URL-Enhanced Adaptive Page-Refresh Models

Robert Warren, Dana Wilkinson, Alejandro López-Ortiz
School of Computer Science, University of Waterloo
{rhwarren, d3wilkin, alopez-o}@uwaterloo.ca

Technical Report CS-2005-16
May 2, 2005

University of Waterloo

Abstract

We study the refresh model required to keep an up to date copy of a web page. This has applications for more efficient and frequent crawling, in the case of search engines, and for higher hit rates for proxy servers with pre-fetching. Two main models have been proposed namely the uniform refresh model (Cho and Garcia-Molina, 2000) and the adaptive page refresh model (Edwards et al., 2001), with some debate as to the relative value of each model.

In this work we show that adaptive page refresh models can be further improved by careful selection of initial page-refresh rates of newly added links as indicated by our page evolution studies showing that page-change rates (and consequently page-refresh rates) are dependent on top-level domain, category and page depth.

Keywords: crawling, refresh rate, rate of change, web evolution, web pages

1 Introduction

The web continues to grow at a rapid pace. For example, in 1998, Lawrence and Giles estimated the existence of 320 million web pages [24]. They revised this estimate to 800 million in 1999 [25], and again to 1 billion in early 2000 [20]. Currently, the popular Google search engine [18] indexes over 3.3 billion pages, which implies over 18 billion pages if the estimates of Lawrence and Giles that the best search engines only cover about 16% of the indexable web [25] still hold. Setting aside sampling techniques, it is safe to assume that the process of discovering new web pages requires a major logistical effort based simply on the growth rate of the web. Other issues, such as web page mirroring [5] and unreachable portions of the net (Dark-Net [23]), compound the problem.

Additionally, all of these estimates of the number of web pages are for the *publically indexable web* which excludes pages that are not normally considered for indexing by web search engines, such as pages with authorization requirements, dynamically generated pages, and pages hidden behind search forms (i.e. those whose "content" is stored in databases). These pages have been referred to as the *deep web*. Bergman [3] has estimated that there are more than 400 times as many pages in the deep web as there are in the publically indexable web—more than 7,500 terabytes of data. Recently, some search engines have attempted to index some of the deep web as well.

Clearly the size and growth rate of the web constitutes a major hurdle in both the creation and maintenance of an effective search engine. In this paper we study the rate at which a web page changes and provide evidence that the rate is dependent on certain attributes of the page such as top-level domain, category and page depth. Consequently, these attributes may be useful in choosing when to refresh the pages stored in a search engine's database.

Much work has been done on page discovery (see for example Menczer et al. [26] on link following strategies). There is also a wide variety of research on effective and efficient indexing and re-indexing of pages (see e.g. Page et al. [32] on PageRank and the excellent review by Florescu, Levy and Mendelzon [17]).

Several studies have considered the rate of change of pages on the net (e.g. [12, 16, 6, 36]). Notably, Cho and Garcia-Molina [9] studied the frequency of change of a web page across the entire web as well as per domain. Table 1 summarizes their observations.

Time between updates	overall %	com	net/org	edu	gov
1 day or less	24%	42%	12%	2%	1%
1-7 days	15%	21%	20%	6%	6%
1-4 weeks	15%	14%	22%	16%	20%
1-4 months	16%	12%	13%	24%	22%
4 months or more	30%	14%	34%	50%	53%

Table 1: Rate of change per domain (Cho and Molina 2000)

Douglis et al. [13] considered the rate of change of pages on the web in the context of caching. Additionally, several authors have reported that over 20% of web pages changed whenever visited (see for example Arasu et al. [1]), possibly because they were dynamically generated.

The change process of a web page is generally modelled as a Poisson process. In a different paper, Cho and Garcia-Molina [8] take a look at a number of different mathematical estimators that are all very close to those of the Poisson distribution.

The death rate of web pages was studied by Pitkow and Pirolli [33] who used survival analysis from statistical theory. One of the most interesting conclusions from their research was that the use of the information within an organization was the most likely indicator of the probability of survival. A web page which is used exclusively within an organization is most likely to be deleted, while a resource which is read mostly from outside the organization has the highest survival probability. Web pages which are read equally from within and without the organization have a middle survival rate, but are most likely to linger for an extended amount of time. Finally, an interesting side note is that HTML-type material was most likely to be deleted as opposed to non-HTML which was most likely to be left to linger on the server.

These continuous page-change and page-death processes are what drives the need to refresh the search engine's page collection. Because of the scale of the information involved, the refreshing process becomes an important consideration.

