What's New on the Web? The Evolution of the Web from a Search Engine Perspective

Alexandros Ntoulas
UCLA Computer Science
ntoulas@cs.ucla.edu

Junghoo Cho
UCLA Computer Science
cho@cs.ucla.edu

Christopher Olston
Carnegie Mellon University
olston@cs.cmu.edu

ABSTRACT

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, the rate of creation of new pages and new distinct content on the Web, and the rate of change of the content of existing pages under search-centric measures of degree of change.

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. For pages that persist over time we found that, perhaps surprisingly, the degree of content shift as measured using TF.IDF cosine distance does not appear to be consistently correlated with the frequency of content updating. Despite this apparent noncorrelation, the rate of content shift of a given page is likely to remain consistent over time. That is, pages that change a great deal in one week will likely change by a similarly large degree in the following week. Conversely, pages that experience little change will continue to experience little change. We conclude the paper with a discussion of the potential implications of our results for the design of effective Web search engines.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia; K.4.m [**Computers and Society**]: Miscellaneous; H.4.m [**Information Systems**]: Miscellaneous

General Terms

Experimentation, Measurement, Algorithms

Keywords

Web evolution, Web characterization, Web pages, rate of change, degree of change, change prediction, link structure evolution, search engines

WWW2004, May 17–20, 2004, New York, New York, USA.
ACM 1-58113-844-X/04/0005.

1. 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 [4], 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.

A number of existing studies have already investigated the evolution of the Web [19, 9, 23, 15, 18]. While some parts of our study have commonalities with the existing work, we believe that the following aspects make our study unique, revealing new and important details of the evolving Web.

- *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 [11]. 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.
- *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, how much new "content" is being introduced and what are the characteristics of the newly-created pages.
- *Search-centric change metric*: We study the changes in the existing Web pages using metrics directly relevant to search engines. Search engines typically use variations of TF.IDF distance metric to evaluate the relevance of a page to a query, and they often use an inverted index to speed up the relevance computation. In our pa-

per, we measure the changes in the existing pages using both 1) the TF.IDF distance metric and 2) the number of new words introduced in each update. The study of the TF.IDF distance will shed light on how much "relevance" change a page goes through over time. The number of new words will tell us what fraction of an inverted index is subject to updating.

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:

1. *What's new on the Web?*
   - We estimate that new pages are created at the rate of 8% per week. Assuming that the current Web has 4 billion pages [2], this result corresponds to 320 million new pages every week, which is roughly 3.8 terabytes in size.[1] We also estimate that only 20% 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.
   - While a large number of new pages are created every week, the new pages seem to "borrow" a significant portion of their content from exiting pages. In our experiments, we observe that about 5% of "new content" is being introduced every week.[2] Given 8% new pages and 5% new content, we estimate that at most $5\%/8\% = 62\%$ of the content in the newly created pages is "new." After a year, about 50% of the content on the Web is new.
   - 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.
2. *How much change?*
   - Our results indicate that once a page is created, the page is likely to go through either minor changes or no change at all. Out of all pages that are still available after one year, half of them do not change at all during that year. Even for the pages that do change, the changes are very minor. For example, after one week, 70% of the changed pages show less than 5% difference from their initial version under the TF.IDF metric. Even after one year, less than 50% of the changed pages show more than 5% difference under the TF.IDF metric. This result is roughly in line with the findings reported in [19] and strongly indicates that creation of new pages is a much more significant source of change on the Web than the changes in the existing pages.
3. *Can we predict future changes?*

   Since search engines have limited network and download resources, they try to download pages that changed most in order to detect as much change as they can. We investigated two ways of predicting how much a page may have changed: the *frequency of change* and the *degree of change*. The frequency of change means how many times a page changed within a particular interval (for example, three changes in a month). The degree of change means how much change a page went through within an interval (for example, 30% difference under the TF.IDF metric in a week).
   - *Frequency of change*: Our result indicates that the frequency of change is not a good "predictor" of the degree of change. We could not observe meaningful correlation between them. For example, even if two pages exhibit a similar frequency of change in the past, say, 10 changes in one week, their future degree of change can be very different. Given this result, we expect that existing refresh algorithms for search engines may not be a good choice if we want to maximize the degree of changes that the search engines detect. Most existing algorithms use the frequency of change as their prediction mechanism [15, 17].
   - *Degree of change*: The past degree of change exhibits a strong correlation with the future degree of change. That is, if a page changed by 30% in the last week (say, under the TF.IDF metric), the page is very likely to change 30% in the next week again. Similar result has been reported by [19], but we also observe that the correlation varies significantly between the sites. While some sites exhibit a very strong correlation some sites do not.

2. 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 [7] 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.

2.1 Selection of the sites

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. [3] 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 [6], and maintains a hierarchical listing of Web sites categorized by topic. Sites

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

[2] More precise definition of "new content" will be given later.

[3] Our dataset constitutes a representative sample of the topical categories on the Web. As we will see later in Section 3.2.1, the trends that we observed from our dataset still hold for a completely random sample of the Web.

Domain	Fraction of pages in domain
.com	41%
.gov	18.7%
.edu	16.5%
.org	15.7%
.net	4.1%
.mil	2.9%
misc	1.1%

Table 1: Distribution of domains in our crawls.

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 [7].

2.2 Download of pages

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[4], 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.

Table 1 reports the fraction of pages included in our study that belong to each high-level domain. The `misc` 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 [8].

3. WHAT'S NEW ON THE WEB?

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 much new content is created? How many new links? We begin by studying the weekly birth rate of pages. For our analysis in Sections 3.1 and 3.2 we treat each unique URL as a distinct unit. Then, in Section 3.3 we measure the shift in the collective content of all pages, to filter out the effect of content duplication among URL's.

