Web Structure, Age and Page Quality*

Ricardo Baeza-Yates Felipe Saint-Jean Carlos Castillo
Computer Science Department, University of Chile
Blanco Encalada 2120, Santiago, Chile
E-mail: {rbaeza,fsaint,ccastill}@dcc.uchile.cl

Abstract

This paper is aimed at the study of quantitative measures of the relation between Web structure, age, and quality of Web pages. Quality is studied from different link-based metrics and their relationship with the structure of the Web and the last modification time of a page. We show that, as expected, Pagerank is biased against new pages. As a subproduct we propose a Pagerank variant that includes age into account and we obtain information on how the rate of change is related with Web structure.

1 Introduction

The purpose of a Web search engine is to provide an infrastructure that supports relationships between publishers of content and readers. In this space, as the numbers involved are very big (500 million users [3] and more than 3 billion pages[1] in 36 million sites [4] at this time) it is critical to provide good measures of quality that allow the user to choose "good" pages. We think this is the main element that explain Google's [1] success. However, the notion of what is a "good page" and how is related to different Web characteristics is not well known.

Therefore, in this paper we address the study of the relationships between the quality of a page, Web structure, and age of a page or a site. Age is defined as the time since the page was last updated. For web servers, we use the oldest page in the site, as a lower bound on the age of the site.

The specific questions we explore are the following:

- How does the position of a web site in the structure of the Web depends on the Web site age? Depends the quality of a Web page on where is located in the Web structure? We give some experimental data that sheds some light on these issues.
- Are link-based ranking schemes providing a fair score to newer pages? We find that the answer is no for Pagerank [12], which is used by Google [1], and we propose alternative ranking schemes that takes in account the age of the pages, an important problem according to [11].

*Funded by Millenium Project "Center for Web Research", Mideplan, Chile.

[1] This is a lower bound that comes from the coverage of a search engine.

Our study is focused in the Chilean Web, mainly the .cl domain on the two different times: first half of 2000, when we collected 670 thousand pages in approximately 7,500 web sites (Set1) and the last half of year 2001, when we collected 795 thousand pages, corresponding to approximately 21.200 Web sites (Set2). This data comes from the TodoCL search site (www.todocl.cl) which specializes on the Chilean Web and is part of a family of vertical search engines built using the Akwan search engine [2].

Most statistical studies about the web are based either on a "random" subset of the complete web, or on the contents of some web sites. In our case, the results are based on the analysis of the TodoCL collection, a search engine for Chilean Web pages. As this collection represents a large % of the Chilean web, we think that our sample is coherent, because it represents a well defined cultural context.

The remaining of this paper is organized as follows. Section 2 presents previous work and that main concepts used in the sequel of the paper. Section 3 presents several relations among Web structure, age, and quality of Web pages. Section 4 presents the relation of quality of Web pages and age, followed by a modified Pagerank that is introduced in Section 5. We end with the some conclusions and future work.

2 Previous Work

The most complete study of the Web structure [7] 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 in [5] that the structure in Chile at the Web site level was similar to the global Web[2] and hence we use the same notation of [7]. The components are:

(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).

In [5] we analyzed Set1 and we extended 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;

[2] Another example of the autosimilarity of the Web, which gives a scale invariant.

(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.

We also gathered time information (last-modified date) for each page as informed by the Web servers. How Web pages change is studied in [9, 6, 8], but here we focus on Web page age, that is, the time elapsed after the last modification. As the Web is young, we use months as time unit, and our study considers only the three last years as most Web sites are that young. The distribution of pages and sites for Set1 with respect to age is given in Figure 1.