From these measurements two main models for page refreshing were proposed, namely, the uniform page refresh model [9] (which calls for the refresh of all pages at a uniform rate) and the adaptive page refresh model [15] (which refreshes pages at different rates depending on the history of the page in the database).

In the uniform refresh model the collection is traversed as fast as resource constraints allow and is updated without prioritizing—in other words, all pages are essentially refreshed at approximately the same rate. In contrast, in the adaptive refresh model pages are initially crawled at uniform rates. The refresh rate is then updated to reflect the observed frequency of change between periodic crawls.

In this paper we study the rate of change and death in a random sample of 115,471 URLs taken from the Open Directory Project [31]. From the statistics gathered we argue that the Adaptive Refresh Model, can be further refined by choosing different initial refresh rates based on secondary characteristics such as page depth, top level domain, and content classification type (if available). To our knowledge page depth and content type have not been studied before in this context.

In our particular study we derive content type from the Open Directory classification. In practice this could be derived from analysis of the document itself. Interestingly, Fetterly et al. [16] considered document length as a content property which can be used to predict rates of change. They report that documents of length at most 2^{12} bytes change substantially less often than average. This observation then can be used to further refine initial refresh rates in the adaptive model.

1.1 Collection Refreshing Concerns

For the purposes of this paper, the word *collection* is used to refer to the set of web pages currently in use by the search engine for queries. The word *freshness* refers to whether a page stored by the search engine is the same as the one on the web.

Because the page-lifetime process is unknown to the search engine, in order to refresh the page it has no choice but to re-fetch the page and compare it with the indexed version. Steps can be taken to reduce the amount of network chatter using HTTP headers, but this still requires a connection to the web server. Furthermore, Douglis et al. [13] report that fewer than 80% of web servers report a proper Last-Modified time-stamp. 20% of 2 billion documents is still a sizeable 40 million web pages which can only be checked through data transfer.

With this in mind, there are at least three different aspects of collection refreshing that have been looked at in detail—these will now be summarized.

The first is whether to refresh on a batch-mode basis (where the entire collection is updated at a specific time interval as quickly as possible) or on a steady basis (where the collection is updated continually over time, one page at a time). If we measure the performance of both methods over the same time period, both methods would achieve the same average freshness over the collection. However, the steady refresh method would have the advantage of spreading the computational and network load over a long period of time and would not "spike", or overload, any part of the system.

Secondly, the updated collection of web pages may not be immediately made available to the search system until all of the pages have been refreshed. This may be done for internal consistency or index generation reasons. Alternatively, refreshed pages can be replaced directly within the search system as they are retrieved from the network.

Finally, the refreshing of the collection maybe done at a single uniform frequency for all pages within the collection such as in the case of a batch-mode system. Another method is for the search engine to refresh different pages at different frequencies in order to keep active pages as updated as possible or so as not to overload certain web sites. This issue is closely related to the design of the crawl queue (see for example Menczer et al. [26]). In this paper we propose a improvements on the scheduling model for the crawl queue.

2 Uniform vs. Adaptive Page Refreshing

The simplest scheduling methodology is the Uniform Refresh Model where the collection is traversed as fast as resource constraints allow and is updated without prioritizing—in other words, all pages are essentially refreshed at approximately the same rate. The advantage of this model is that it makes a "best-effort" attempt at keeping the collection fresh while still being flexible at the amount of resources being consumed. Hence, priority can be given to time-sensitive activities, such as serving user queries, while crawling the web takes place utilizing unused resources.

An ideal solution would be available if the actual rate of change of a page was known. Individual page-refresh rates could then be determined based on these rates of change. The problem does not become trivial, however, as there is a an important distinction between the rate of change of a page and the point in time at which a page *is* changed. If we refresh the page just before it is updated by its author (thereby refreshing at the same rate as the page changes), the search engine's copy will continually be obsolete! Note that this is not unlike well known Digital Signal Processing (DSP) problems with respect to phase error and sampling rates.

Unfortunately, the actual rate of change of a page cannot generally be known and must somehow be estimated. Edwards et al. [15] present an adaptive refresh model for the scheduling of page refresh that takes into account, among other things, bandwidth and discovery constraints. The rate of page refresh is determined by an estimate of the page-change rate based on how often the page has changed between past crawls. Manually defined limits ensure that a web page is crawled within a set maximum delay.

Some researchers suggest that the Uniform Refresh Model is sufficient in light of scheduling difficulties and resource hogging on the part of ever-changing pages. Cho and Molina [10], for example, use the Poisson process assumption to argue that the Uniform Refresh Model is effectively superior to an Adaptive Refresh Model.