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

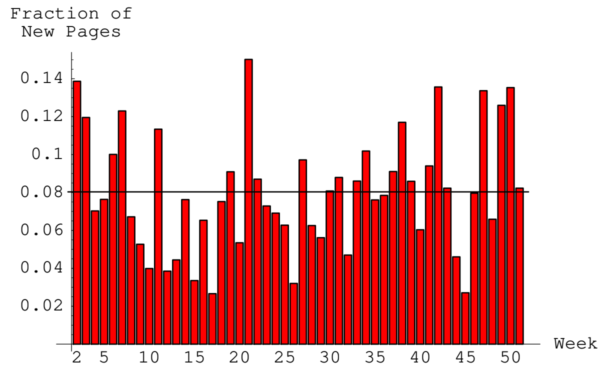

Figure 1: Fraction of new pages between successive snapshots.

3.1 Weekly birth rate of pages

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. (Later in Section 3.3 we measure how much new "content" is introduced every week, which factors out this effect.)

In Figure 1 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 four billion pages [5], we conjecture that there may be around 320 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.

3.2 Birth, death, and replacement

In our next experiment, we study how many new pages are created and and how many disappear over time. [6] We also measure what fraction of pages on our Web sites is replaced

[5]As reported by Google [2].

[6]We assume that a page disappeared if our crawler received an

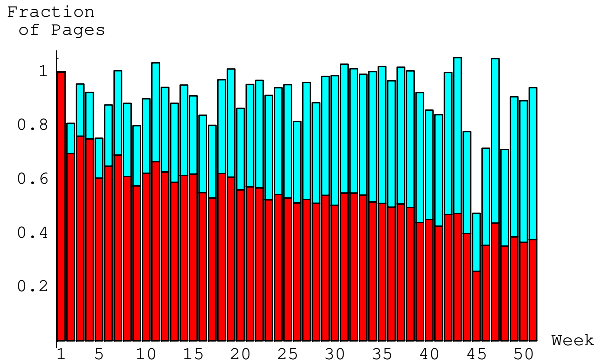

Figure 2: Fraction of pages from the first crawl still existing after n weeks (dark bars) and new pages (light bars).

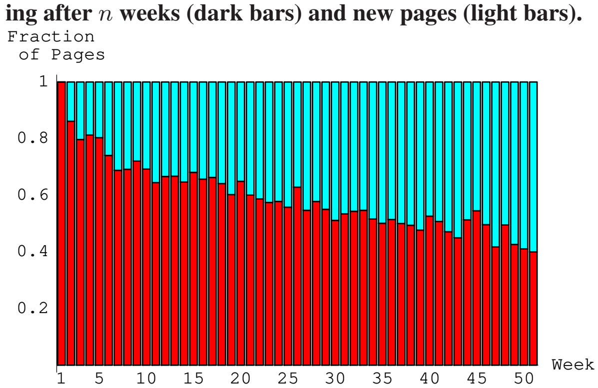

Figure 3: Normalized fraction of pages from the first crawl still existing after n weeks (dark bars) and new pages (light bars).

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

Figure 2 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 2 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 3, 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 3 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.

3.2.1 Generalizing to the entire Web

Our results can be combined with results from recent study by the Online Computer Library Center (OCLC) [5] to get a picture of the rate of change of the entire Web. The OCLC collects an annual sample of the Web and studies various trends pertaining to the nature of Web sites. One of the experiments that the OCLC has conducted over the last few years is to estimate how the number of available Web sites changes over time. From years 1998 to 2002, OCLC has performed systematic polling of IP addresses to estimate the total number of available Web sites. They have also measured what fraction of Web sites are still available after k years.

The result of this OCLC study is shown on Figure 4. In the figure, the horizontal axis represents the year of measurement. The overall height of each bar shows the total number of Web sites available at the given year, relative to the number of sites available in 1998. In 1998 the number of the publicly-accessible Web sites was estimated to be 1.4 million. The dark bottom portion of the bar represents the fraction of the Web sites that existed in 1998 and were still available in the given year. The light portion represents the fraction of new Web sites that became available after 1998. From the graph, we can see that about 50% of Web sites go offline every year. For example, in 1999, half of the 1998 Web sites were still accessible.

Combining this result with ours, we may get an idea of how many pages on the entire Web will still be available after a certain period of time. The OCLC study shows that about 50% of

HTTP 404 response for that particular page, or we could not download the page (due to timeouts etc.) after three attempts.

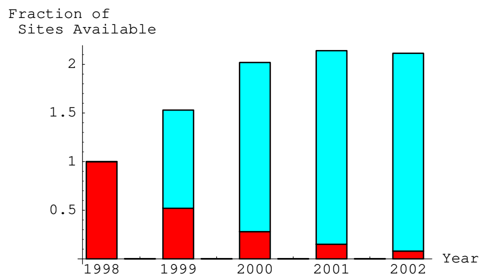

Figure 4: Percent of IP addresses identifying a Web site in Year A also identifying a Web site in Year B. For example, 56% of IP addresses identifying a Web site in the 1998 sample also identified one in 1999 sample. Taken from `http://wcp.oclc.org/`.

existing Web sites remain available after one year. Our study shows that roughly 40% of the pages in each Web site remain available after one year. Therefore, we can speculate that only about $50\% \times 40\% = 20\%$ of today's Web pages will be accessible after one year.

Given this low rate of "survival" of Web pages, historical archiving as performed by, e.g., the Internet Archive [3], is of critical importance for enabling long-term access to historical Web content. A significant fraction of pages accessible today are unlikely to be available after one year. Another implication of our finding applies to standard search engines that do not focus on access to historical content. Since search engine users tend to be very intolerant of broken links in search results, it is very important for search engines to keep abreast of page deletions and omit deleted pages from search results (or, alternatively, point to "cached" copies as Google [2] and other search engines sometimes do, which effectively extends the lifetime of pages).

Finally, from Figure 4, we observe that growth in the number of the Web sites has slowed significantly in recent years. While the number of available Web sites increased by 50% between year 1998 and 1999, the total number of available sites did not change much since year 2000. Given that the number of pages within popular sites does not appear to grow significantly over time (our finding discussed above), there are two remaining potential sources of growth on the Web.