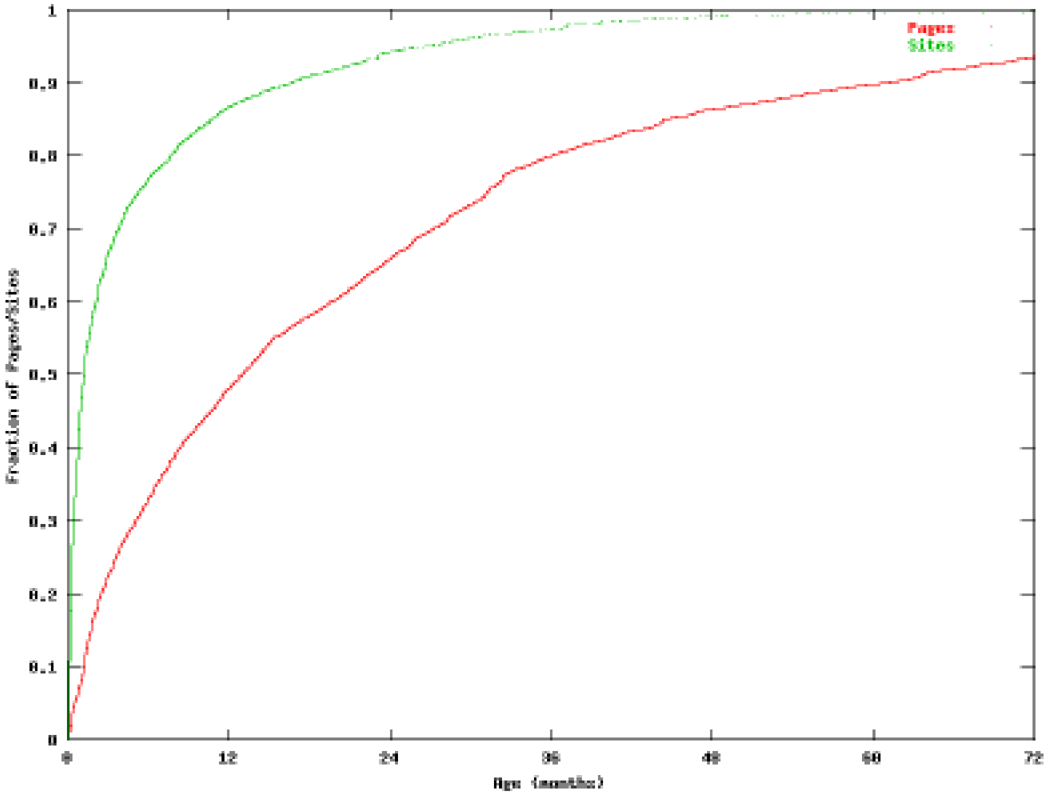

Figure 1: Cumulative distribution of pages (bottom) and sites (top) in function of age for Set1.

The two main link based ranking algorithms known in the literature are Pagerank [12] and the hub and authority measures [10].

Pagerank is based on the probability of a random surfer to be on a page. This probability is modeled with two actions: the chance of the surfer to get bored and jump randomly to any page in the Web (with uniform probability), or choosing randomly one of the links in the page. This defines a Markov chain, that converges to a permanent state, where the probabilities are defined as follows:

$$PR_i = q + (1-q) \sum^{k} \frac{PR_{m_j}}{L_{m_j}}$$

where q is the probability of getting bored (typically 0.15), m_j with $j \in (1..k)$ are the pages that point to page i, and L_j is the number of outgoing links in page j.

The hub and authority are complementary functions. A page will have a high hub rank if it points to good content pages. In the similar way a page will have a high authority rank if it is referred by pages with good links. In this way the authority of a page is defined as the sum of the hub ranks of the pages that point to it, and the hub rank of a page is the sum of the authority of the pages it points to.

When considering the rank of a Web site, we use the sum of all the ranks of the pages in the site, which is equivalent to the probability of being in any page of the site [5].

3 Relations to the Web Structure

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 2 shows the relation between the macro-structure of the Web using the number of Web sites in each component to represent the area of each part of the diagram for Set1. The colors represent Web site age (oldest, average, and newest page), such that a darker color represents older pages. 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 3 plots the cumulative distribution of the oldest page in each site for Set 1 in each component of the Web structure versus date in a logarithmic scale (these curves have the same shape as the ones in [7] for pages). The central part is a line and represents the typical power laws that appear in many Web measures.

