Analysis of Link Based Ranking for the Web

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

In the last years, several techniques based in link analysis have been proposed and used in search engines to rank Web pages. As links are generated by humans, link based ranking seems to give better results than traditional techniques such as vector based ranking. However, no studies have been done about their real impact. In this paper we extend global page ranking techniques to Web site ranking, and do a first analysis of link ranking regarding the structure and dynamics of the Web.

1 Introduction

The Web became popular in less than ten years and has grown exponentially to an estimated number of pages of over two billions. This exponential growth poses a difficult scalability problem to Web search engines, particularly in the coverage of it and also in how to rank reasonably well large answers. The later issue has been partially solved by using link based ranking in Web search engines such as Google [goo98] or TodoBr [tod99]. In fact, a recent comparison of ranking techniques for Web pages show that link ranking improves precision and recall [SRNC+00], and an even recently survey on this topic highlights its importance [Hen01] . However, no studies about how these techniques relate to the real Web have been done. In this paper we use the Chilean Web pages (.cl domain) to explore how link based ranking relates to Web structure and dynamics. Although this is a small subset of the Web, it is not a sample of the global Web as in most other Web 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, which resembles the whole Web considering the high degree of auto-similarity that we have found [BYC00].

As pages are not always logical documents, we also consider Web sites as our logical basic units. We consider three link based ranking techniques: PageRank [BP98] and authorities and hubs [Kle98]. We extend these link based ranking techniques to Web sites and we study their relation with the structure of the Web and site age. As a result of our study, we find some known relations, but we also discover some new relations. Between the main results we can mention that:

- PageRank is biased against new pages and sites, which benefits older sites.

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

- There are more good hubs (good directories) than authorities (good pages), so finding them is easier.
- That link based ranking is much more uniform over Web sites than over Web pages, and that PageRank is very different from hub or authority ranking.

The paper is organized as follows. Section 2 presents the scope of our study, the extension of page based ranking to Web site ranking, as well as previous work. Section 3 presents an empirical analysis of page link ranking, as well as the extension of link based ranking to Web sites. Section 4 explores the relations of link ranking with the structure of the Web, while section 5 looks at the relation to the Web dynamics. The final section discusses some of the results obtained and ongoing work.

2 Scope of the Study

Our study is focused in the Chilean Web, mainly the .cl domain on the first half of year 2000, when we collected 670 thousand pages, corresponding to approximately 7.500 Web sites. About 93% of the pages were in Spanish, while most of the rest were in English, with an average page size of about 15Kb. The .cl domain at the end of year 2000, had a bit more than one million pages and more than 30 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]. A complete characterization of the Chilean Web was presented in [BYC00].

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

(a) MAIN, sites that are in the strong connected component of the connectivity graph of sites (72.7%);

(b) IN, sites that can reach MAIN but cannot be reached from MAIN (4.8%);

c) OUT, sites that can be reached from MAIN, but there is no path to go back to MAIN (19.0%); 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). All these sites represent only the 3.5%.

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 (21.0%);

(b) MAIN-IN, which are sites that can be reached directly from the IN component but are not in MAIN-MAIN (10.8%);

(c) MAIN-OUT, which are sites that can reach directly the OUT component, but are not in MAIN-MAIN (20.2%);

(d) MAIN-NORM, which are sites not belonging to the previously defined subcomponents (20.7%).

Figure 1 shows this structure using number of Web sites and number of pages for each component to represent the area of each part of the diagram. In the sequel we use the diagram based in the number of sites (right), because is our logical unit and because the areas of the components are more balanced.

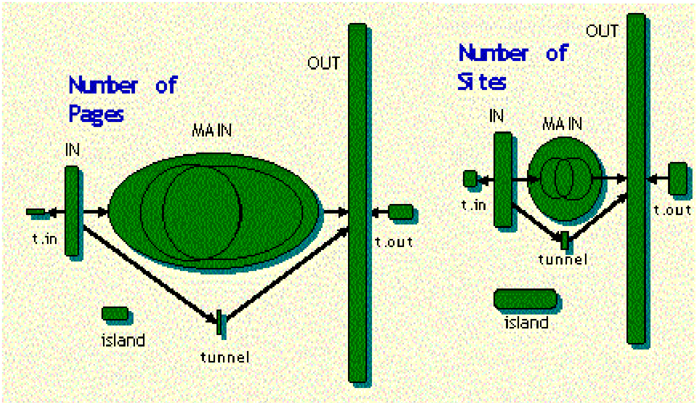

Figure 1: Connectivity structure of the Chilean Web with component areas proportional to the number of pages (left), and the number of sites (right).