First, it may be the case that relatively unpopular sites are growing. Although our study does not focus on the behavior of unpopular sites, as a small side project we did measure growth for a small random sample of 100 sites on the Web over a period of two months. Our findings for those sites matched those for popular sites: the overall number of pages remained nearly constant over time. Further measurements over a larger scale are needed to verify this preliminary and as yet inconclusive finding.

The second potential source of growth on the Web may stem from increase in the size of pages. While the total number of pages may be leveling off, perhaps it is the case that pages are growing larger over time. In Figure 5 we plot the average page size of each of our snapshots. The horizontal axis plots the week and the vertical axis shows the average page size in a given week, normalized so that average size in the first week equals one. While we see wide fluctuations, a clear upward trend exists in the graph, although it has very mild slope. Given these results, we suspect that the current growth of the Web is mainly driven by the increase in the size of pages over time.

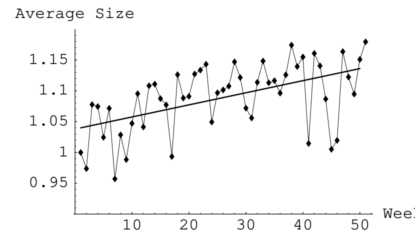

Figure 5: Average page sizes in our snapshots over time.

3.3 The creation of new content

While we measured the rates of creation and deletion of Web pages in previous sections, we did not address how much "new content" is being introduced over time. That is, even when a new page is created, the page may be a mere copy of an existing page in which case it does not contribute any new content to the Web. To quantify the amount of new content being introduced, we use the shingling technique as described in [10, 13]. From every page we exclude the HTML markup and we view the page as an ordered sequence of lower-cased words. A w-shingle is a contiguous ordered subsequence of w words. That is, we group w adjacent words of the page to form a w-shingle, possibly wrapping at the end of the page, so that all words in the page start a shingle.

To measure how much content is created we computed the shingles for all pages included in our study and compared how many new unique shingles are introduced over time. We wanted to answer the following questions: Out of all unique shingles that existed in the first week, how many of them still exist in the nth week? How many unique shingles in the nth week did not exist in the first week? By measuring the number of unique existing and newly appearing shingles, we can learn how much "new content" is being introduced every week. For the experiments presented in this section we used a shingle size of $w = 50$, which roughly corresponds to the number of words in a typical paragraph.

The result of our shingle measurements is shown in Figure 6. The horizontal axis plots time in weeks and the vertical axis shows the total number of unique shingles present each week, relative to the first week. The first week has approximately 4.3 billion unique shingles. The darkly colored, lower portion of each bar shows the number of first-week shingles available in the nth week. The lightly colored, upper portion shows the number of new shingles that did not exist in the first week. To factor out fluctuation in the total number of shingles and focus on the trends in relative terms, we show a normalized version of the graph in Figure 7, where the total number of unique shingles in each weekly crawl is normalized to one.

By comparing Figure 7 with Figure 3, we can see that new shingles are created at a slower rate than new pages. It takes nine months for 50% of the pages to be replaced with new ones, but even after nearly one year, more than 50% of the shingles are still available. On average, each week around 5% of the unique shingles were new, i.e., not present in any previous week. It is interesting to contrast this figure with our finding from Section 3.1 that, on average, roughly 8% of pages each week were new (when identified solely based on

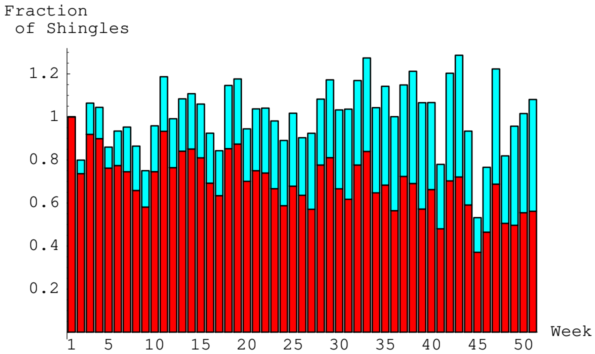

Figure 6: Fraction of shingles from the first crawl still existing after n weeks (dark portion of bars) and shingles newly created (light portion).

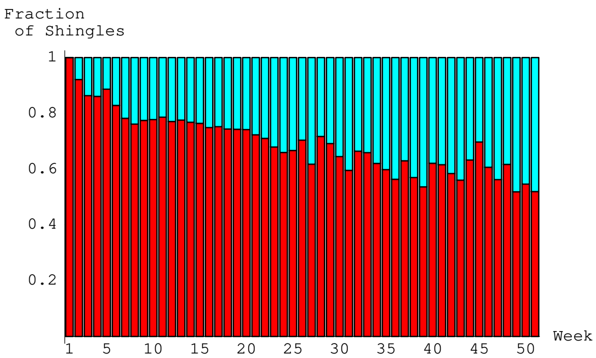

Figure 7: Normalized fraction of shingles from the first crawl still existing after n weeks (dark portion of bars) and shingles newly created (light portion).

URLs). By combining the two results, we determine that at most $5\%/8\% = 62\%$ of the content of new URLs introduced each week is actually new.

3.4 Link-structure evolution

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 8. 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

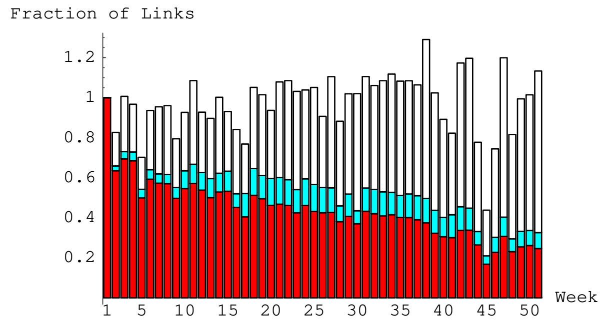

Figure 8: Fraction of links from the first weekly snapshot still existing after n weeks (dark/bottom portion of the bars), new links from existing pages (grey/middle) and new links from new pages (white/top).