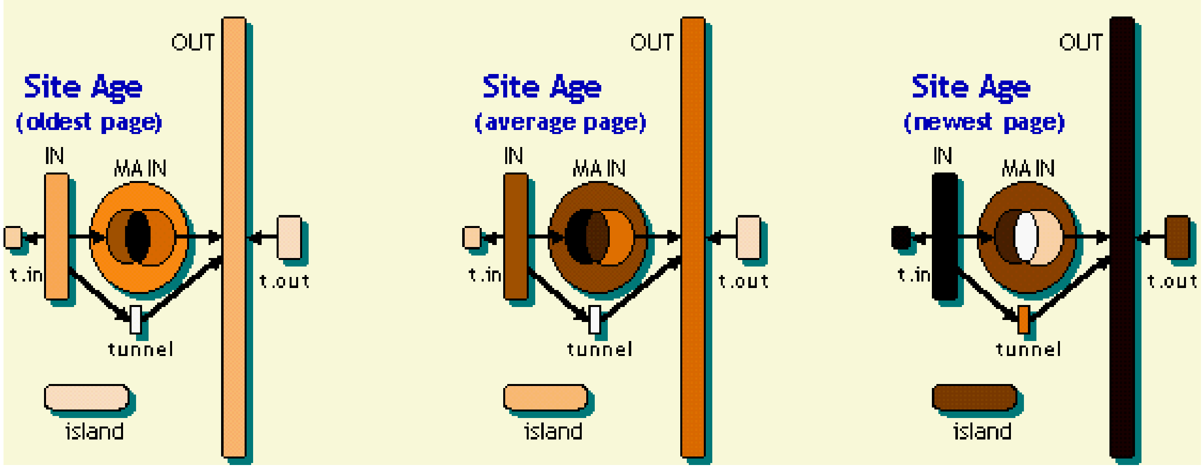

Figure 2: Visualization of Web structure and Web site age.

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 high in MAIN-MAIN and MAIN-OUT, while sites in IN and OUT are updated less frequently.

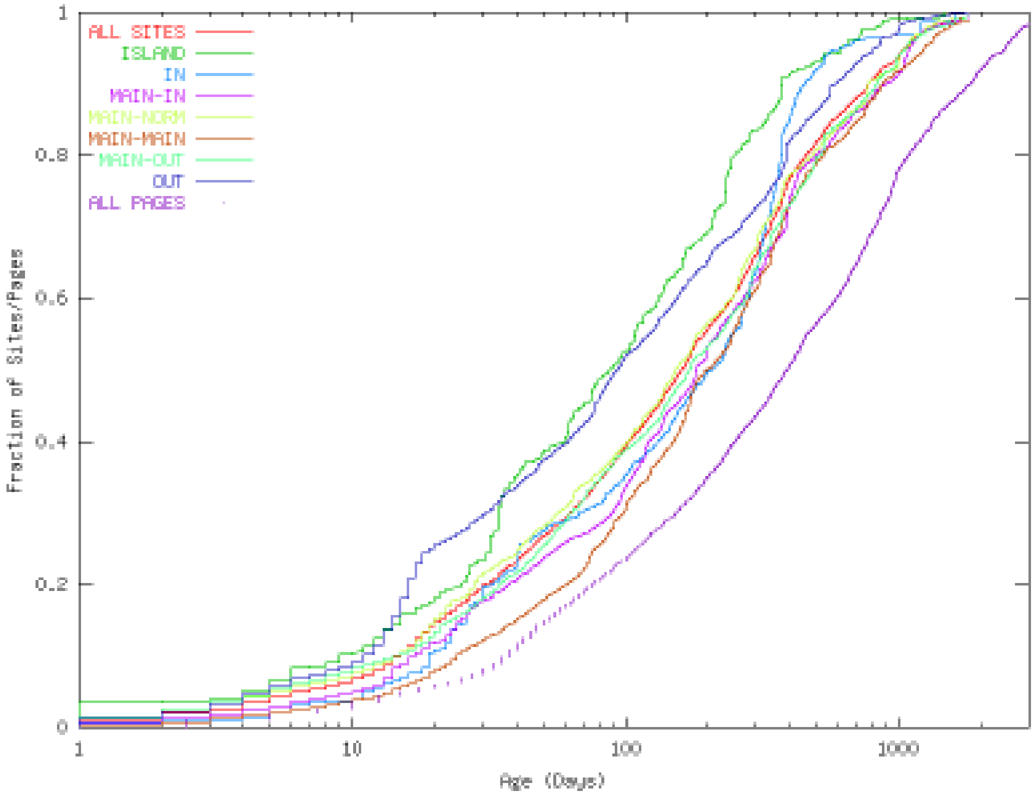

Figure 3: Web site age in the different components and page age (rightmost curve).

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.

In Table 1 we give the numerical data for the average age as well as the Web quality (sum for all the sites) in each component of the macro-structure of the Web, as well as the percentage change among both data sets in more than a year. Although Set1 did not include all the ISLANDS at that time (we estimate that Set1 was 70% of the sites), we can compare the core. The core has the smaller percentage but it is larger as Set2 triples the number of sites of Set1. OUT also has increased, which may imply a degradation of some part of the Web. Inside the core, MAIN-MAIN has increased in expense of MAIN-NORM. Overall, Set2 represents a Web much more connected than Set1.