We also gathered time information (last-modified information) for each page, to try to correlate dynamic information in out study. 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).

3 Link Based Ranking

Search engines are one of the most visited Web sites and several studies show that most visits are the result of a Web search. An interesting relation between Web structure and search engines 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, independently of the queries. 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. Pages with authority have good content and good hubs are pages that have links to pages with authority. 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. In this case the ranking is dynamic, because the link analysis is computed over a set of pages satisfying the query.

Now, we adapt link analysis for Web sites. PageRank models a user surfing the Web in a random fashion, such that, 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 where you are (removing self links). Hence, the rank of a 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.

Following Kleinberg's idea, in the case of authorities and hubs, we computed the global authority and hub values per page using the original algorithm. That is, the authority of a page is the sum of the hub values of the pages pointing to it, and its hub value is the sum of the authority of the pages that point. Although our computation is global, notice that authorities and hubs depend on a subgraph of the Web that represents certain knowledge, and hence a local computation will be similar. Let s_i be the pages pointing to page p, and as before r_i be the pages pointed by p. Then, we have, before normalization, that

$$A(p) = \sum_i H(s_i), \qquad \text{and} \quad H(p) = \sum_i A(r_i)$$

Figure 2 shows the cumulative page rank distribution in our data for the cases mentioned before. They show that most pages have a meaningful PageRank, with the best pages concentrated in less than 1% of the total. However, PageRank is almost uniform for the vast majority of cases, being q/T the minimal ranking value. We also see that less than 10% of the pages were meaningful hubs (because about half of the sites do not have links to other sites), while less than 3% of the pages had some authority (because about one third of the sites are not pointed by any other site). 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, a duplication problem that is not always easy to solve. Hubs and authorities are much more discriminating than PageRank, and the results suggest that would be better to use a hub based ranking for two reasons: (a) there are more pages with good links, and (b) many users would prefer a good set of related links instead of a few pages with good authority.

To verify what was the relation of the each ranking for the same page, we plot the ranking for all pages using the PageRank order. The result is given in Figure 3, and shows that PageRank is very different from authorities or hubs, which is counter intuitive.

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);

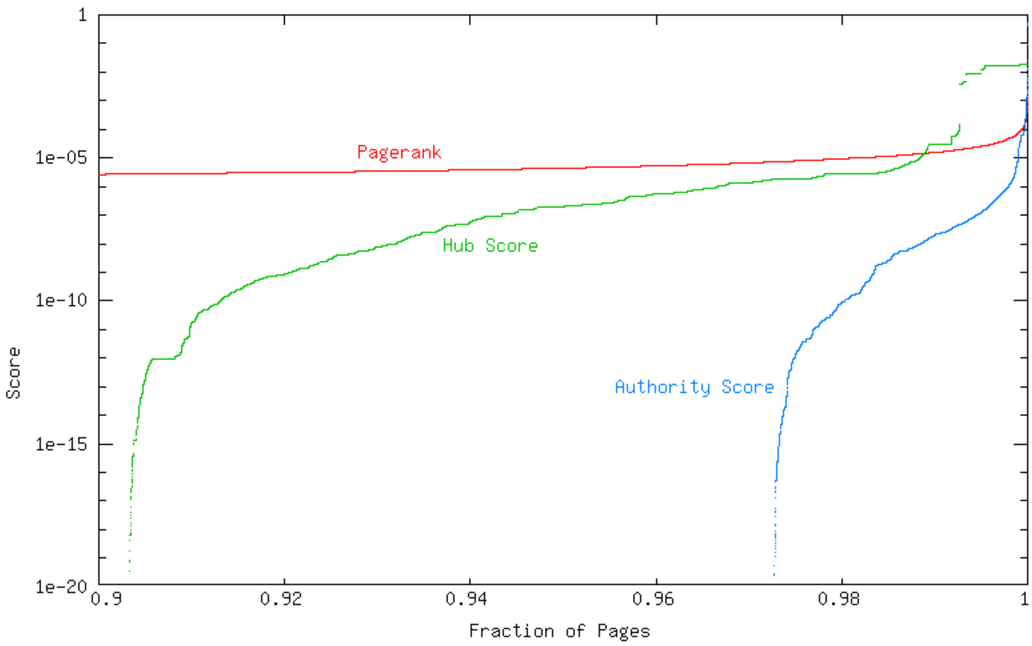

Figure 2: Cumulative distribution of PageRank, hubs, and authorities for pages.

