A STUDY ON THE EVOLUTION OF THE WEB

Alexandros Ntoulas[1], Junghoo Cho[1], Hyun Kyu Cho[2], Hyeonsung Cho[2], and Young-Jo Cho[2]

Summary

We seek to gain improved insight into how Web search engines should cope with the evolving Web, in an attempt to provide users with the most up-to-date results possible. For this purpose we collected weekly snapshots of some 150 Web sites over the course of one year, and measured the evolution of content and link structure. Our measurements focus on aspects of potential interest to search engine designers: the evolution of link structure over time and the rate of creation of new pages on the Web. Our findings indicate a rapid turnover rate of Web pages, *i.e.*, high rates of birth and death, coupled with an even higher rate of turnover in the hyperlinks that connect them. We conclude the paper with a discussion of the potential implications of our results for the design of effective Web search engines.

Introduction

As the Web grows larger and more diverse, search engines are becoming the "killer app" of the Web. Whenever users want to look up information, they typically go to a search engine, issue queries and look at the results. Recent studies confirm the growing importance of search engines. According to [3], for example, Web users spend a total of 13 million hours per month interacting with Google alone.

Search engines typically "crawl" Web pages in advance to build local copies and/or indexes of the pages. This local index is then used later to identify relevant pages and answer users' queries quickly. Given that Web pages are changing constantly, search engines need to update their index periodically in order to keep up with the evolving Web. An obsolete index leads to irrelevant or "broken" search results, wasting users' time and causing frustration. In this paper, we study the evolution of the Web from the perspective of a search engine, so that we can get a better understanding on how search engines should cope with the evolving Web. We believe that the following aspects make our study unique, revealing new and important details of the evolving Web:

- *New pages on the Web*: While a large fraction of existing pages change over time, a significant fraction of "changes" on the Web are due to new pages that are created over time. In this paper, we study how many new pages are being created every week and what are the characteristics of the newly-created pages.
- *Link-structure evolution*: Search engines rely on both the content and the link structure of the Web to select the pages to return. For example, Google uses PageRank as their primary ranking metric, which exploits the Web link structure to evaluate the importance of a page [8]. In this respect, the evolution of the link structure is an important aspect that search engines should know, but not much work has been done before. As far as we know, our work is the first study investigating the evolution of the link structure.

In this paper, we study the above aspects of the evolving Web, by monitoring pages in 154 Web sites on a weekly basis for one year and analyzing the evolution of these sites. We can summarize some of the main findings from this study as following:

- We estimate that new pages are created at the rate of 8% per week. Assuming that the current Web has 8 billion pages [2], this result corresponds to 640 million new pages every week, which

[1]University of California Los Angeles (UCLA), Los Angeles, CA 90095, USA

[2]Electronics and Telecommunications Research Institute (ETRI), 161 Gajeong-Dong, Yuseong-Gu, Daejeon, 305-350, Republic of Korea

is roughly 7.6 terabytes in size.[1] We also estimate that only 40% of the pages available today will be still accessible after one year. Given this result, we believe that creation and deletion of new pages is a very significant part of the changes on the Web and search engines need to dedicate substantial resources detecting these changes.

- The link structure of the Web is significantly more dynamic than the content on the Web. Every week, about 25% new links are created. After a year, about 80% of the links on the Web are replaced with new ones. This result indicates that search engines need to update link-based ranking metrics (such as PageRank) very often. Given 25% changes every week, a week-old ranking may not reflect the current ranking of the pages very well.

Experimental setup

To collect Web history data for our evolution study, we downloaded pages from 154 "popular" Web sites (e.g., acm.org, hp.com, oreilly.com; see [5] for a complete listing) every week from October 2002 until October 2003, for a total of 51 weeks. In this section, we explain how we selected the sites for our study and describe how we conducted the crawls of those sites. We also present a few general statistics about the data we collected.

In selecting the sites to monitor, we wanted to pick a "representative" yet "interesting" sample of the Web. By representative, we mean that our sample should span various parts of the Web, covering a multitude of topics.[2] By interesting, we mean that a reasonably large number of users should be interested in the sites, as search engines typically focus their resources on maintaining these sites the most up to date.