Several observations can be made from Table 1. First, sites in MAIN have the higher Pagerank, and inside it, MAIN-MAIN is the subcomponent with highest Pagerank. In a similar way MAIN-MAIN has the largest authority. This makes MAIN-MAIN a very important segment of the Web. Notice that IN has the higher hub which is natural because sites in MAIN have the higher authority.

Component	size(%,Set1)	size(%,Set2)	age (days)	Pagerank	hub	authority
MAIN	23%	9.25%	429	0.0002	0.0053	0.0009
IN	15%	5.84%	295	8.02e-05	0.0542	9.24e-08
OUT	45%	20.21%	288	6.12e-05	5.71e-08	1.00e-05
TUNNEL	1%	0.22%	329	2.21e-05	7.77e-08	3.78e-08
TENTACLES-IN	3%	3.04%	256	3.45e-05	1.83e-12	1.53e-06
TENTACLES-OUT	9%	1.68%	293	3.5e-05	4.12e-07	5.41e-09
ISLANDS	4%	59.73%	273	1.41e-05	1.10e-12	3.08e-11
MAIN-MAIN	2%	3.43%	488	0.0003	0.01444	0.0025
MAIN-OUT	6%	2.49%	381	0.0001	7.71e-05	4.19e-07
MAIN-IN	3%	1.16%	420	0.0001	1.14e-06	9.82e-06
MAIN-NORM	12%	2.15%	395	8.30e-05	3.31e-06	1.92e-07

Table 1: Age and page quality for Set2 in the different components of the macro-structure of the Chilean Web.

ISLANDS have a low score in every rank.

Studying age, sites in MAIN are the oldest, and inside it, sites in MAIN-MAIN are the oldest. As MAIN-MAIN also has good ranking, seems that older sites have the best content. This may be true when evaluating the quality of the content, but the value of the content, we believe in many cases, could be higher for newer pages, as we need to add novelty to the content.

Therefore there is a strong relation between the macro-structure of the Web and age/rank characteristics. This makes the macro-structure a valid partition of Websites.

4 Link-based Ranking and Age

Now we study the correlation of the mentioned rank algorithms with the age of the pages. In [5] we gave qualitative data that showed that link-based ranking algorithms had bad correlation and that Pagerank was biased against new pages. Here we present quantitative data supporting those observations.

Web pages were divided in 100 time segments of the same weight (that is, each segment has the same number of pages), and we calculated the standard correlation of each group pair of average rank values. Three graphs where obtained: Figure 4 which shows the correlation between Pagerank and authority, Figure 5 the correlation among Pagerank and hub, and Figure 6 shows the correlation of authorities and hubs.

The low correlation between Pagerank and authority is surprising because both ranks are based on incoming links. This means that Pagerank and authority are different for almost every age percentile except the one corresponding to the older and newer pages which have Pagerank and authority rank very close to the minimum.

Notice the correlation between hub/authority, which is relatively low but with higher value for pages about 8 months old. New pages and old pages have a lower correlation. Also notice that hub and authority are not biased with time.

It is intuitive that new sites will have low Pagerank due to the fact that webmasters of other sites take time to know the site and refer to it in their sites. We show that this intuition is correct in Figure 7, where Pagerank is plotted against percentiles of page age. As can be seen, the newest