b) The maximum page rank of all site pages (this is biased to sites that may have few good pages 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 site i to j. We can define the PageRank of a 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 instead of 0.15. This means that most page links are internal, so the difference is small. 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 two values are consistent, because page site connectivity is mainly internal, and then we get bored sooner in a Web site with few or no internal links.

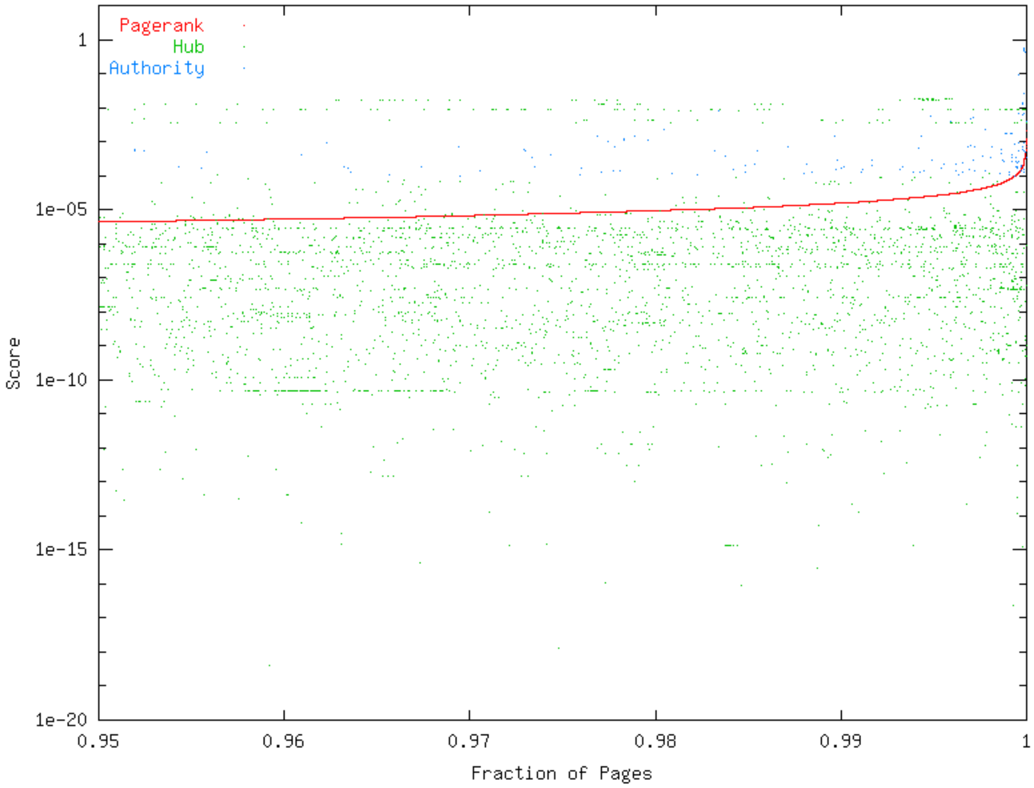

Figure 3: Hubs and authorities sorted by PageRank for all pages.

We can extend hubs and authorities for Web sites in a similar way by using $L_{i,j}$ as weights like in PageRank. Figure 4 shows the distribution of the rank sum site ranking (case (c) before), which is much more uniform than page ranking for all the cases and covers a much larger proportion of sites than in the case of pages. In fact, good hubs are concentrated, because the best page directories (MAIN-OUT) in general have many mages, many of them with good hub score. On the other hand, more of 70% of the sites have some authority. This means that pages with good content are distributed in many more Web sites than good hubs (the average density of authorities is less than half the density of hubs).