To obtain such a sample, we decided to pick roughly the five top-ranked pages from a subset of the topical categories of the Google Directory [1]. Google Directory reuses the data provided by the Open Directory Project [4], and maintains a hierarchical listing of Web sites categorized by topic. Sites within each category are ordered by PageRank, enabling users to identify sites deemed to be of high importance easily. By selecting sites from each topical category, we believe we made our sample "representative." By picking only top-ranked sites, we believe we make our sample "interesting." A complete list of sites included in our study can be acquired from [5].

From the 154 Web sites we selected for our study, we downloaded pages every week over a period of almost one year. Our weekly downloads of the sites were thorough in all but a few cases: starting from the root pages of the Web sites, we downloaded in a breadth-first order either *all* reachable pages in each site, or all pages until we reached a maximum limit of $200,000$ pages per site. Since only four Web sites (out of 154) contained more than $200,000$ pages[3], we have captured a relatively complete weekly history of these sites. Capturing nearly complete snapshots every week is important for our purposes, as one of our main goals is to study the creation of new pages on the Web.

The total number of pages that we downloaded every week ranges from 3 to 5 million pages, with an average of 4.4 million pages. The size of each weekly snapshot was around 65 GB before compression. Thus, we currently have a total of 3.3 TB of Web history data, with an additional 4 TB of derived data (such as links, shingles, etc.) used for our various analyses. When we compress the weekly snapshots using the standard zlib library, the space footprint is reduced to about one third of the original.

Figure 1 reports the fraction of pages included in our study that belong to each high-level domain. The `miscellaneous` category contains other domains including regional ones such as `.uk, .dk, .jp` etc. The distribution of domains for pages in our study roughly matches the general distribution of domains found on the Web [6].

[1] The average page size in the data collection we used for this paper was about 12KB.

[2] Our dataset constitutes a representative sample of the topical categories on the Web.

[3] The sites containing more than $200,000$ pages were `www.eonline.com` `www.hti.umich.edu`, `www.pbs.org` and `www.intelihealth.com`.

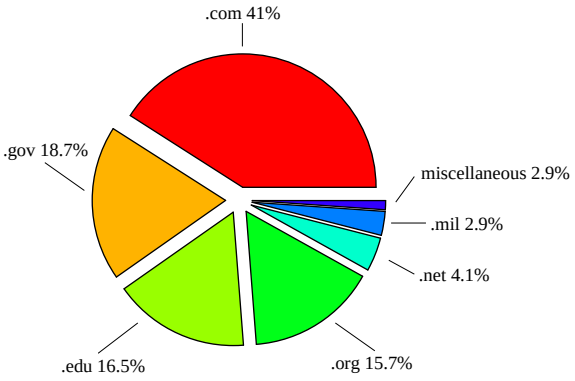

Figure 1: Distribution of domains in our crawls.

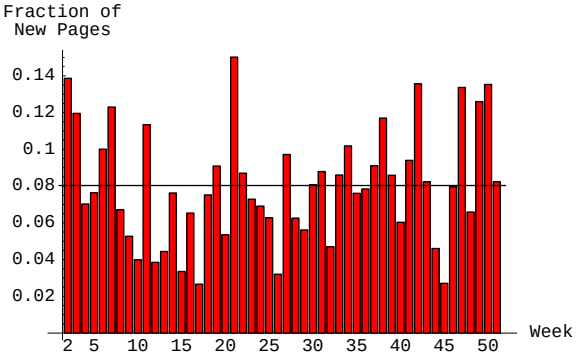

Figure 2: Fraction of new pages between successive snapshots.

Weekly birth, death and replacement rate of pages

In this section, we focus on measuring what is new on the Web each week. In particular, we attempt to answer questions such as: How many new pages are created every week? How many new links? We begin by studying the weekly birth rate of pages. For our analysis in the remainder of the paper we treat each unique URL as a distinct unit.

We first examine how many new pages are created every week. That is, for every snapshot, we measure the fraction of the pages in the snapshot that have not been downloaded before and we plot this number over time. This fraction represents the "weekly birth rate" of Web pages. We use the URL of a page as its identity, and consider a page "new" if we did not download any page with the same URL before. Under this definition, if a page simply changes its location from URL A to URL B, we consider that a new page B has been created. (In [14] we measure how much new "content" is introduced every week, which factors out this effect.)