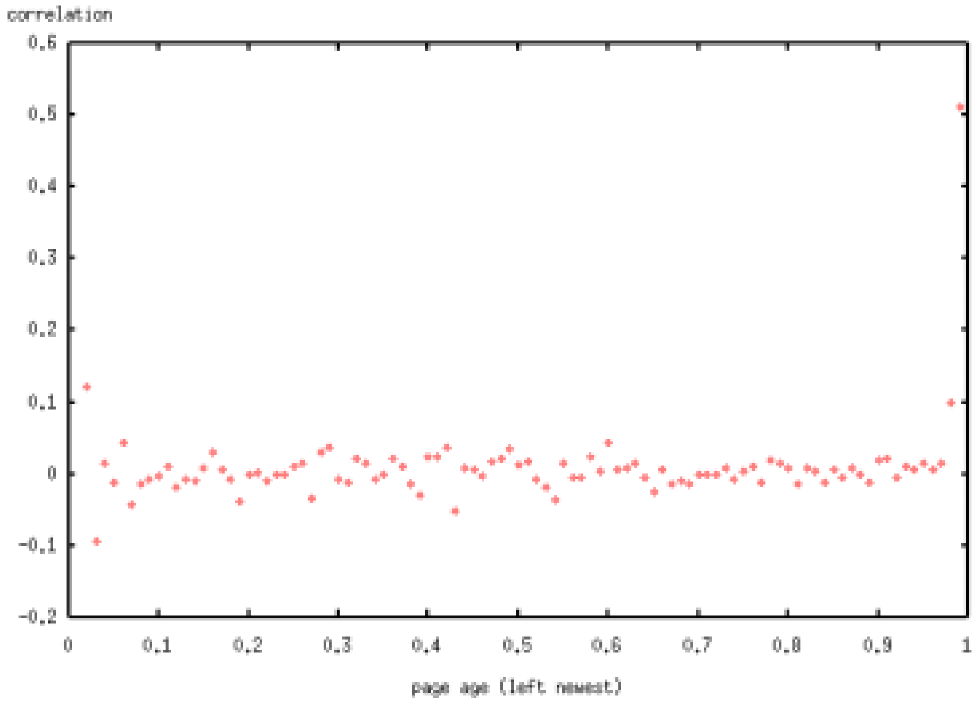

Figure 4: Correlation among Pagerank and authority with age.

pages have a very low Pagerank, similar to very old pages. The peak of Pagerank is in pages of 1.6 months old.

In a dynamic environment as the Web, new pages have a high value so a ranking algorithm should take an updated or new page as a valuable one. Pages with high Pagerank are usually good pages, but the opposite is not necessarily true (good precision does not imply good recall). So the answer is incomplete and a missing part of it is in new pages. In the next section we explore this idea.

5 An Age Based Pagerank

Pagerank is a good way of ranking pages, and Google is a demonstration of it. But as seen before it has a tendency of giving higher ranks to older pages, giving new pages a very low rank. With that in mind we present some ideas for variants of Pagerank that give a higher value to new pages.

A page that is relatively new and already has links to it should be considered good. Hence, the Pagerank model can be modified such that links to newer pages are chosen with higher probability. So, let $f(age)$ be a decreasing function with age (present is 0), and define $f(x)$ as the weight of a page of age x. Hence, we can rewrite the Pagerank computation as:

$$PR_i = q + (1-q)\ f(age_i) \sum_{i=1.\ i\neq i}^{k} \frac{PR_{m_j}}{L_{m_j}}$$

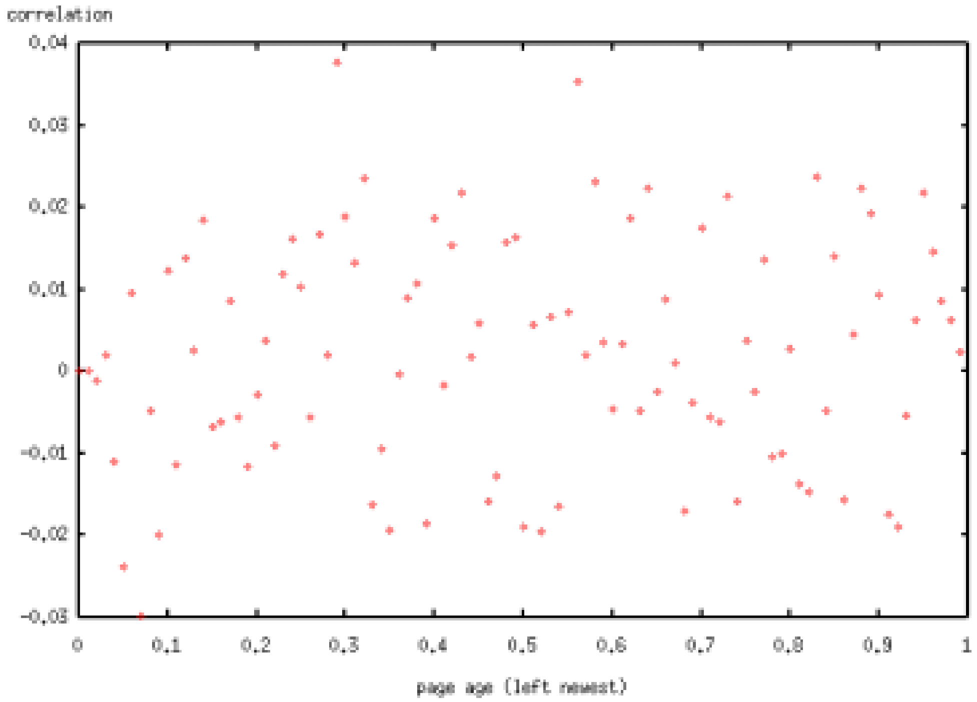

Figure 5: Correlation among Pagerank and hub with age.

where L_{m_j} as before is the number of links in page m_j. At each step, we normalize PR. Figures 8 and 9 shows the modified Pagerank by using $f(age) = (1 + A * e^{-B*age})$, $q = 0.15$, and different values of A and B.