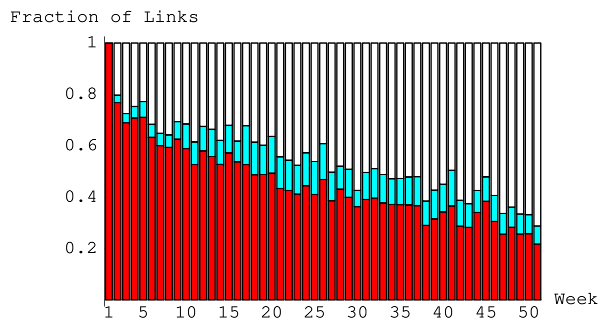

Figure 9: Normalized fraction of links from the first weekly snapshot still existing after n weeks (dark/bottom portion of the bars), new links from existing pages (grey/middle) and new links from new pages (white/top).

coming from the "new" pages (the pages that did not exist in the first week). Figure 9 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.

4. CHANGES IN THE EXISTING PAGES

The previous experiment demonstrated that every week numerous pages disappear from our weekly snapshots and another set of pages is created. The pages that appear repeatedly in our weekly snapshots, however, do not all remain static. In this section we study the way in which the content of pages captured repeatedly by our weekly snapshots changes over time.

4.1 Change frequency distribution

In our first experiment, we investigate how often Web pages change on average. We begin by using the simplest definition

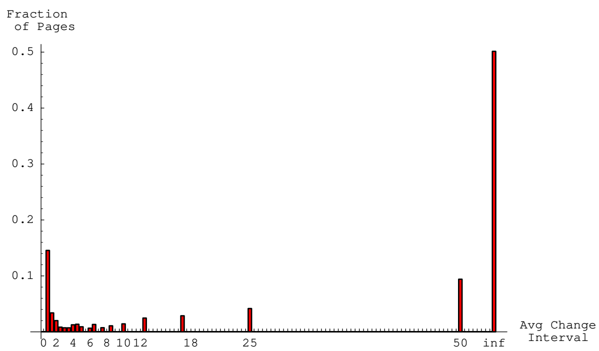

Figure 10: Distribution of the average change intervals of the pages.

of a change: we consider *any* alteration to a page as constituting a change. Later, we will consider a more refined notion of change (Section 4.2).

For this experiment we conducted a scan of our weekly snapshots to determine, for each page that appeared in all snapshots, the average interval between successive changes. For example if a particular page changed twice during our 51-week measurement period, its average change interval is $51/2 = 25.5$ weeks. We then grouped pages by change interval and obtained the distribution shown in Figure 10. Average change interval is plotted on the horizontal axis. The vertical axis shows the fraction of pages having each average change interval. Pages that did not change at all during our 51-week measurement period are counted in the bar on the far right, marked "inf." The large gaps between bars toward the right side of Figure 10 correspond to average change intervals that cannot arise in a 51-week experiment.

From Figure 10 we observe that a significant fraction of pages (around 50%) that occurred in each weekly snapshot remained unchanged throughout the course of our study. Another quite large portion of pages changed very often: approximately 15% of pages underwent at least one change between each weekly download. These two extremes account for more than 65% of the pages. The remaining pages occurring in all snapshots have average change intervals ranging across the spectrum in a roughly U-shaped pattern, with most pages concentrated near one of the two extremes. The tendency is that most pages either change very frequently or very infrequently.

4.2 Degree of change

In our previous experiment we used a very simple definition of change that only captures the *presence* of change, ignoring whether changes are major or minor ones. To delve deeper into the nature of changes undergone by Web pages over time, we now report on experiments designed to measure *degree* of change.

From the point of view of a Web search engine, degree of change is as important as, if not more important than, presence of change. Due to the immense scale and highly dynamic nature of the Web, search engines are faced with a constrained optimization problem: maximize the accuracy of the local search repository and index, given a constrained amount of resources available for (re)downloading pages from the Web and incorporating them into the search index. Search engines that ignore degree of change may waste precious resources downloading pages that have changed in only trivial ways and have little impact on the quality of the search service. Effective search engine crawlers ignore insignificant changes and devote resources to incorporating important changes instead.

Our goal with these experiments was to get a handle on how degree of change may influence the design of highly effective search engine crawlers. Hence, we measured the distribution of degree of change using two metrics that are of relevance to typical search engines:

1. **TF.IDF Cosine Distance** Given two versions of a page p, say p_1 and p_2, we calculate the TF.IDF cosine distance [25] between p_1 and p_2. More precisely, supposing v_1 and v_2 are the TF.IDF weighted vector representations of p_1 and p_2 (excluding any HTML markup), we compute cosine distance as follows:

$$D_{cos}(p_1, p_2) = 1 - \frac{v_1 \cdot v_2}{||v_1||_2 ||v_2||_2}$$

   where $v_1 \cdot v_2$ is the inner product of v_1, v_2 and $||v_i||_2$ is the second norm, or length, of vector v_i.

2. **Word Distance** Given two versions of a page p, p_1 and p_2, we measure how many words of text in p's content have changed (we exclude any HTML markup). The word distance between p_1 and p_2 is defined as:

$$D_{word}(p_1, p_2) = 1 - \frac{2 \cdot |\text{common words}|}{|\text{words in } p_1| + |\text{words in } p_2|}$$

Note that both degree of change metrics are normalized so that all values are between zero and one, with 0 corresponding to no change and 1 indicating that the two versions differ completely.

The TF.IDF cosine distance metric (in its various forms) is the most commonly used method of determining relevance of documents to search queries based on content. Search engines typically rank search results using a combination of cosine distance and other factors (including link-based importance measures as discussed in Section 3.4). A small cosine distance change for a page generally translates to a relatively minor effect on result ranking for most search queries.

Word distance is also important from the perspective of search engine design. Word distance reflects the amount of work required to bring the search index up to date, assuming modifications are made incrementally to allow immediate searchability as in [21]. Both metrics ignore the order in which terms appear on a page, i.e. they treat pages as "bags of words." Doing so is consistent with the way in which typical search engines treat documents (with the exception of phrase matching). In contrast, the shingles metric (which we used in Section 3.3; it is also used by [19]) is highly sensitive to the exact order of terms.

