1

Web Dynamics, Structure, and Page Quality*

Ricardo Baeza-Yates Carlos Castillo Felipe Saint-Jean

Center for Web Research
Computer Science Department, University of Chile
Blanco Encalada 2120, Santiago, Chile
E-mail: {rbaeza,ccastill,fsaint}@dcc.uchile.cl

Summary. This chapter is aimed at the study of quantitative measures of the relation between the dynamics of the Web, its structure, and the quality of Web pages. Quality is studied using different link-based metrics and considering 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, and we obtain information on how recency is related with Web structure.

1.1 Introduction

The purpose of a Web search engine is to provide an infrastructure that supports relationships between publishers of content and readers. In this context, as the numbers involved are very big (550 million users [2] and more than 3 billion pages (a lower bound that comes from the coverage of popular search engines) in 35 million sites [4] on January 2003) it is critical to provide good measures of quality that allow the user to choose "good" pages. We think that this is the main element that explain Google's [3] success. However, the notion of what is a "good page" and how this it is related to different Web characteristics is not well understood.

Therefore, in this chapter we address the study of the relationships between the age of a page or a site, the quality of a page, and the structure of the Web. Age is defined as the time since the page was last updated (recency). 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 website in the structure of the Web depends on the website age? Does the quality of a Web page depend on where it 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 found that the answer is "no" for Pagerank [25], which is very important among the ranking functions used by Google [3].

* Supported by Millenium Nucleus "Center for Web Research" (P01-029-F), Mideplan, Chile.

Our study is focused on the Chilean Web, mainly the .cl domain at two different time instants: first half of 2000, when we collected 670 thousand pages in approximately 7,500 websites (CL-2000), and the last half of year 2001, when we collected 795 thousand pages, corresponding to approximately 21.200 websites (CL-2001). This data comes from the TodoCL search engine [5] which specializes on the Chilean Web and is part of a family of vertical search engines built using the Akwan search engine [1].

Most statistical studies about the Web are based either on a "random" subset of the complete Web, or on the contents of some websites. In our case, the results are based on a Web collection that represents a large % of the Chilean Web, so we believe that our sample is more homogeneous and coherent, because it represents a well defined cultural and linguistic context.

The remaining of this chapter is organized as follows. Section 2 presents models of how a page changes. Section 3 shows how to measure the rate of change in the Web. Section 4 introduces the main results on the Web structure. Section 5 presents several relations among Web structure and age. Section 6 presents the relation of quality of webpages and age (recency). We end with some conclusions and future work.

1.2 Models of Page Change

Crawling the World Wide Web resembles, to some extent, the task of an astronomer that watches the sky. What he sees is not a "snapshot" of the instant state of the universe, but rather the result of light traveling from different distances to him. Most of the stars have experienced huge changes or even disappeared by the time their light reaches the earth [7].

In the same way, by the time a web crawler has finished its crawl, many events can happen on the web. The following discussion applies to any resource on the public web space, such as web pages, documents, images, audio, but we will focus on text. We are interested in the following events:

Creations

When a page is created, it will not be visible on the public web space until it is linked, so we asume that at least one page update -adding a link to the new web page- must occur for a web page creation to be visible.

Updates

Updates are difficult to characterize: an update can be either *minor* -at the paragraph or sentence level, so the page is still almost the same and references to its content are still valid- or *major* -all references to its content are not valid anymore. It is customary to consider **any** update as *major*, as it is difficult to judge automatically if the page's content is semantically the same. Partial changes characterization is studied in [24].

Deletions

A page is deleted if it is removed from the public web, or if all the links to that page are removed. Undetected deletions are more damaging for a search engine's reputation than updates, as they are more evident to the user.

In all cases, there is a cost associated with not detecting an event, and having an outdated copy of a resource. The most used cost functions, defined in [15], are freshness and age:

$$F(page_i;t) = \begin{cases} 1 & \text{if } page_i \text{ is up} - \text{to} - \text{date at time } t \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

Age

In this case, an up-to-date page has a value of 1 and an outdated page has a value that degrades lineary with time:

$$A(page_i;t) = \begin{cases} 0 & \text{if } page_i \text{ is up} - \text{to} - \text{date at time } t \\ t - \text{modification time of } page_i & \text{otherwise} \end{cases} \tag{1.2}$$

The evolution of these two quantities is depicted in Figure 1.1.