Another possibility would be to take in account the age of the page pointing to i. That is,

$$PR_i = q + (1-q) \sum_{j=1,\ j \neq i}^{k} \frac{f(age_{m_j})\ PR_{m_j}}{F_{m_j}}$$

where $F_(j) = \sum_{pages\ k\ linked\ by\ j} f(age_k)$ is the total weight of the links in a page. The result does not change to much, but the computation is slower.

Yet another approach would be to study how good are the links based in the modification times of both pages involved in a link. Suppose that page P_1 has an actualization date of t_1, and similarly t_2 and t_3 for P_2 and P_3, such that $t_1 < t_2 < t_3$. Let's assume that P_1 and P_3 reference P_2. Then, we can make the following two observations:

1. The link (P_3, P_2) has a higher value than (P_1, P_2) because at time t_1 when the first link was made the content of P_2 may have been different, although usually the content and the links of a page improves with time. It is true that the link (P_3, P_2) could have been created before t_3, but the fact that was not changed at t_3 validates the quality of that link.

2. For a smaller $t_2 - t_1$, the reference (P_1, P_2) is fresher, so the link should increase its value. On

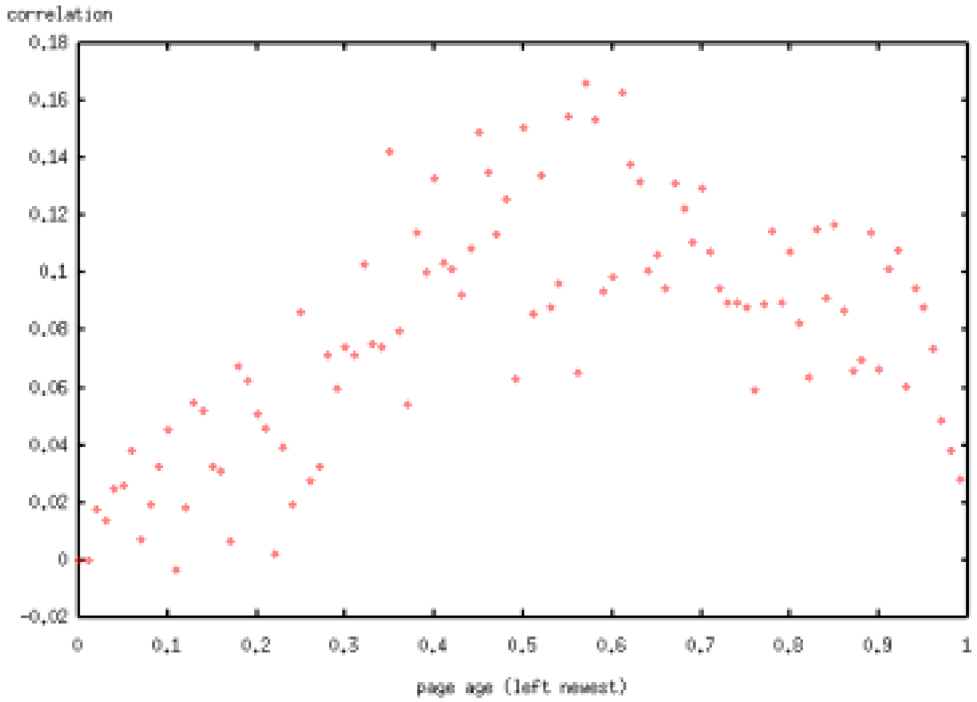

Figure 6: Correlation among hubs and authorities with age.

the other hand, the value of the link (P_3, P_2) should not depend on $t_3 - t_2$ unless the content of P_2 changes.

A problem with the assumptions above is that we do not really know when a link was changed and that they use information from the servers hosting the pages, which is not always reliable. These assumptions could be strengthened by using the estimated rate of change of each page.