4 Web Structure

We start by correlating the Web site structure with its connectivity. In each Web site we consider 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. Figure 5 show these relations.

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 deepest. This is in contrast with connectivity, because the higher number of in-links are in MAIN-IN and MAIN-MAIN, while the

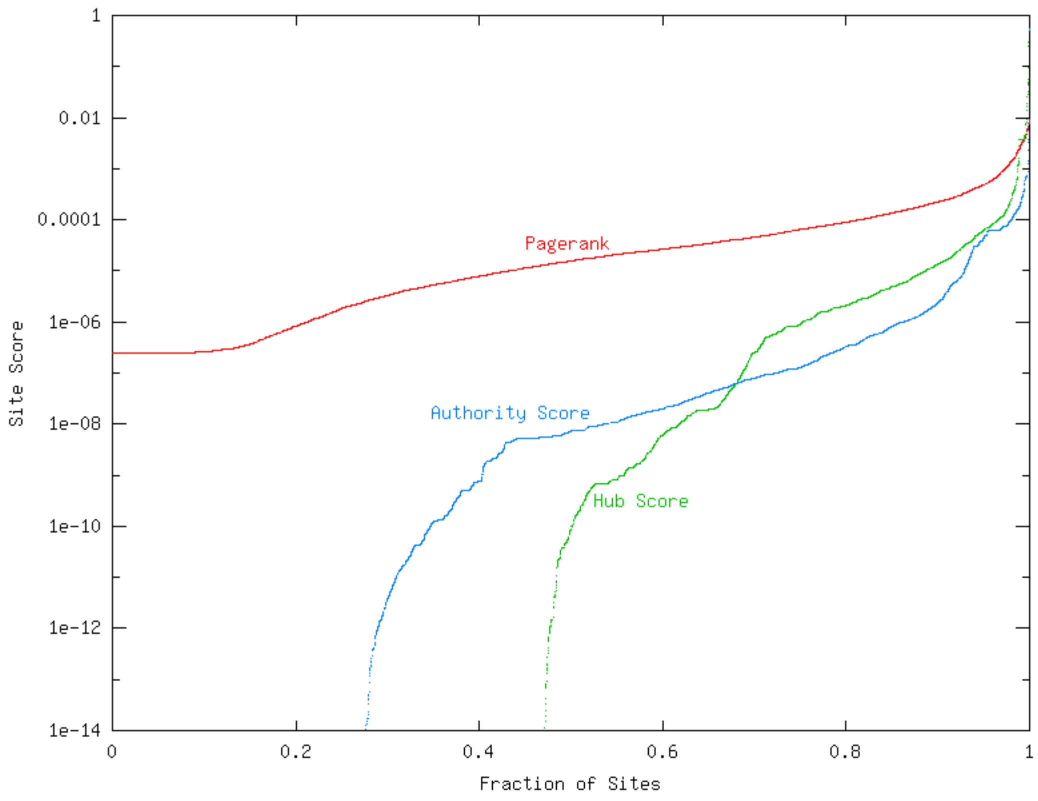

Figure 4: Cumulative distribution of PageRank, hubs, and authorities for sites.

out-degree is concentrated in MAIN-MAIN and MAIN-OUT. The later means that those sub-components may have "better" directories (hubs). 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 (which reflects the popularity of some sites).

We 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 structure diagrams, as well as the total ranking for the component (rightmost column). Each color, from white (minimum) to black (maximum) represents a value using a linear mapping.

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-degree connectivity, but also the IN component have good directories. On the other hand, the best content (authorities) is concentrated in OUT and MAIN-MAIN, while according PageRank, the same holds, but all the MAIN component has also good content.