In Figure 2 we show the weekly birth rate of pages, with week along the horizontal axis. The line in the middle of the graph gives the average of all the values, representing the "average weekly birth rate" of the pages. From the graph we can observe that the average weekly birth rate is about 8%. That is, 8% of pages downloaded by an average weekly crawl had not been downloaded by any previous crawl. Scaling up from our data (which, by design, is biased toward popular pages), and assuming the entire Web consists of roughly eight billion pages[4], we conjecture that there may be around 640 million new pages created every week (including copies of existing pages and relocated pages). Admittedly, this number may not be fully accurate because our study focuses on popular pages. However, it does give us a ball-park figure.

We also observe that approximately once every month, the number of new pages being introduced is significantly higher than in previous weeks. For example, the bars are higher in weeks 7, 11, 14, etc. than their previous weeks. Most of the weeks with the higher birth rate fall close to the end of a calendar month. This fact implies that many Web sites use the end of a calendar month to introduce new pages. Manual examination of the new pages in these "high birth rate" weeks revealed that a number of such pages contain job advertisements or portals leading to archived pages in a site. For the most part, however, we could not detect any specific pattern or topical category for these pages.

In our next experiment, we study how many new pages are created and and how many disappear over time.[5] We also measure what fraction of pages on our Web sites is replaced with new pages after a certain period. For these purposes, we compare our weekly snapshots of the pages against the first snapshot and measure 1) how many pages in the first snapshot still remain in the nth-week snapshot, and 2) how many pages in the nth week snapshot do not exist in the first snapshot. For all

[4] As reported by Google [2] at the time of this writing.

[5] We assume that a page disappeared if our crawler received an HTTP 404 response for that particular page, or we could not download the page (due to timeouts etc.) after three attempts.

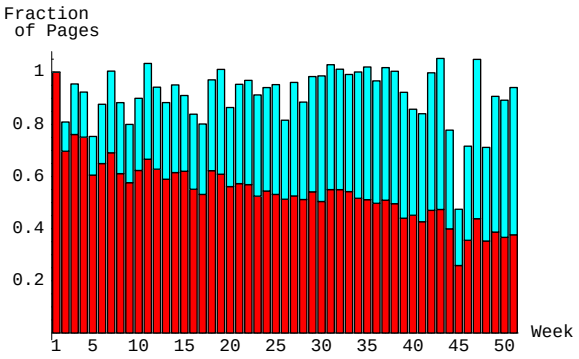

Figure 3: Fraction of pages from the first crawl still existing after n weeks (dark bars) and new pages (light bars).

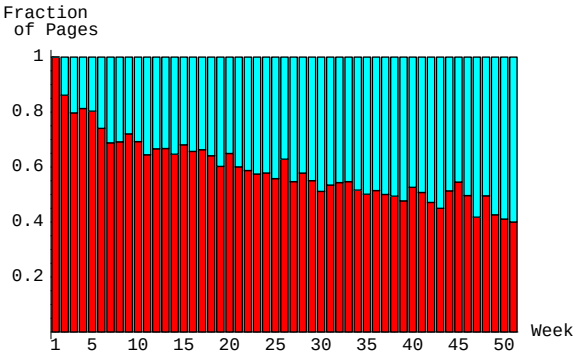

Figure 4: Normalized fraction of pages from the first crawl still existing after n weeks (dark bars) and new pages (light bars).

the comparisons presented here, the URLs of the crawled pages were canonicalized.

Figure 3 shows the result. The horizontal axis of this graph plots the week and the vertical axis shows the number of pages that we crawled in the given week. The bars are normalized such that the number of pages in the first week is one. (We downloaded 4.8 million pages in the first week.) The dark bars represent the number of first-week pages that were still available in the given week. The light bars represent the number of pages that were created since the first week (i.e., the pages that exist in the given week but did not exist in the first week). For example, the size of the second-week snapshot was about 80% of that of the first week, and we downloaded about 70% of the first-week pages in the second week.