Let $w(t, s)$ be the weight of a link from a page with modification time t to a page with modification time s, such that $w(t, s) = 1$ if $t \geq s$ or $w(t, s) = f(s - t)$ otherwise, with f a fast decreasing function. Let W_j be the weight of all the out-links of page j, then we can modify Pagerank using:

$$PR_i = q + (1 - q) \sum_{j=1,\ j\neq i}^{k} \frac{w(t_j, t_i)\ PR_{m_j}}{W_{m_j}}$$

where t_j is the modification time of page j. One drawback of this idea is that changing a page may decrease its Pagerank.

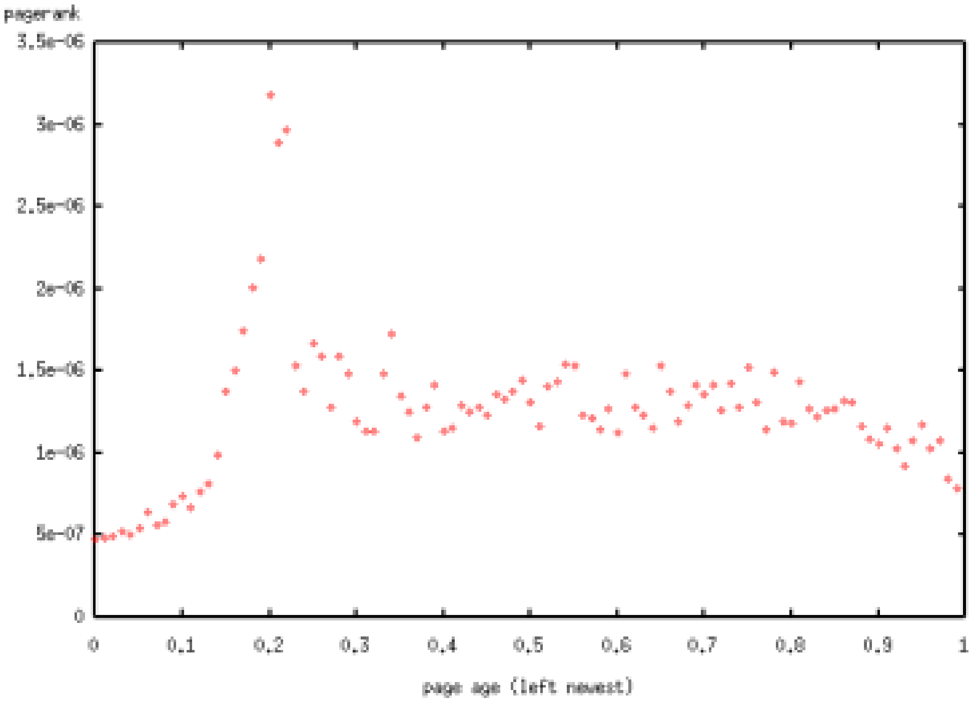

Figure 7: Pagerank as a function of page age.

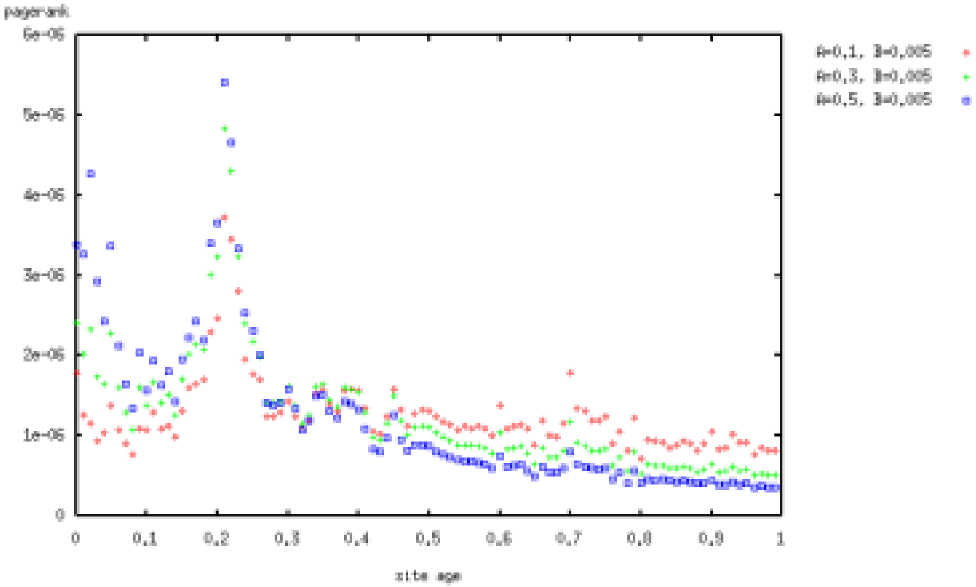

Figure 8: Modified PageRank taking in account the page age (constant B).