Another possible model utilizes a priority-queueing system so as not to waste bandwidth on unchanging pages and to keep pages which change more often as current as possible. In order to ensure that all pages do eventually get refreshed, the minimum-refresh rate is set to once per collection update cycle (see Risvik and Michelsen [34]). In a different paper, Cho et al. [11] review crawl queueing systems on the basis of links weighting and topic, but do not cover the queueing of pages to be refreshed—in fact, little work seems to have been done in this area.

The refresh scheduling of a search engine's collection is a problem in which much work remains to be done. Often the particular strategies used depend much more on design parameters than on a particular theoretical performance metric. The problem is compounded by the fact that it is difficult to predict the actual change rate of the information on the web and hence some initial adjustment period must be taken into account. Edwards [14] suggests that the approach taken by most engines is some form of adaptive sampling based on categorical information—the oft-quoted rate of 4-6 weeks for re-indexing of an entire collection seems to be the minimally guaranteed rate of refresh for a page, with certain pages being re-indexed more often.

Although it may be the case that certain pages cannot be refreshed at a rate approaching their change rate it may still be beneficial to use the adaptive model for pages which do not change as frequently, hence freeing results to more frequently refresh rapidly changing pages.

3 URL Adaptive Refresh Model

One attribute that can be immediately derived from a page with a DNS entry is the top-level domain (.com, .org, etc.). This is easy enough to determine from the URL using simple string parsing and should be findable for a majority of newly discovered pages. Cho et al. [9] use a similar approach and other empirical evidence supports the hypothesis that page categorization is a viable way to predict page-change rate (see e.g. [16]).

A second attribute that can be gleaned from the URL of a page is the page depth. This can quickly be found by creating a normalized canonical form of the URL, and counting the number of slashes (/) in the page's URL.

Thirdly, it should be possible in many cases to assign a page to a category based either on the content of the page or, in some cases, based in part on how the page was found (e.g. in a directory such as Yahoo, or depending on what pages link to it).

Finally, a useful metric for determining initial refresh rates is the probability of page death. If a page is more likely to die than another then it may rate less attention (and consequently less bandwidth and/or other resources) than a page that is more likely to be around in the future. Consequently for each of the groupings mentioned in section 3 we would like to determine the percentage of dead pages in each group in order to assess the probability that a page with that particular attribute will end up dead.

3.1 Design of Study

A random sample of 115,471 URLs was taken from a list obtained from the Open Directory Project [31] The Open Directory Project is a large directory of the world-wide web compiled by a community of volunteers. In Section 5 we discuss the validity of the sampling technique used. We also provide a rationale as to why it was chosen instead of other known sampling alternatives.

The HTTP headers of each URL were examined to determine whether or not the link was dead. We considered a page dead only if it returned a "404-Not Found" error. We then retrieved the HTML content for each URL from the Internet Archive Wayback Machine and stored them in a database. The Internet Archive is a nonprofit organization that periodically crawls the web and stores copies of the crawled pages. These copies are *not* overwritten with each update, rather the new copy is stored as well resulting in numerous "snapshots" of each URL at various dates [21]. After the Internet Archive had been scanned the database contained 2,126,659 total observations, compared to the original 115,471 URLs.

Each URL was then examined and assigned membership into different groups. Top-level domain was determined by parsing the text after the last period (.). The category of each link was indicated in the file obtained from the Open Directory Project [31]. The link depth was determined by counting the number of forward slashes (/) in the URL. For each page, each copy was compared with the copy that preceded it temporally. Each observation was then denoted as either changed (if it was different from the previous observation) or unchanged (if there was no difference from the previous observation).

The change statistics gathered are under-estimations of actual rates of exchange, as more than one change might have occurred between two consecutive snapshots. As well, the possibility that a change occurred which was later reversed before the a subsequent snapshot was taken cannot be dismissed. In this case, no change would be reported. However we venture to guess that this is not a likely scenario.

Although there were many different top-level domains in our sample the majority of them contained less than a hundred pages. In the histograms that follow, only the more common ones are shown (.com,

.net, .org, .edu, .ca and .us). Similarly, we removed the few (less than a hundred) pages that fell into the “Netscape” category from our histograms. The number of pages with depth greater than five was also very small (again less than a hundred) so they are also not shown.

