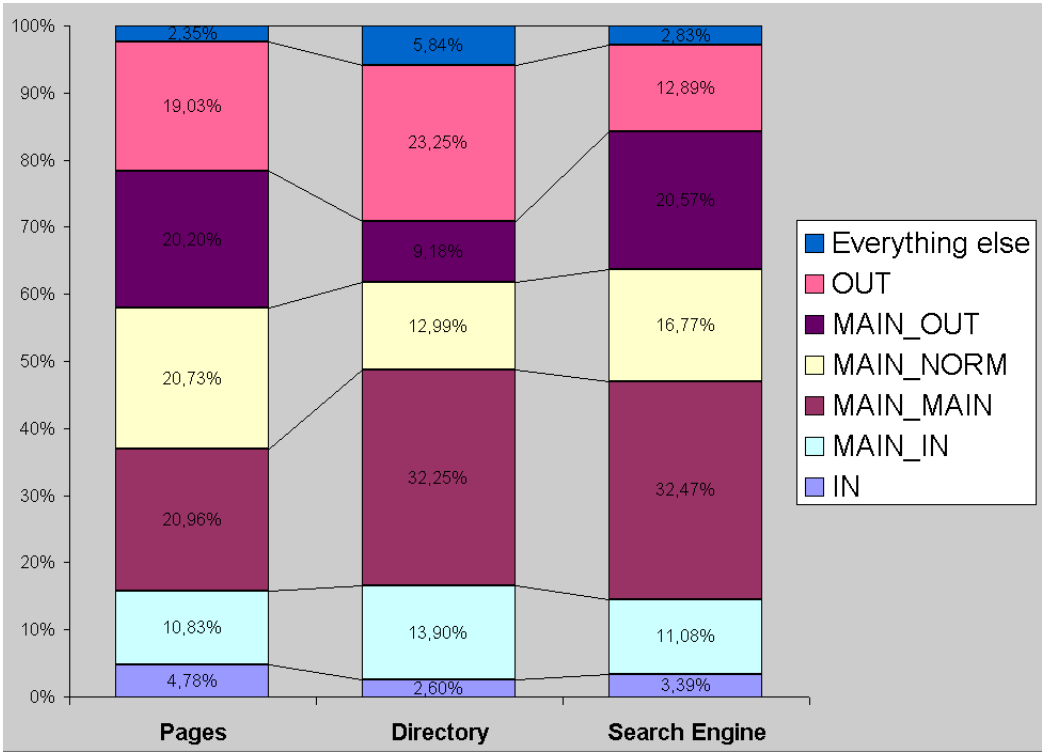

Figure 12: Percentages of pages chosen by the searchers and the ODP editors in the different components.