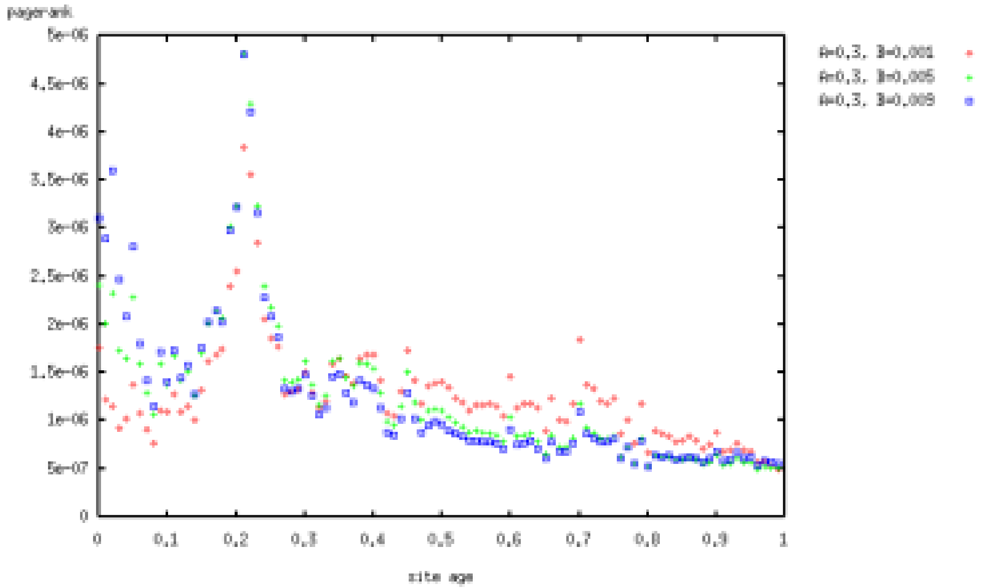

Figure 9: Modified PageRank taking in account the page age (constant A).

6 Conclusions

In this paper we have shown several relations between the macro structure of the Web, page and site age, and quality of pages and sites. Based on these results we have presented a modified Pagerank that takes in account the age of the pages. Google might be already doing something similar according to a BBC article[3] pointed by a reviewer, but they do not say how. We are currently trying other functions, and we are also applying the same ideas to hubs and authorities.

There is lot to do for mining the presented data. Further work includes how to evaluate the real goodness of a Web page link based ranking. Another line of research includes the analysis of search engines logs to study user behavior with respect to time.

References

[1] Google search engine: Main page. http://www.google.com/, 1998.

[2] Akwan search engine: Main page. `http://www.akwan.com`, 2000.

[3] Nua internet - how many online. `http://www.nua.ie/surveys/how_many_online/`, 2001.

[4] Netcraft web server survey. http://www.netcraft.com/survey/, 2002.

[5] Baeza-Yates, R., and Castillo, C. Relating web characteristics with link analysis. In *String Processing and Information Retrieval* (2001), IEEE Computer Science Press.

[3] http://news.bbc.co.uk/hi/english/sci/tech/newsid_1868000/1868395.stm (in a private communication with Google staff they said that journalist had a lot of imagination.

[6] Brewington, B., Cybenko, G., Stata, R., Bharat, K., and Maghoul, F. How dynamic is the web? In *9th World Wide Web Conference* (2000).

[7] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., and Tomkins, A. Graph structure in the web: Experiments and models. In *9th World Wide Web Conference* (2000).

[8] Cho, J., and Garcia-Molina, H. The evolution of the web and implications for an incremental crawler. In *The VLDB Journal* (2000).

[9] Douglas, F., Feldmann, A., Krishnamurthy, B., and Mogul, J. Rate of change and other metrics: a live study of the world wide web. In *USENIX Symposium on Internet Technologies and Systems* (1997).

[10] Kleinberg, J. Authoritative sources in a hyperlinked environment. In *9th Symposium on discrete algorithms* (1998).

[11] Levene, M., and Poulovassilis, A. Report on international workshop on web dynamics, london, january 2001.

[12] Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation algorithm: bringing order to the web. In *7th World Wide Web Conference* (1998).