The observable fluctuations in our weekly crawl sizes (most noticeable for week 45) are primarily due to technical glitches that are difficult to avoid completely. While collecting our data, to minimize the load on the Web sites and our local network, we ran our crawler in a slow mode. It took almost a full week for the crawler to finish each crawl. During this time, a Web site may have been temporarily unavailable or our local network connection may have been unreliable. To be robust against short-lived unavailabilities our crawler makes up to three attempts to download each page. Still, in certain cases unavailabilities were long-lived and our crawler was forced to give up. Since these glitches were relatively minor in most cases (except in the 45th week when one of our crawling machines crashed), we believe that our results are not significantly affected by them.

By inspecting the weeks with the highest bars in Figure 3 and taking glitches with our crawling into account, we find that the total number of pages available from the 154 sites in our study remained more or less the same over the entire 51-week period of our study. However, they are not all the same pages. Instead, existing pages were replaced by new pages at a rapid rate. For example, after one month of crawling (week 4), only 75% of the first-week pages were still available (dark portion of the graph at week 4), and after 6 months of crawling (week 25), about 52% are available.

A normalized version of our graph is shown in Figure 4, with the numbers for each week normalized to one to allow us to study trends in the fraction of new and old pages. After six months (week 25), roughly 40% of the pages downloaded by our crawler were new (light bars) and around 60% were pages that also occurred in our first crawl (dark bars). Finally, after almost a year (week 51) nearly 60% of the pages were new and only slightly more than 40% from the initial set was still available. It took about nine months (week 39) for half of the pages to be replaced by new ones (i.e., half life of 9 months).

To determine whether the deletion rate of pages shown in Figure 4 follows a simple trend, we used linear regression to attempt to fit our data using linear, exponential, and inverse-polynomial functions. The deletion rate did not fit any of these trends well. The best match was with a linear trend, but the R-squared value was still very low at 0.8.

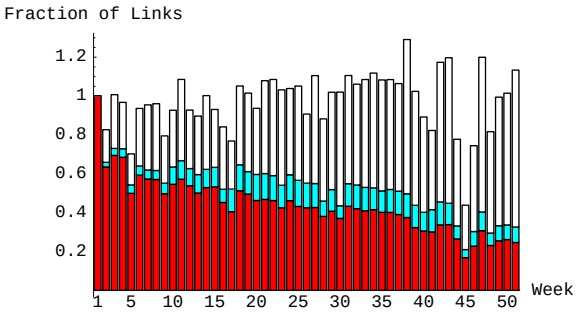

Figure 5: Fraction of links from the first weekly snapshot still existing after n weeks (dark/bottom portion of the bars), new links from existing pages (grey/middle) and new links from new pages (white/top).

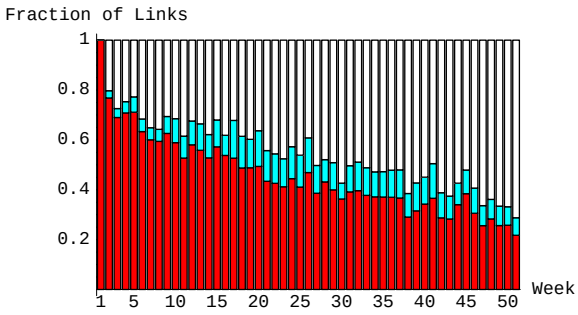

Figure 6: Normalized fraction of links from the first weekly snapshot still existing after n weeks (dark/bottom portion of the bars), new links from existing pages (grey/middle) and new links from new pages (white/top).

Evolution of the Web link structure

The success of Google has demonstrated the usefulness of the Web link structure in measuring the importance of Web pages. Roughly, Google's PageRank algorithm estimates the importance of a page by analyzing how many other pages point to the page. In order to keep up with the changing importance and popularity of Web pages, it is thus important for search engines to capture the Web link structure accurately. In this section we study how much the overall link structure changes over time. For this study, we extracted all the links from every snapshot and measured how many of the links from the first snapshot existed in the subsequent snapshots and how many of them are newly created.

The result of this experiment is shown in Figure 5. The horizontal axis shows the week and the vertical axis shows the number of links in the given week. The height of every bar shows the total number of links in each snapshot relative to the first week. The dark-bottom portion shows the number of first-week links that are still present in the given week. The grey and white portions represent the links that did not exist in the first week: The grey portion corresponds to the new links coming from the "old" pages (the pages that existed in the first week), while the white portion corresponds to the new links coming from the "new" pages (the pages that did not exist in the first week). Figure 6 is the normalized graph where the total number of links in every snapshot is one.