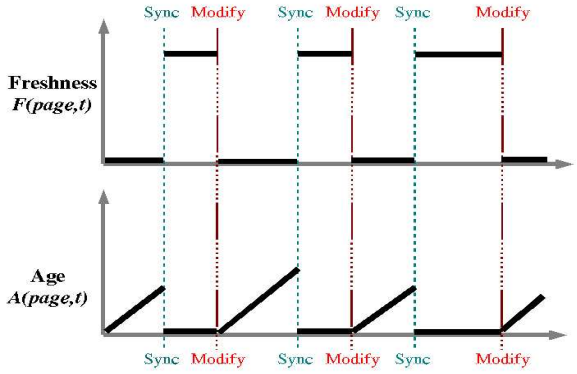

Fig. 1.1. Evolution of freshness and age with time. Two kinds of event can occur: modification of a web page in the server, and downloading of the modified page by the crawler (*sync*). When the modified page is downloaded, freshness becomes 1 and age becomes 0.

We gathered time information (last-modified date) for each page as informed by the webservers. We focus on webpage age, that is, the time elapsed after the last modification (recency). As the Web is young, we use months as time unit, and our data considers only the three last years as most websites are that young. The distribution of pages and sites for CL-2000 with respect to age is given in Figure 1.4.

1.3 Measuring and Estimating the Rate of Change

Sampling

One of the main difficulties involved in any attempt of Web characterization studies is how to obtain a good sample. As there are very few important pages lost in a vast amount of unimportant pages (according to any metric: pagerank, reference count, page size, etc.), just taking a URL at random is not enough. Pages must be sampled based on their importance, and not all pages are known in advance, so a way of estimating the importance of each page based on partial information is needed [20].

There are two aproaches for gathering data: using a log of the transactions in the proxy of a large organization or ISP, or using a Web crawler. There are advantages and disadvantages for each method: when monitoring a proxy it is easy to find popular pages, but the revisit period is impossible to control, as it depends on users; using a crawler the popularity of pages has to be estimated but the revisit period can be fine-tuned.

Data

The data is obtained by repeated access to a large set of pages during a period of time. Notice that in all cases the results will be only estimation because they are obtained by **polling** for events (changes), not by the resource **notifying** events.

For each $page_i$ and each visit there is:

- The access timestamp of the page $visit_i$.
- The last-modified timestamp (given by most web servers; about 80%-90% of the requests in practice) $modified_i$.
- The text page itself, that can be compared to an older copy to detect changes, specially if $modified_i$ is unknown.

There are other data that can be estimated sometimes, specially if the revisiting period is short:

- The creation timestamp of the page, $created_i$.
- The delete timestamp of the page, when the page is no longer available, $deleted_i$.

Metrics

There are different time-related metrics for a web page, the most common are:

- Age, $visit_i - modified_i$.
- Lifespan, $deleted_i - created_i$.
- Number of modifications during the lifespan, $changes_i$.
- Change interval, $lifespan_i / changes_i$.

For the entire web or for a large collection, useful metrics are:

- Distribution of change intervals.

- Time it takes for 50% of the web to change.
- Average lifespan of pages.

One of the most important metrics is the change interval; Figure 1.2 was obtained in a study from 2000 [13]. An estimation of the average change interval is about 4 months.

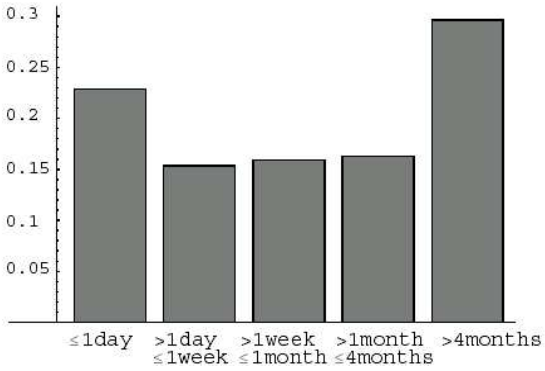

Fig. 1.2. Fraction of pages with given average interval.

Estimating

The probability that a copy of $page_i$ is up-to-date at time t, $p_{i,t} = P(F(page_i; t) = 1)$ decreases with time. Page change can be modeled as a Poisson process [10], so if t units of time has passed since the last visit:

$$p_{i,t} = e^{-\lambda_i t} \tag{1.3}$$

The parameter λ_i characterizes the rate of change of the page i and can be estimated based on previous observations, specially if the web server gives the last modification date of the page each time it is visited. The estimation for λ_i is [14]:

$$\lambda_i \approx \frac{(X_i - 1) - \frac{X_i}{N_i \log(1 - X_i/N_i)}}{S_i T_i} \tag{1.4}$$

- N_i number of visits to $page_i$.
- S_i time since the first visit to $page_i$.
- X_i number of times the server has informed that the page has changed.
- T_i total time with no modification, according to the server, summed for all the visits.

If the server does not give the last-modified time, we can still check for modifications by comparing the downloaded copies at two different times, so X_i now will be the number of times a modification is detected. The estimation for the parameter in this case is:

$$\lambda_i \approx \frac{-N_i \log(1 - X_i/N_i)}{S_i} \tag{1.5}$$

There have been many large-scale studies about web dynamics in 1997 [18], 1999 [21], 2000 [10, 11] and a recent study in 2003 [19].