The distribution of TF.IDF cosine distance for all changes is shown in Figure 11. To ensure proper comparability across multiple weeks whose snapshots contained different numbers of pages, we selected a representative week (week 21) from which to obtain IDF weights to use in all of our TF.IDF calculations. The horizontal axis of Figure 11 shows cosine distance and the vertical axis shows the fraction of changes corresponding to the given distance. The dark bars show the distribution of cosine distances; the light bars give the cumulative distribution.

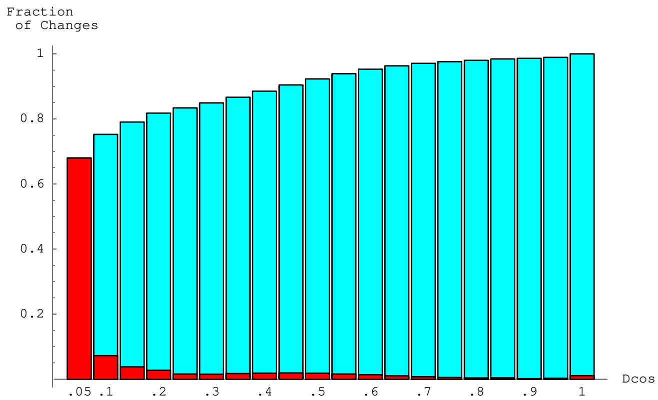

Figure 11: Distribution of cosine distance for all changes. Each dark bar corresponds to changes with cosine distance between the respective x-axis value and the previous one. For example, bin 0.1 corresponds to changes with cosine distance between 0.05 and 0.1. The light bars show the cumulative distribution.

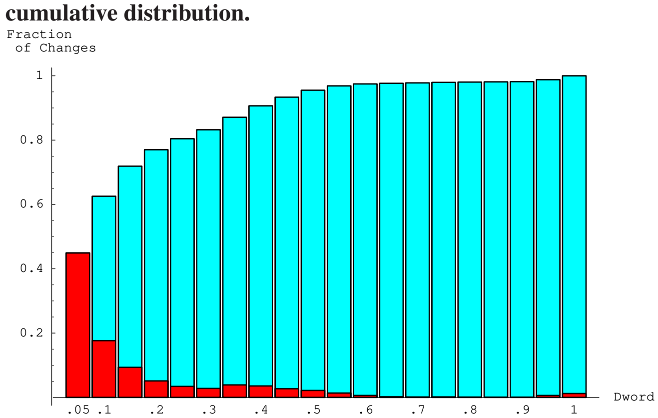

Figure 12: Distribution of word distance values for all changes. Each dark bar corresponds to changes with word distance between the respective x-axis value and the previous one. For example, bin 0.1 corresponds to changes with word distance between 0.05 and 0.1. The light bars show the cumulative distribution.

By examining Figure 11 we can see that most changes are very small, and concentrated on the far left portion of the graph. More than 80% of all changes resulted in a new version whose cosine distance was less than 0.2 from the old version. In fact, 65% of the changes had a cosine distance of less than 0.05. In light of this finding, we conclude that over half of the total changes recorded by our repeated crawls were induced by operations that altered the content of a page very slightly. Such operations might be modifications to advertising material, counters, "last updated" tags, etc. We provide evidence to support this conjecture in Section 4.3.

Our finding coincides with that of [19], which measured degree of change by counting the number of "shingles" affected by a change occurrence. However, we feel that the metric we have selected, TF.IDF cosine distance, may be more directly relevant to search engine crawler design. The observation that most Web page modifications influence cosine distance (and thus ranking accuracy for most queries) very little may have important implications. In particular, in light of this fact it is crucial that search engine crawlers facing resource limitations consider the degree of change, not just the presence of change, when choosing among pages to download and incorporate into the search index. Of course, fine-grained degree of change information can only be leveraged in a traditional search engine context if it is amenable to prediction. We study predictability of degree of change later in Section 5.

We now turn to measurements of word distance. Figure 12 shows the distribution of word distances for all changes detected. We again see that the vast majority of changes are relatively minor ones, although this phenomenon is less pronounced for word distances than it is for cosine distances. The difference between Figures 11 and 12 indicates that a moderate fraction of changes induce a nontrivial word distance (such as 0.2) while having almost no impact on cosine distance. These are changes that primarily affect words with high document frequency that are typically given low weight in search result ranking functions. Incorporating such changes into a search index incurs moderate overhead while adding relatively little benefit in terms of search result quality. This phenomenon and its implications for search engine design merit further study.

4.3 Degree and frequency of change

Our findings in Section 4.2 indicate that most changes are relatively minor ones, so it is important for search engine crawlers to take degree of change into account. We now investigate whether there is any correlation between frequency of change and degree of change. If so, then perhaps degree of change can be estimated indirectly by measuring frequency of change. For example, perhaps it is the case that pages that are modified very often (say, once every day) usually experience only a minor degree of change with each modification (e.g., swapping advertisements). Conversely, perhaps pages that are only modified occasionally (say, twice per year) undergo a large-scale overhaul with each modification.

In our next experiment we aimed to determine whether such a correlation exists. To check for correlation we grouped pages based on their average frequency of change (based on the "all-or-nothing" notion of change) and computed the average degree of change for changes in each group. Degree of change was measured using both the TF.IDF cosine distance and word distance metrics described in Section 4.2.

The result is shown in Figure 13. The horizontal axis represents the number of times (1 to 50) a page changed over the course of our 51 downloads. The vertical axis shows the average degree of change, for each change undergone by pages in each category. The line for D_{cos} corresponds to the cosine distance metric; the line for D_{word} corresponds to word distance. Under both metrics, the highest average degree of change per change occurrence is experienced by pages that either change very frequently (far right of the graph) or change very rarely (left side of the graph). This fact implies that the content of the pages that change very frequently (at least once per week) is significantly altered with each change. The same is true for the pages that change infrequently. Otherwise, no discernible trend is apparent.