5 Web Dynamics

One of the initial motivations of our study was to see if the IN and OUT components were related to Web growth (or Web dynamics) or just due to bad Web sites. In fact, Web sites in IN could be

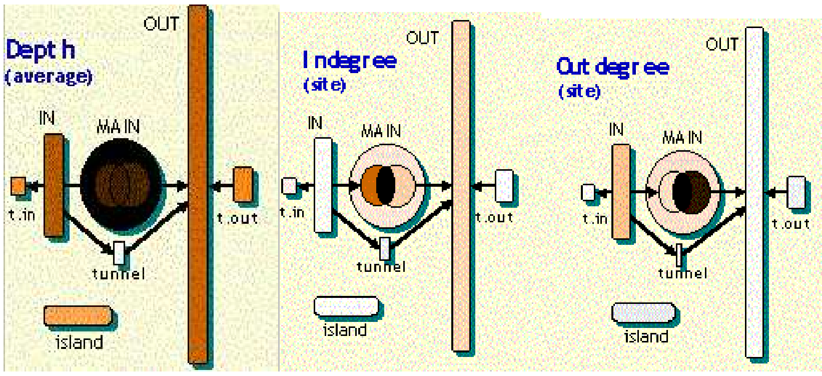

Figure 5: Web structure vs. Web site connectivity.

considered as new sites which are not linked because of temporal 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 is a measure of update frequency on a site. 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.

Doing a more detailed analysis, we found that 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 link ranking and age? Figure 8 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 PageRank 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 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. This graph also shows that the Web grow with periodic bursts of new pages (each vertical line).

On the other hand, many new pages have good hub ranking as seen on Figure 9, while for authority ranking there is a mixed behavior, there are good and bad new pages, as shown in Figure 10. These two results may imply the good hubs have to be updated frequently and that the quality of new pages is distributed on both sides.