An important motivation for them is determining an optimal scheduling policy for a web crawler, as in the works by Cho [13, 16, 14, 15] and Coffman [17]; in this area, there are some proposals for changing from a polling scheme to a notify scheme [26], in which servers cooperate with crawlers [9].

1.4 Structure of the Web

The most complete study of the Web structure [12] focus on the connectivity of web pages. This study starts with a large-scale web graph and then identifies a single large strongly connected component (MAIN). All the structures in the graph are related to MAIN, as explained later.

A page can describe several documents and one document can be stored in several pages, so we decided to study the structure of how websites were connected, as websites are closer to real logical units. Not surprisingly, we found in [6] that the structure in the .cl (Chile) domain at the website level was similar to the global Web -another example of the autosimilarity of the Web, which gives a scale invariant- and hence we use the same notation of [12]. The components are:

(a) 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);
(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, where T is an abbreviation of tentacle), sites in paths between IN and OUT (TUNNEL), sites that only reach OUT (T.OUT), and unconnected sites (ISLANDS).

In [6] we analyzed CL-2000 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;
(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 1.3 shows all these components.

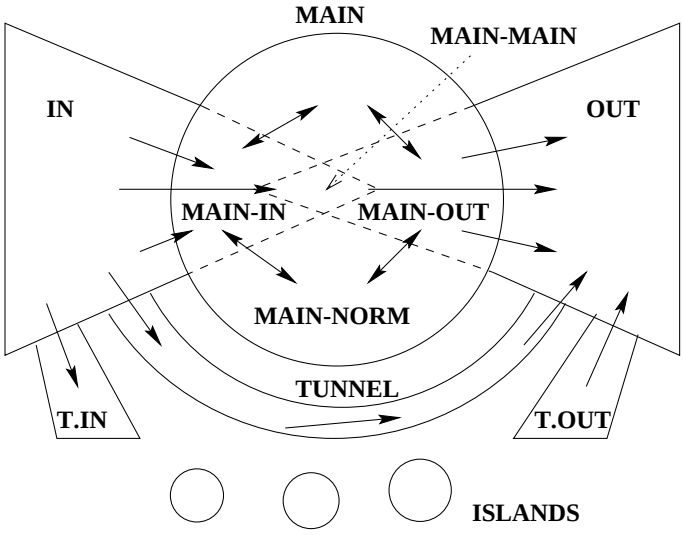

Fig. 1.3. Structure of the Web.

An important observation is that in random graphs the size of the strongly connected component tends to be much larger, over 90% of the nodes. This is not the case of the Web. This can be explained because the Web is not a random graph: it grows in a way that is compatible with models of preferred attachment, in which a highly linked page tends to attract even more links [22]. However, these models do not consider the part of the Web also disappears. More on generative models for the Web can be found in the previous chapter.

1.5 Web Structure and Age

An initial motivation is to find if the IN and OUT components were related to Web dynamics or just due to bad website design. In fact, websites 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 1.5 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 [12] for pages). The central part is a line and represents the typical power laws that appear in many Web measures.

Figure 1.6 shows the relation between the macro-structure of the Web using the number of websites in each component to represent the area of each part of the

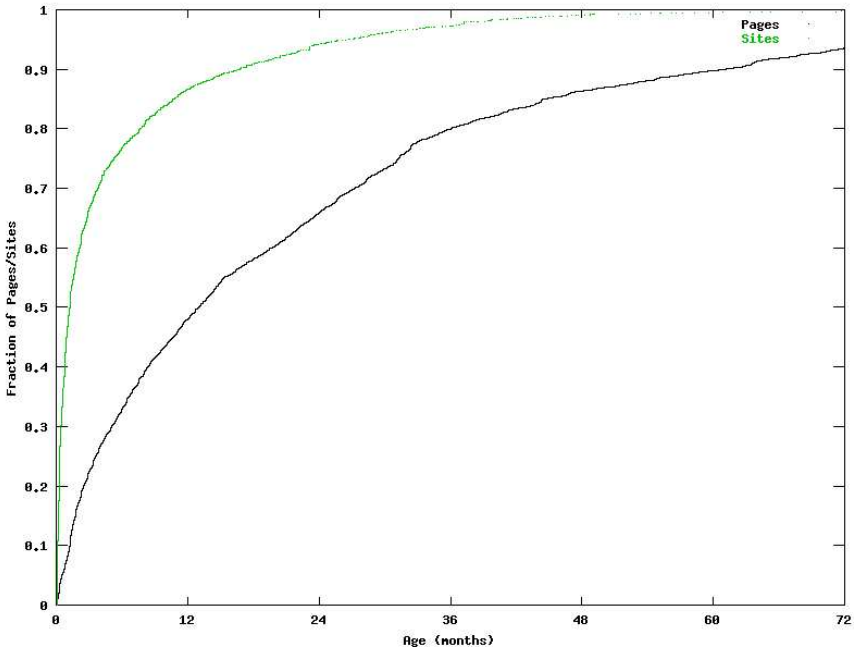

Fig. 1.4. Cumulative distribution of pages (bottom) and sites (top) in function of age for CL-2000.