To study the relationship between degree and frequency of change in more depth, we also measured *cumulative* degree of change grouped by change frequency: Figure 14 plots overall degree of change between the first and last version of each page, averaged across all pages within each change frequency

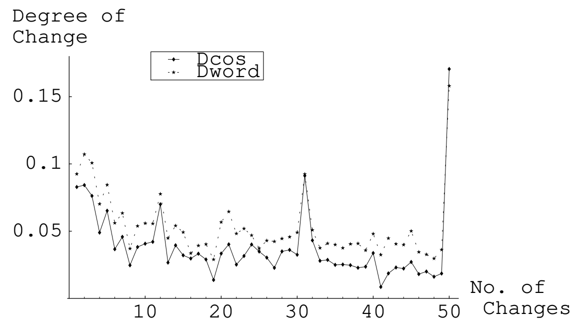

Figure 13: Relationship between degree of change and frequency of change.

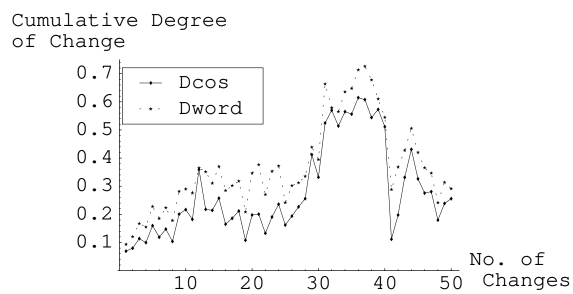

Figure 14: Relationship between *cumulative* degree of change and frequency of change.

group (horizontal axis).

By comparing Figures 13 and 14 we can see that for pages that changed at least weekly (i.e. 50 times during our 51-week measurement period), although each week a roughly 17% degree of change was measured on average (Figure 13), the cumulative degree of change after 50 weeks was only around 30% under both metrics (Figure 14). This finding suggests that the vast majority of modifications to these frequently updated pages tend to occur in the same portion(s) of the page. We inspected a small sample of such pages by hand, and found that in many cases repeated modification of a restricted portion of the content corresponds to aspects such as: weather, stock market, "news of the day" reports, counters, advertisement material containing small text strings (recall that our measurements factor out html tags, images, etc.), and "last updated on..." snippets (these are often generated automatically and in some cases do not coincide with any actual content modification). For the purposes of most search engines, changes such as these can safely be ignored.

In stark contrast, pages that underwent modifications between 30 and 40 times during our 51-week measurement period tended to exhibit a significant cumulative degree of change (above 50% on average), even though on average each modification only incurred a $5-10\%$ degree of change. This discrepancy implies that although these pages tend to change only moderately with each modification, successive modifications often target different portion(s) of the page. As a result, the cumulative degree of change increases substantially over time. Pages in this category that are prone to experience the most substantive and durable alterations in content, even though the frequency of change is not the highest. From a search engine perspective, these moderately frequent changes are likely to be worthwhile to capture, whereas many of the very high frequency changes may not be, as suggested above.[7] The question of how well the two classes of changes can be differentiated based solely on frequency of change statistics remains open.

5. PREDICTABILITY OF DEGREE OF CHANGE

As we concluded in Section 4, most of the changes detected in our experiments were very minor. Search engines may be able to exploit this fact by only redownloading pages that have undergone significant revision since the last download. When resources are scarce it is important to avoid wasting resources on pages whose minor changes yield negligible benefit when incorporated into the search index. However, given the pull-oriented nature of the Web, the capability to differentiate between minor and major changes hinges on the ability to predict degree of change successfully. In this section we study the *predictability* of degree of change in Web pages. In particular, we seek to determine whether past degree of change is a good indicator of future degree of change, in terms of TF.IDF cosine distance. Our results obtained for the word distance metric were very similar so we omit them.

5.1 Overall predictability

We begin our analysis of predictability by studying the overall trends across all pages collected by our crawler. Later, in Section 5.2, we will extend our analysis to a finer granularity by inspecting individual sites. Figure 15(a) shows three scatter plots, each plotting cosine distance measured over a particular interval of time (one week, one month, and three months) on the horizontal axes. The vertical axes plot cosine distance measured over the ensuing time interval of the same duration. Each page contributes exactly one dot to each plot (although the vast majority are masked due to occlusion). These plots enable us to gauge the degree of correlation between successive time intervals. Points aligned along the diagonal (i.e. ones that satisfy the equation $y = x$) exhibit the same degree of change in successive measurement periods, and hence highly predictable degree of change. Therefore, our dataset manifests high predictability if most dots lie close to the diagonal.

To help us judge degree of predictability, we rank pages according to its straight-line distance from the diagonal ($y = x$) and divide them into four groups:

- **Group A:** The top 80% of pages in terms of proximity to the diagonal.
- **Group B:** Pages that fall between the top 80% and the top 90% in terms of proximity to the diagonal.
- **Group C:** Pages that fall between the top 90% and the top 95% in terms of proximity to the diagonal.

[7]Note that [16] suggests that search engines optimizing for overall freshness should, when resources are scarce, ignore high-frequency modifications so that resources can be used more profitably, even when all modifications are assumed to incur the same degree of change. Here, we are pointing out that many high-frequency modifications may be of little interest to search engines *intrinsically*, not just because resources can be saved by not incorporating them.