From the figure, we can see that the link structure of the Web is significantly more dynamic than the pages and the content. After one year, only 24% of the initial links are available. On average, we measure that 25% new links are created every week, which is significantly larger than 8% new pages and 5% new content. This result indicates that search engines may need to update link-based ranking metrics (such as PageRank) very often. For example, given the 25% new links every week, a week-old ranking may not reflect the current ranking of the pages very well.

Related Work

Others have studied Web evolution. We are not aware of any prior work on characterizing the evolution of the link structure of the Web experimentally. However, previous studies do touch upon aspects related to our measurements of the birth, modification, and death of individual pages over time. Here we discuss prior studies that exhibit some commonalities with our own.

Recently, Fetterly et al. [12] repeatedly downloaded some 151 million Web pages and measured, among other things, degree of change by counting the number of changed "shingles." The study of [12] spanned a larger collection of pages than ours, but over a shorter period of time (eleven downloads over a period of roughly two months). Aside from those differences, our study differs from [12] in two significant ways. First, by recrawling sites from scratch each week, we were able to

measure rates of web page creation (Fetterly et al. only measured deletion rates), which, interestingly, appear to match deletion rates closely. Second, our study concentrates specifically on aspects relevant to search engine technology, bringing out many implications for the design of search engine crawlers.

Although there has been a rich body of theoretical work on Web growth models, e.g., [9, 10, 13] to the best of our knowledge, our work is the first to study the evolution of Web link structure experimentally.

An earlier large-scale study of the evolutionary properties of the Web was performed by Brewington et al. [7]. That study focused on page modification rates and times, and did not consider link structure evolution.

In [11], lifespans and rates of change of a large number of Web pages were measured in order to assess the viability of adopting an "incremental" strategy for Web crawling. Changes were detected by comparing checksums, and were thus restricted to "all or nothing".

Conclusion and Future Work

We have studied aspects of the evolving Web over a one-year period that are of particular interest from the perspective of search engine design. Many of our findings may pertain to search engine crawlers, which aim to maximize search result quality by making effective use of available resources for incorporating changes. In particular, we found that existing pages are being removed from the Web and replaced by new ones at a very rapid rate. Since some search engines exploit link structure in their ranking algorithms, we also studied the evolution of links the Web. We determined that the link structure is evolving at an even faster rate than the pages themselves, with most links persisting for less than six months.

It is our hope that our findings will pave the way for improvements in search engine technology. Indeed, as future work we plan to study ways to exploit knowledge of document and hyperlink evolution trends in crawlers and ranking modules for next-generation search engines.

References

[1] Google Directory `http://dir.google.com`.

[2] Google Search. `http://www.google.com`.

[3] Nielsen NetRatings for Search Engines. avaiable from searchenginewatch.com at `http://searchenginewatch.com/reports/article.php/2156451`.

[4] Open Directory Project `http://www.dmoz.org`.

[5] The WebArchive Project, UCLA Computer Science, `http://webarchive.cs.ucla.edu`.

[6] Z. Bar-Yossef, A. Berg, S. Chien, J. Fakcharoenphol, and D. Weitz. Approximating aggregate queries about web pages via random walks. In *Proceedings of Twenty-Sixth VLDB Conference, Cairo, Egypt*, 2000.

[7] B. E. Brewington and G. Cybenko. How dynamic is the web? In *Proceedings of the Ninth WWW Conference*, Amsterdam, The Netherlands, 2000.

[8] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the Seventh WWW Conference*, Brisbane, Australia, 1998.

[9] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proceedings of the Ninth WWW Conference*, Amsterdam, The Netherlands, 2000.

[10] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's link structure. *Computer*, 32(8):60–67, 1999.

[11] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *Proceedings of the Twenty-Sixth VLDB Conference*, pages 200–209, Cairo, Egypt, 2000.

[12] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the Twelfth WWW Conference*, Budapest, Hungary, 2003.

[13] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2000.

[14] A. Ntoulas, J. Cho, and C. Olston. What's new on the web? the evolution of the web from a search engine perspective. In *Proceedings of the Thirteenth WWW Conference*, New York, USA, 2004.