diagram for CL-2000. The gray scale represent website age, such that a darker tone represents older websites. We consider three ages: the oldest page that is a lower bound to the website age, the average age that can be considered as the freshness of a site, and the newest page which is a measure of update frequency on a site. Figure 1.6 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.

We also obtain some confirmation to what can be expected. Newer sites are in the ISLANDS 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: we have observed that some of these websites are from e-commerce sites whose policy is not to link other websites.

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 1.1 we give the numerical data for the average of website age (using the oldest page) as well as the overall 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 CL-2000 did not include all the ISLANDS

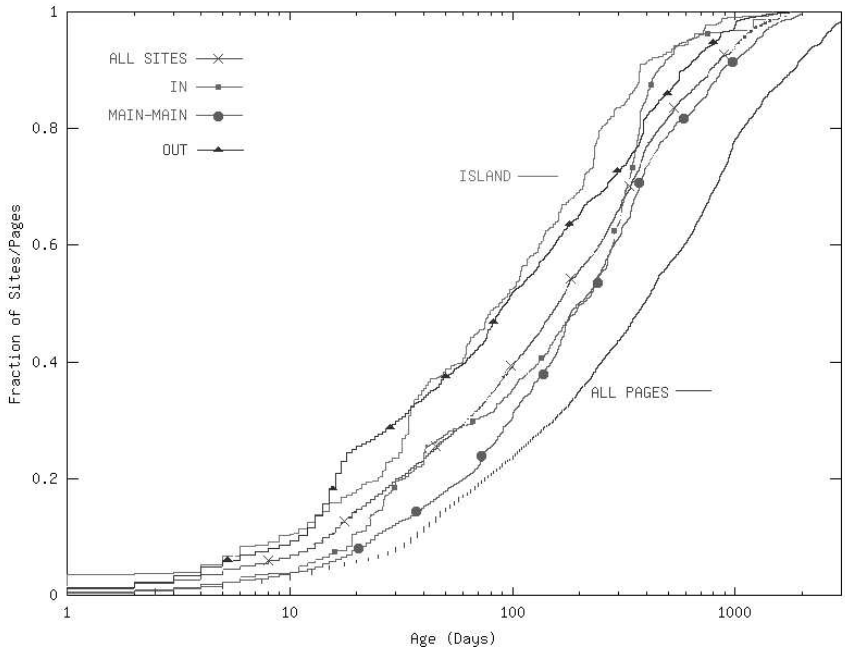

Fig. 1.5. Website age in the different components and webpage age (rightmost curve).

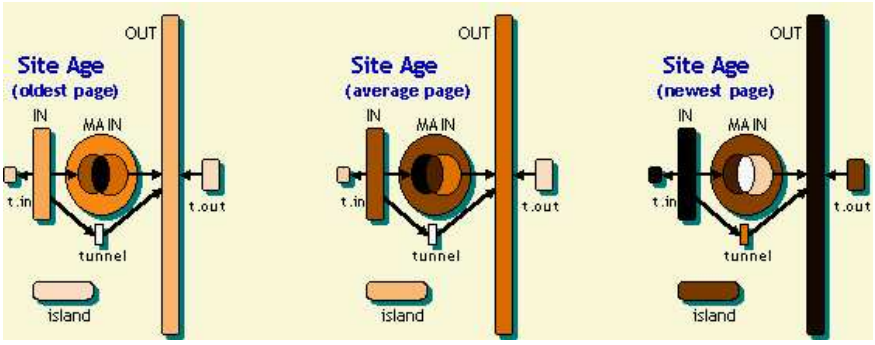

Fig. 1.6. Visualization of Web structure and website age.

at that time (we estimate that CL-2000 was 70% of the sites), we can compare the core. The core has the smaller percentage but it is larger as CL-2001 triples the number of sites of CL-2000. 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, CL-2001 represents a Web much more connected than CL-2000.

Several observations can be made from Table 1.1. First, sites in MAIN have the higher Pagerank, and inside it, MAIN-MAIN is the subcomponent with high-

Component	size (%,CL-2000)	size (%,CL-2001)	age (days)
MAIN	23%	9%	429
IN	15%	6%	295
OUT	45%	20%	288
TUNNEL	1%	0%	329
T.IN	3%	3%	256
T.OUT	9%	2%	293
ISLANDS	4%	60%	273
MAIN-MAIN	2%	3%	488
MAIN-OUT	6%	2%	381
MAIN-IN	3%	1%	420
MAIN-NORM	12%	2%	395

Table 1.1. Distribution and age CL-2001 in the different components of the macro-structure of the Chilean Web.

est 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. ISLANDS have a low score in all cases.