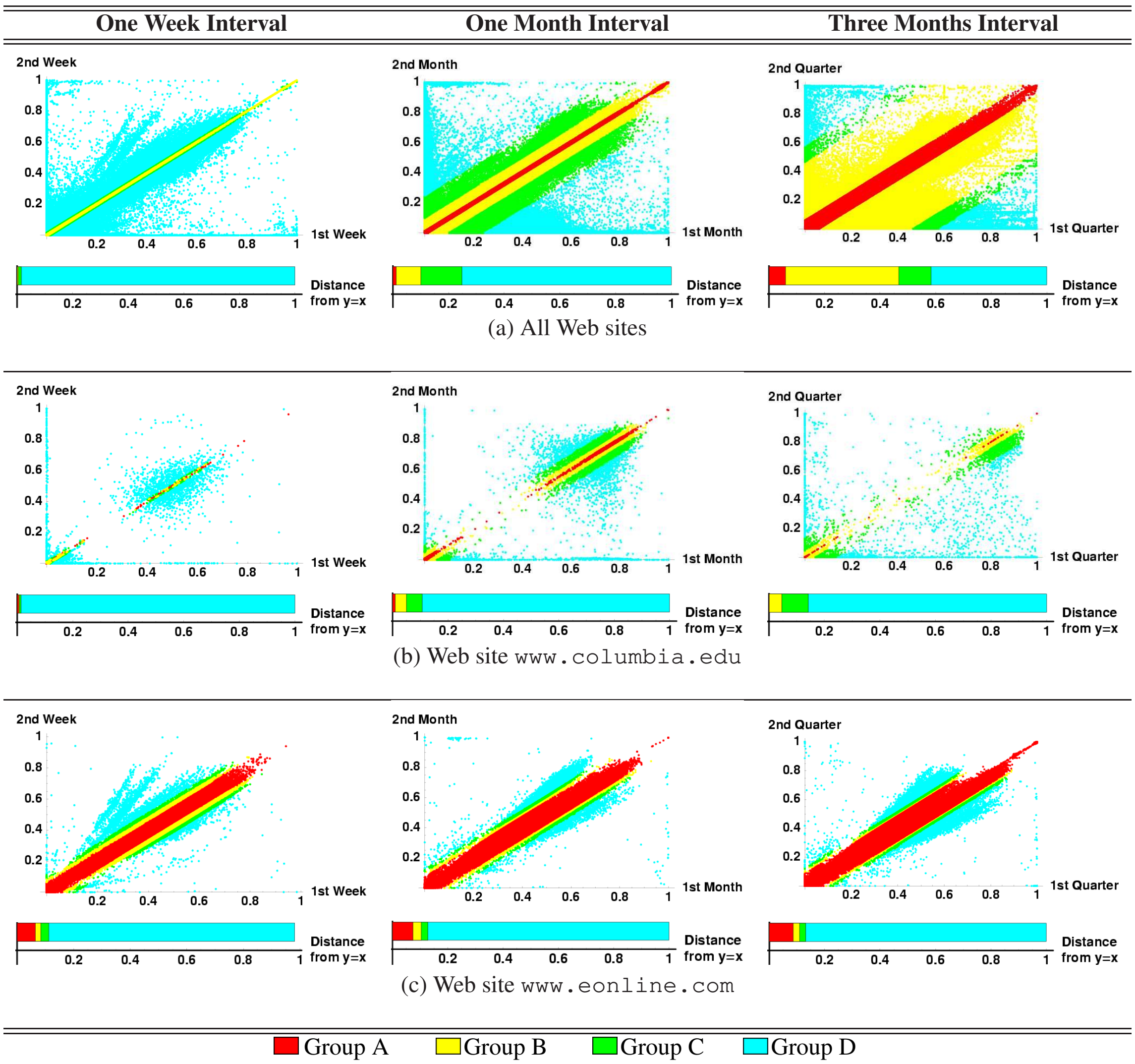

Figure 15: (a) Cosine distance predictability over time for all pages, (b) just www.columbia.edu, and (c) just www.eonline.com. The first column represents changes falling one week apart, the second column shows changes one month apart, and the last column shows changes three months apart.

- **Group D:** All remaining pages.

We plot each group using a different color in Figure 15: red for Group A, yellow for Group B, green for Group C, and blue for Group D. For example, from the red dots in left-most graph of Figure 15(c), we see that Group A (top 80% of pages) from `www.eonline.com` lie in a band between $y = x + 0.06$ and $y = x - 0.06$ from the diagonal. The narrower this band is, the more predictable the degree of change is. To make the "width" of each band easier to see, we plot the distance of members of each group from the diagonal (shown below each scatter plot; distance is normalized to the range [0, 1]).

It is clear from Figure 15(a) that most pages captured in our study change in a highly predictable manner, in terms of cosine distance. For example, from the yellow band (Group B, top 90%) in the second graph of Figure 15(a), we can see that for 90% of the pages, we can predict their future degree of change with +/-8% error; yellow dots lie in the band $y = x$ +/-0.08. Not surprisingly, the degree of predictability decreases somewhat as we move to longer time intervals. The widths of the bands grow larger as the interval becomes longer. The fact that degree of change appears to be highly predictable, especially over the short term, is good news for search engines. By relying on simple predictive methods to estimate degree of change accurately, search engines may be able to take degree of change into account when selecting pages to redownload and reincorporate into the search index. As suggested by our results in Section 4, doing so may enable search engines to use resources more effectively and ultimately achieve higher quality search results.

There are some caveats, however. First, as can be seen in Figure 15(a), the ability to predict degree of change accurately degrades mildly over time: the distance of every group from the diagonal grows over time. Second, a small but non-negligible fraction of pages defy even short-term prediction.

Third, our conclusions are only valid over the pages considered in our study, which are drawn from popular Web sites. Further investigation will be necessary to determine whether our conclusions extend to less popular sites.

Fetterly et al. [19] drew similar conclusions as to the predictability of degree of change, using number of changed shingles to measure change. They only studied short-term predictability over two successive week-long periods, however. Our results in Figure 15(a) show that even over a fairly lengthy span of time (successive three-month quarters), 80% of the pages are within a radius of 5% of the diagonal (recall that the diagonal represents exact predictability for cosine distance). It may not be feasible for certain search engines to monitor all pages of concern on a weekly basis, and the ability to rely on long-term predictability may help considerably with download scheduling and resource management.