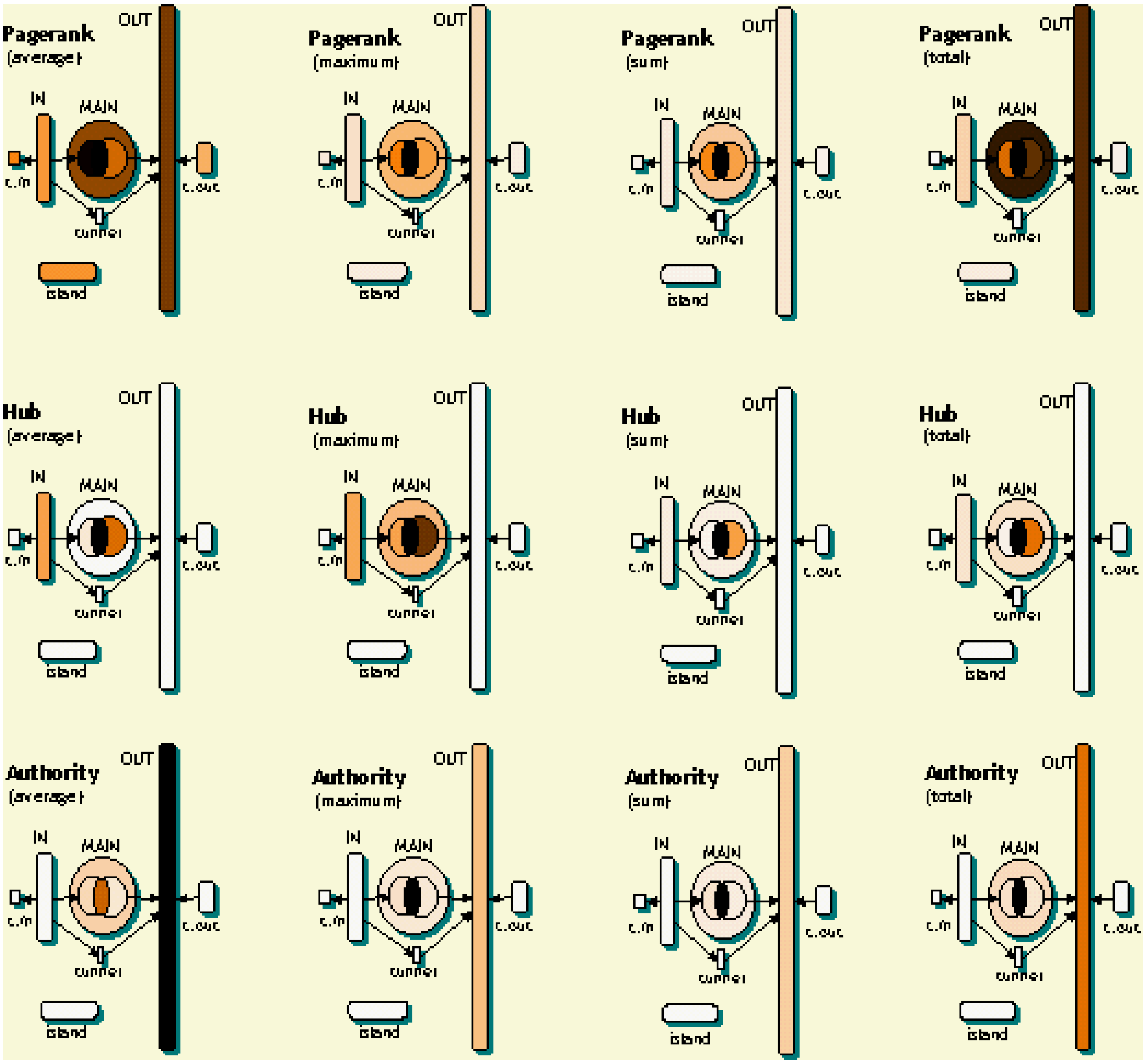

Figure 6: Web structure vs. Web site rank.

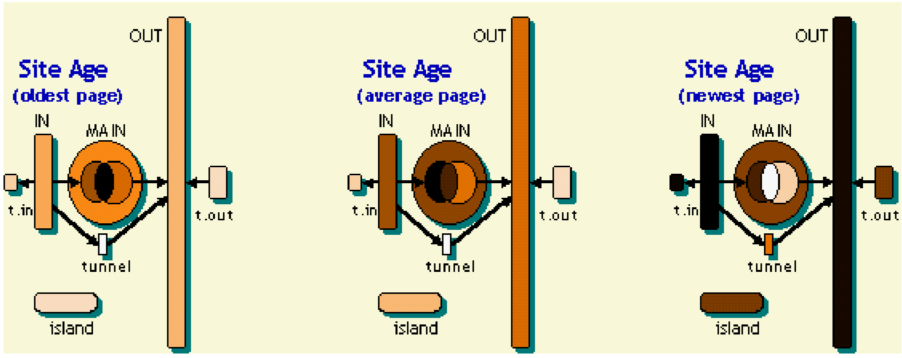

Figure 7: Web Structure vs. Web site age.

6 Concluding Remarks

In this paper we have attempted a first study to correlate link based ranking with 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 and fair.

We have defined the rank of a Web site based on link connectivity. In addition to use it for this kind of studies, Web site ranking can be used for automatic ordering of sites in Web directories, as an alternative to lexicographical or manually based orders (the later is better, but it is not scalable). In particular, for this case, using a hub based measure seems to be the best choice.

Perhaps the most interesting relation affecting the final user are the dependencies between page ranking and dates due to specific ranking algorithms that are not fair in the time dimension, like PageRank, which is used in Google. Considering that most visits are the product of a search, this dependency can have a large impact in electronic commerce as they benefit older sites. The effect is reversed if hub ranking is used.

Our results can help to devise new ranking techniques based in link analysis, giving more weight to hubs or letting the user to decide if he/she wants popular pages, good pages or good directories. Even more, studies of this kind can suggest how many pages must be considered to have meaningful computations of hubs and authorities in dynamic link ranking (that is, when the subset of the Web depends on the query).

In [Hen01] hubs and authorities are criticized because their values for a given subgraph (due to an specific query) can be manipulated by adding edges to a few nodes, and that the answer can suffer a topic drift. The first issue is valid for small answers, but many studies show that user queries have on average two or three words, which yields subgraphs of many thousand pages, which are not easy to manipulate (in addition, it is difficult to add edges in more than one Web site). The second issue only happens if we use additional nodes which may not contain the query (a neighborhood graph). Hence, according to our results, a good mixture of hub and authority based ranking (or one of them selected by the user) will give better results than PageRank. If we add content evidence to that, the result might be even better [SRNC+00].