Studying age, sites in MAIN are the oldest, and inside it, sites in MAIN-MAIN are the oldest. As MAIN-MAIN also has good quality, it 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.

Table 1.2 indicates the percentage of sites coming from CL-2000 in each part of the structure for CL-2001. Note that some are new sites (NEW) and that other sites from CL-2000 have disappeared (GONE). Figure 1.7 shows this change graphically using relative percentage changes in each component and absolute percentage changes in the arrows. Each column has new sites and the distribution of from where are coming old sites. The main flows are from MAIN, IN and OUT to ISLANDS or from MAIN to OUT (probably sites that become outdated), and sites that disappear in OUT and ISLANDS (probably new sites that were not successful). On the other hand, it is interesting to notice the stability of the MAIN component. At the same time, all these changes show that the Web is very unstable as a whole.

Therefore there is a strong relation between the macro-structure of the Web and age/quality characteristics. This implies that the macro-structure is a valid partition of websites regarding these characteristics.

1.6 Link-based Ranking and Age

The two main link based ranking algorithms known in the literature are Pagerank [25] and the hub and authority measures of the HITS algorithm [23].

Pagerank is based on the probability of a random surfer to be visiting a page. This probability is modeled with two actions: the chance of the surfer to get bored

2000 - 2001	MAIN	IN	OUT	ISLANDS	GONE
MAIN	36.36	5.31	27.46	11.57	19.30
IN	5.19	15.71	11.85	37.15	30.09
OUT	8.12	1.62	31.21	31.21	27.83
ISLANDS	3.31	2.58	22.84	39.23	32.04
NEW	5.20	6.30	14.40	74.10	
Rest	3.79	11.76	29.41	1.26	53.78

Table 1.2. Relative percentage of sites coming from different parts of the structure, including new sites and sites that have disappeared among CL-2000 and CL-2001.

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:

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

where T is the total number of webpages, 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 scores 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.

Hub and authorities were defined for subsets of the web graph induced by a textual query, and this is usually called "dynamic ranking". In this study, we use them in terms of the whole graph ("static ranking"). Table 1.3 shows the average page static ranking in each component of the structure.

When considering the rank of a website, 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 in the case of Pagerank [6]. Using this is fairer that using the best page or the average of all pages of a site.

In [6] 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.

Webpages sorted by recency were divided in 100 group segments of the same weight (that is, each segment has the same number of pages), obtaining a time division that is not uniform. Then we calculated the standard correlation (which is defined as $\hat{\rho}(x,y) \equiv \frac{\hat{cov}(x,y)}{\hat{\sigma}_x \hat{\sigma}_y}$ where x and y are two random variables, $\hat{cov}$ is the covariance, and σ is the standard deviation) of each pair of average rank values. Three graphs were obtained: Figure 1.8 which shows the correlation between Pagerank and authority, Figure 1.9 the correlation among Pagerank and hub, and Figure 1.10 shows