5.2 Predictability for individual sites

Having shown that degree of change tends to be highly predictable for most pages included in our study, we turn to an investigation of predictability at the granularity of individual Web sites. After examining our data, we selected two sites that are representative of the range of site-level predictability present in our data: `www.columbia.edu` (an educational site) and `www.eonline.com` (an entertainment magazine site). Scatter plots of cosine distance predictability for the two sites are shown in Figures 15(b) and (c). (For brevity we omit similar plots obtained for other sites.) Both sites exhibit good predictability overall, but the degree of predictability of pages from `www.columbia.edu` is significantly better than that of pages from `www.eonline.com`. Moreover, in the short term, pages from `www.eonline.com` tend to change much less predictably than the majority of pages in our overall study (Figure 15(a)). The degree of change of many pages on `www.eonline.com` either accelerated or decelerated between the first and second weeks, as is especially apparent for group D in the graphs. Perhaps this characteristic, which represents an outlier among the sites we studied, can be attributed to the fast-paced nature of trends and hot topics in the entertainment world.

From these examples we conclude that the ability to predict future degree of change from past behavior can vary a fair amount from site to site. (We confirmed that there is moderate variation in the general degree of predictability across other sites not discussed here due to space constraints.) Therefore, search engines may want to avoid heavy reliance on prediction for certain sites, and indeed for certain "rogue" pages that defy prediction even when other pages on the same site exhibit highly predictable behavior. Establishing reliable methods for identifying Web sites and individual pages for which prediction of degree of change is not likely to succeed (i.e., predicting predictability) is an important topic for future work.

6. 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.

In the first related study we are aware of to focus on degree of change, Lim et al. [22] measured edit distance between two successive versions of around 6000 Web pages. More recently, Fetterly et al. [19] 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 [19] 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 [19] 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.

In particular, when measuring degree of change we focused on TF.IDF weighted cosine distance, which typically forms the basis for search engine ranking functions. Some of our results mirror those of [19, 22] under a complementary distance metric, strengthening our shared conclusions. In addition, our work probes the following issues impacting search engine crawler design: We measured the correlation (or lack thereof) between frequency and degree of change. Furthermore, we studied the predictability of degree of change at a fine granularity, which turns out to vary significantly across domains. Finally, we measured the evolution of the hyperlink structure of the Web, which is a vital concern for modern search engines that combine link-based importance measures with traditional relevance scoring in their ranking functions [11]. Although there has been a rich body of theoretical work on Web growth models, e.g., [12, 14, 20] 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. [9]. That study focused on page modification rates and times, and did not consider link structure evolution. A Boolean, "all or nothing" notion of page modification was used, in contrast to our study which measured degree of change in continuous domains. Using statistical modeling, [9] estimated the growth rate of the Web and determined the growth in the number of pages to be exponential, under the assumption of exponential growth in the number of Web hosts. A white paper from Cyveillance, Inc. [23] published in the same year also reported superlinear growth in the number of pages on the Web (although [23] does not reveal the methods used). These results are in opposition with our finding, which is based on analysis of our data in conjunction with new evidence of a stagnation in the growth rate of the number of Web hosts in recent years [5].

In [15], 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 was thus restricted to "all or nothing." Degree of change was not measured. Earlier, Douglis et al. [18] also studied Web page modification rates, gathering statistics useful from the point of view of designing an effective Web caching proxy. As a result of their focus on caching, the measurements of [18] centered around the extent of replication of content across multiple pages and interactions between access patterns and modification patterns. Again, degree of change was not measured. Furthermore, neither [15] nor [18] measured page creation rates or link structure evolution.

Pitkow et al. [24] studied statistical characteristics of a single Web site for which they had access to user activity logs in addition to content snapshots. They characterized pages in terms of user access patterns and co-citation patterns. The cocitation analysis was performed over a static snapshot of the site—analysis of link structure evolution was not undertaken. However, the evolution of individual pages in the site was studied, and correlations were found between frequency of modification and page lifetime, and between source of access (internal versus external) and page lifetime. Creation rates of new pages and degree of change were not measured in [24].

7. 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. However, new pages tend to "borrow" their content heavily from existing pages. The minority of pages that do persist over extended periods of time typically exhibit very little substantive change (although many undergo superficial changes). For the exceptional pages that change significantly over their lifetimes, the degree of change tends to be highly predictable based on past degree of change. However, past *frequency* of change does not appear to be a good all-around predictor of degree of change.

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.

8. REFERENCES

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

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

[3] The Internet Archive `http://www.archive.org`.

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

[5] Online Computer Library Center `http://wcp.oclc.org`.

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

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

[8] 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.

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

[10] S. Brin, J. Davis, and H. García-Molina. Copy detection mechanisms for digital documents. In *Proceedings of the ACM SIGMOD Annual Conference*, 1995.

[11] 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.

[12] 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 Nineth WWW Conference*, Amsterdam, Netherlands, 2000.

[13] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. In *Proceedings of the Sixth WWW Conference*, 1997.

[14] 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.

[15] 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.

[16] J. Cho and H. Garcia-Molina. Synchronizing a database to improve freshness. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 117–128, Dallas, Texas, 2000.

[17] E. Coffman, Jr., Z. Liu, and R. R. Weber. Optimal robot scheduling for web search engines. *Journal of Scheduling*, 1(1):15–29, June 1998.

[18] F. Douglis, A. Feldmann, and B. Krishnamurthy. Rate of change and other metrics: a live study of the world wide web. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, Monterey, 1997.

[19] 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.

[20] 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.

[21] L. Lim, M. Wang, S. Padmanabhan, J. S. Vitter, and R. Agarwal. Dynamic maintenance of web indexes using landmarks. In *Proceedings of the Twelfth WWW Conference*, Budapest, Hungary, 2003.

[22] L. Lim, M. Wang, S. Padmanabhan, J. S. Vitter, and R. C. Agarwal. Characterizing web document change. In *Proceedings of the Second International Conference on Advances in Web-Age Information Management*, pages 133–144. Springer-Verlag, 2001.

[23] B. H. Murray and A. Moore. Sizing the internet. White paper, Cyveillance, Inc., 2000.

[24] J. Pitkow and P. Pirolli. Life, death, and lawfulness on the electronic frontier. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, Atlanta, Georgia, 1997.

[25] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, first edition, 1983.