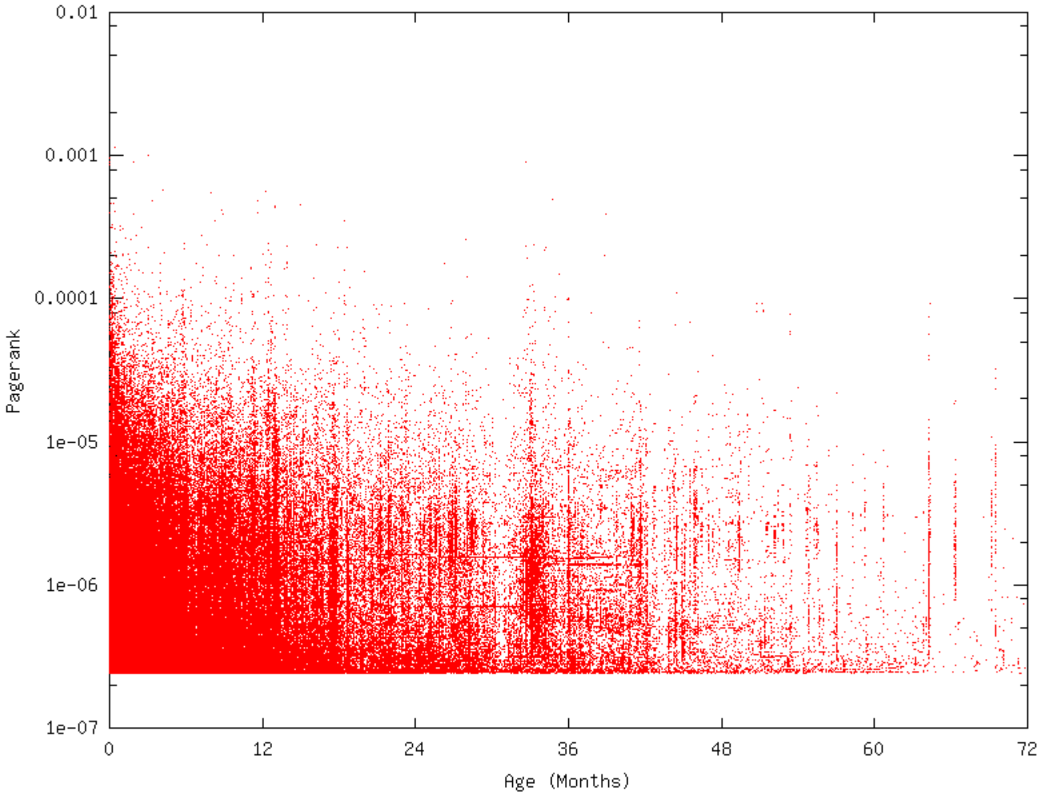

Figure 8: PageRank vs. Age.

Future work includes a sensibility analysis of these ranking techniques to formally assess which technique is easier to be manipulated by Web positioning companies (which try to improve the ranking of a site in specific search engines). Further study is also needed to assess which technique is really better respect to precision and recall.

References

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

[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.