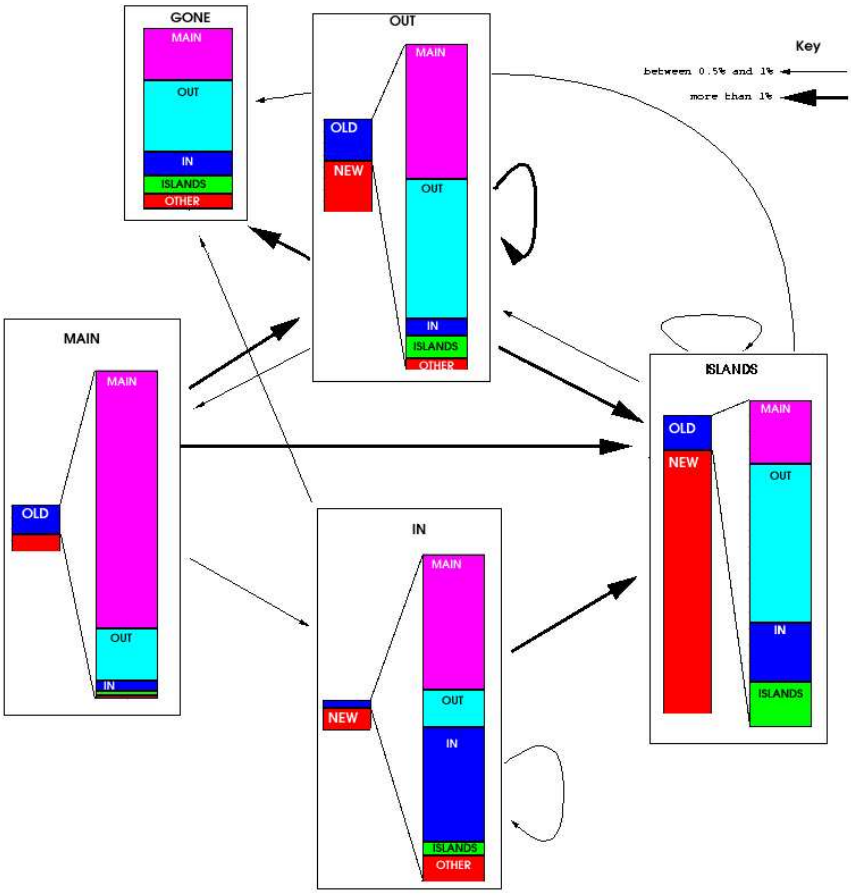

Fig. 1.7. Site flow among the Web structure from 2000 to 2001 (the left column indicates new and old sites, the second column indicates from where the old sites come.)

the correlation of authorities and hubs. Notice that the horizontal axis has two scales: at the top the fraction of groups and at the bottom the recency in months.

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. This is true also for other ranking schema based in incoming links. We show that this intuition is

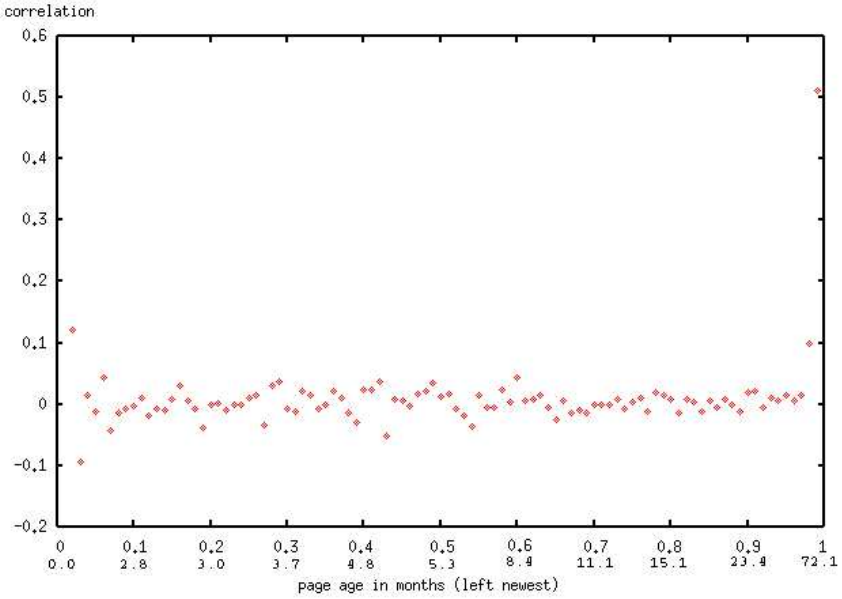

Fig. 1.8. Correlation among Pagerank and authority with age.

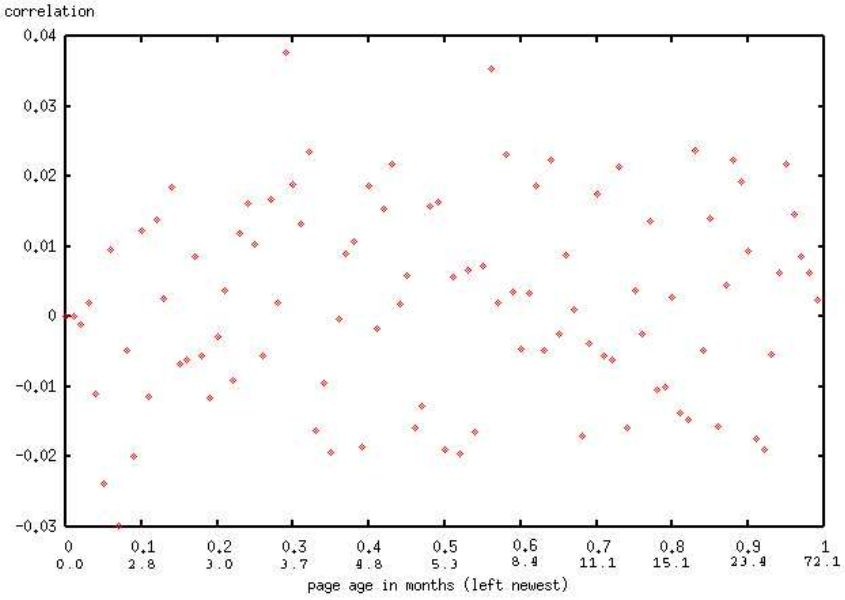

Fig. 1.9. Correlation among Pagerank and hub with age.