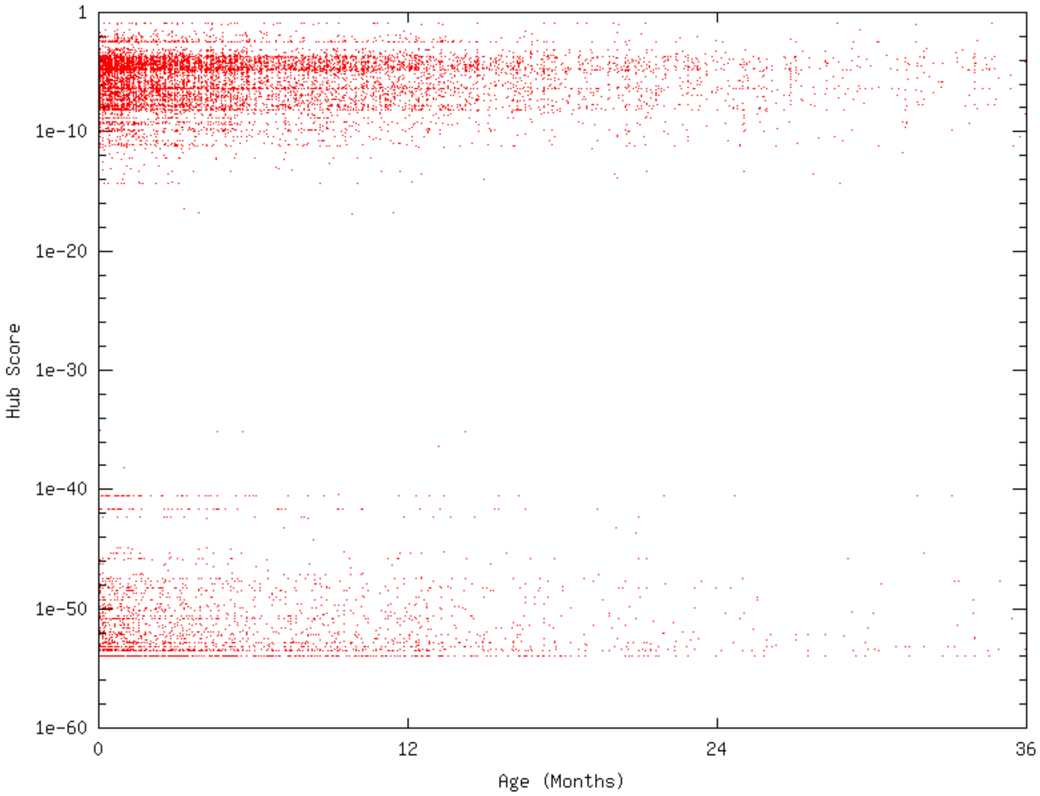

Figure 9: Hub ranking vs. Age.

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

[Hen01] Monika R. Henzinger. Hyperlink analysis for the Web (survey). *IEEE Internet Computing*, 5(1):45–50, Jan/Feb 2001.

[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.

[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.

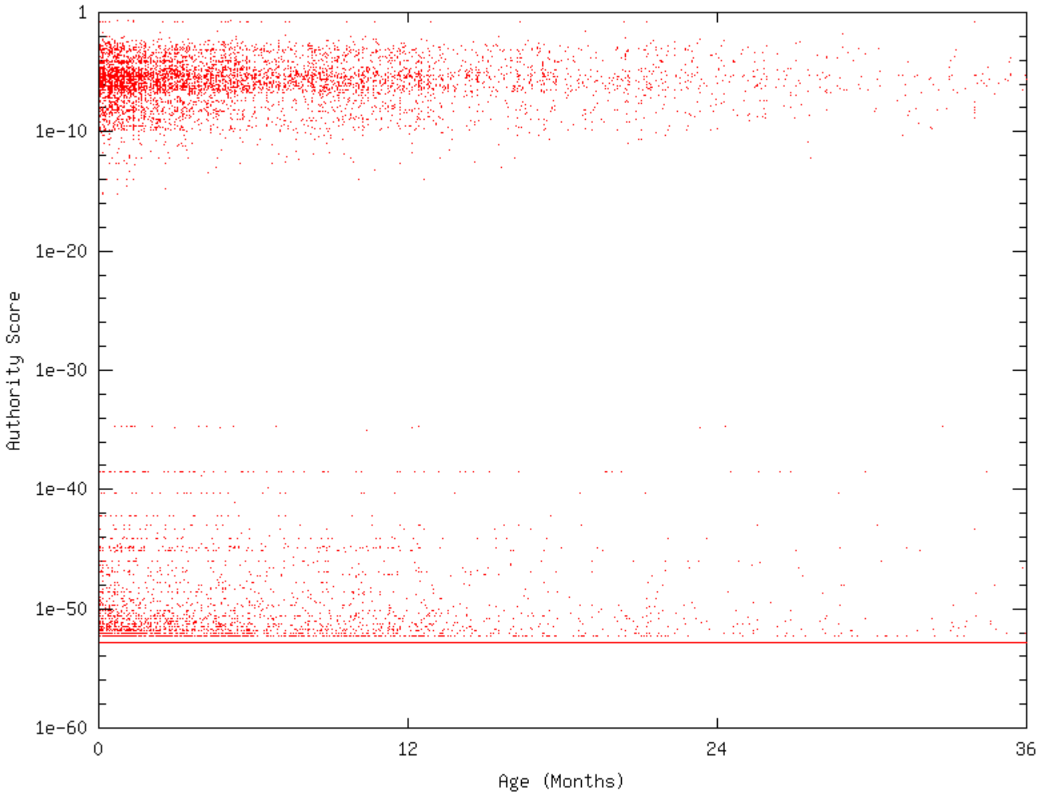

Figure 10: Authority ranking vs. Age.