Component	Pagerank	Hub	Authority
MAIN	2e-04	53e-04	9e-04
IN	0.8e-04	542e-04	1e-07
OUT	0.6e-04	1e-07	0.1e-04
TUNNEL	0.2e-04	1e-07	0e-07
T.IN	0.3e-04	0e-07	0e-07
T.OUT	0.4e-04	0e-07	0e-07
ISLANDS	0.1e-04	0e-07	0e-07
MAIN-MAIN	3e-04	144e-04	25e-04
MAIN-OUT	1e-04	1e-04	4e-07
MAIN-IN	1e-04	0e-07	98e-07
MAIN-NORM	0.8e-04	0e-07	2e-07

Table 1.3. Page quality for CL-2001 in the different components of the macro-structure of the Chilean Web.

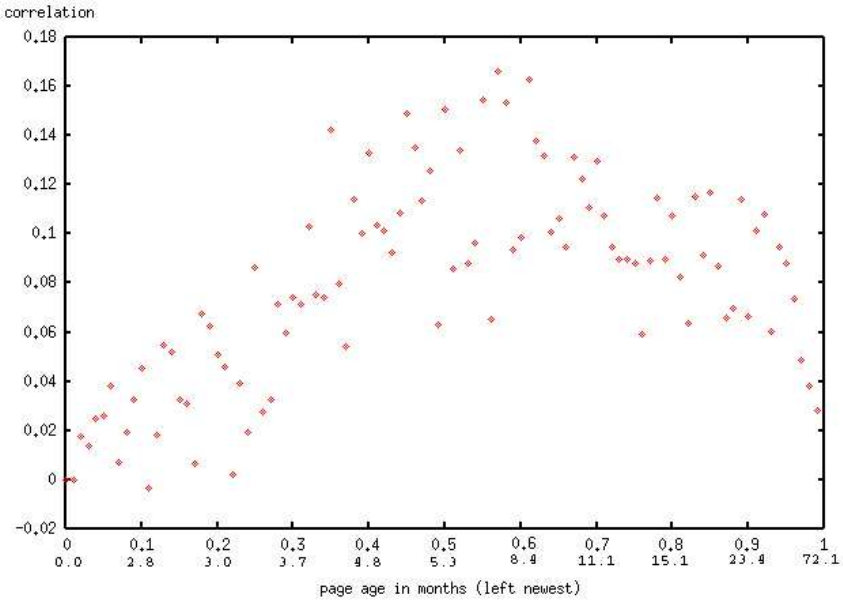

Fig. 1.10. Correlation among hubs and authorities with age.

correct in Figure 1.11, where Pagerank is plotted against percentiles of page age. As can be seen, the newest pages have a very low Pagerank, similar to very old pages. The peak of Pagerank is in three months old pages.

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. An age based Pagerank based on these results is presented in

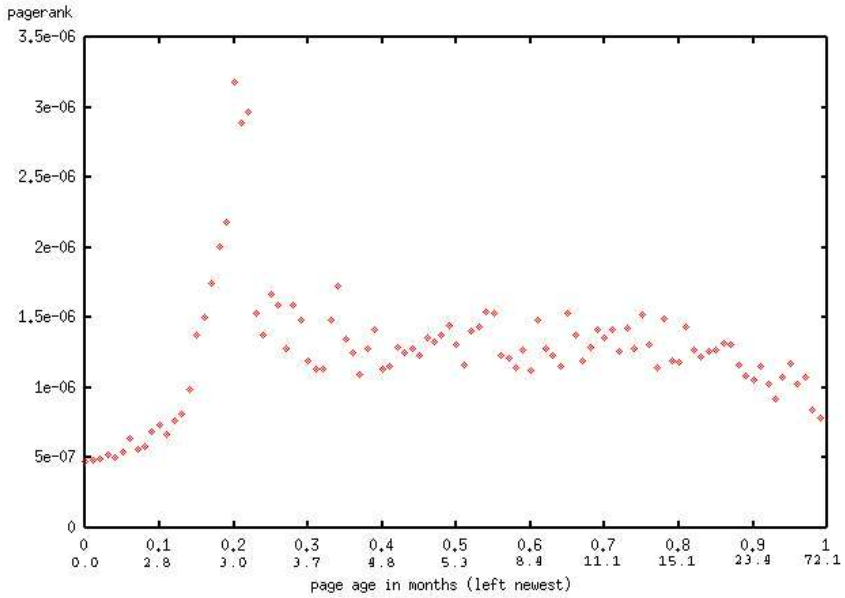

Fig. 1.11. Pagerank as a function of page age.

[8], using the following expression

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

where $f(age_i)$ is the last-modified date of the page. In [8] we use $f(age) = (1 + A * e^{-B*age})$ with positive values of A and B to obtain a Pagerank scheme biased to new pages. The idea is that each time a page is changed, implies a reaffirmation of the links on that page. Hence if the page is new, the value of the link is $1 + A$. If the page is never updated, the value of the link tends to 1.

1.7 Conclusions

In this chapter we have shown several relations between the macro structure of the Web, page and site age, and quality of pages and sites. There is plenty to do for mining the presented data and this is just the beginning of this kind of Web mining. For example, it is usually assumed that Pagerank never decreases, as incoming links should only increase. However, as shown previously, part of the Web disappears, so incoming links to a page may decrease. Including Web death in generative models and analysis of ranking measures are interesting open problems.

Further related work includes how to evaluate the real goodness of a webpage link based ranking and the analysis of search engine logs to study user behavior with respect to time.

References

1. Akwan search engine: Main page. `http://www.akwan.com`, 1999.
2. Nua internet - how many online. www.nua.ie/surveys/how_many_online/, 2001.
3. Google search engine. www.google.com/, 2002.
4. Netcraft web server survey. www.netcraft.com/survey/, 2002.
5. Todocl - todo chile en internet. www.todocl.cl/, 2002.
6. BAEZA-YATES, R., AND CASTILLO, C. Relating web characteristics with link based web page ranking. In *String Processing and Information Retrieval* (2001), IEEE Cs. Press.
7. BAEZA-YATES, R., AND RIBEIRO-NETO, B. *Modern Information Retrieval*. Addison-Wesley & ACM Press, Harlow, UK, 1999.
8. BAEZA-YATES, R., SAINT-JEAN, F., AND CASTILLO, C. Web structure, dynamics and page quality. In *Proceedings of the 9th Symposium on String Processing and Information Retrieval (SPIRE'2002)* (2002), Springer Lecture Notes in Computer Science.
9. BRANDMAN, O., CHO, J., GARCIA-MOLINA, H., AND SHIVAKUMAR, N. Crawler-friendly web servers. In *Proceedings of the Workshop on Performance and Architecture of Web Servers (PAWS)* (Santa Clara, California, 2000).
10. BREWINGTON, B., CYBENKO, G., STATA, R., BHARAT, K., AND MAGHOUL, F. How dynamic is the web? In *World Wide Web Conference* (2000).
11. BREWINGTON, B. E., AND CYBENKO, G. Keeping up with the changing Web. *Computer 33*, 5 (2000), 52–58.
12. 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).
13. CHO, J. The evolution of the web and implications for an incremental crawler. In *The VLDB Journal* (2000), pp. 200 – 209.
14. CHO, J., AND GARCIA-MOLINA, H. Estimating frequency of change. In *Technical Report* (2000).
15. CHO, J., AND GARCIA-MOLINA, H. Synchronizing a database to improve freshness. In *ACM International Conference on Management of Data (SIGMOD)* (2000), pp. 117–128.
16. CHO, J., AND NTOULAS, A. Effective change detection using sampling. In *Proceedings of the 28th Very Large Databases Conference* (2002).
17. COFFMAN, E. G., LIU, Z., AND WEBER, R. R. Optimal robot scheduling for web search engines. Tech. Rep. RR-3317, INRIA, 1997.
18. DOUGLIS, F., FELDMANN, A., KRISHNAMURTHY, B., AND MOGUL, J. C. Rate of change and other metrics: a live study of the world wide web. In *USENIX Symposium on Internet Technologies and Systems* (1997).
19. FETTERLY, D., MANASSE, M., NAJORK, M., AND WIENER, J. L. A large-scale study of the evolution of web pages. In *World Wide Web Conference* (2003).
20. HENZINGER, M., HEYDON, A., MIZENMACHER, M., AND NAJORK, M. On near–uniform url sampling. In *World Wide Web Conference* (2000).
21. HUBERMAN, B., AND ADAMIC, L. Evolutionary dynamics of the world wide web. Tech. rep., Xerox Palo Alto Research Center, 1999.

22. HUBERMAN, B., AND ADAMIC, L. Growth dynamics of the world wide web. *Science* (1999).
23. KLEINBERG, J. Authoritative sources in a hyperlinked environment. In *9th Symposium on discrete algorithms* (1998).
24. LIM, L., WANG, M., PADMANABHAN, S., VITTER, J. S., AND AGARWAL, R. Characterizing Web document change. *Lecture Notes in Computer Science 2118* (2001).
25. PAGE, L., BRIN, S., MOTWAIN, R., AND WINOGRAD, T. The pagerank citation algorithm: bringing order to the web. In *7th World Wide Web Conference* (1998).
26. REDDY, S., AND FISHER, M. Event notification protocol. Tech. rep., WEBDAV Working Group Internet Draft, 1998.