4 Results and Discussion

4.1 Top-Level Domain Results

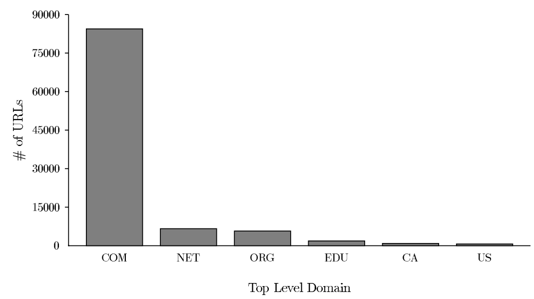

Figure 1: Number of URLs in each top-level domain.

To begin with, it is worth noting that the distribution of the histogram in Figure 1 (which shows the number of URLS from the sample in each top-level domain) matches the power law distribution of data from previous experiments. By far the most sites are in the .com domain. The next largest domains, smaller by an order of magnitude, are .net and .org. Smaller again by an order of magnitude are the .edu, .ca and .us domains. This helps argue the validity of the sample.

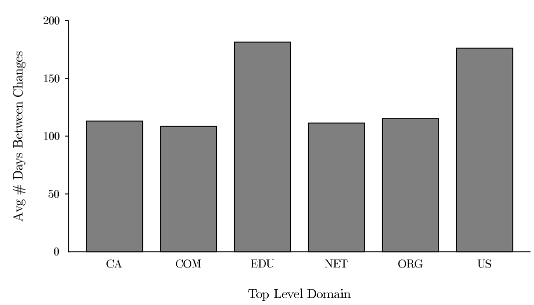

Figure 2: Average time (in days) between changes for each top-level domain.

An immediate and significant difference is noticed (see Figure 2) in the average days between changes for pages in the .edu and .us domains as opposed to those in the remaining domains (.com, .net, .org and .ca). This certainly provides good support for our hypothesis that page-change rate is somehow dependent on top-level domain. More research on why this is the case for these particular domains would be interesting. The observation that the .com pages change faster than the .edu pages matches the previous observations made by Cho and Molina [9]. Table 1, which summarizes the results from their paper, clearly shows a tendency for the .com pages to change at rates measured in days while the .edu pages are more likely to have rates of change measured in months.

Despite the fact that the .com pages demonstrate one of the higher rates of change their death rate is low. Figure 3 shows that the percentage of dead pages in each top-level domain. This observed low percentage of dead pages in the .com domain is actually contrary to the previous reported results of Cho and Molina [9] who predicted that .com pages were more likely to die than pages from other domains. This may be explained by the recent high levels of competition in the market for commercial domain names—even if a page dies, it may not be long before another commercial venture re-opens

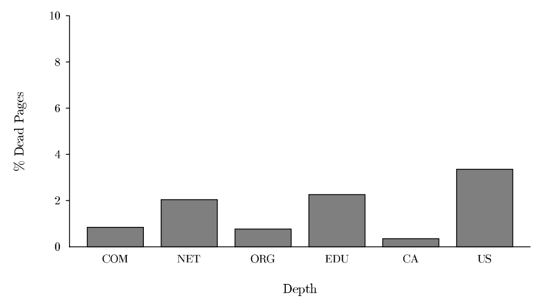

Figure 3: Percentage of dead pages in each top-level domain.

it. If this is the case it leads to speculation that links in certain domains should be periodically crawled even after they have been determined dead as there is greater possibility that they may "return to life" in the future—albeit in a completely different form. As the majority of pages are in the .com domain, more research on this "rebirth" concept could be quite illuminating. Why this phenomena is also observed for the .org and .ca pages in the sample is unclear.

4.2 Category Data

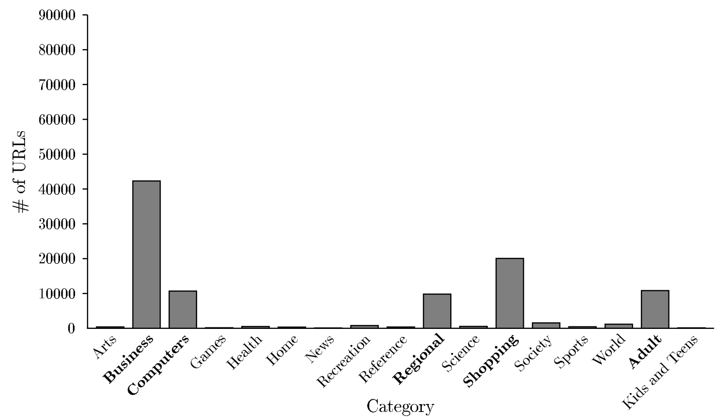

Figure 4: Number of URLs in each category.

Although there are many categories listed in the Open Source Directory, the one containing the most pages by far is the "Business" category (see Figure 4 which shows the number of URLs from the sample in each category). The next most-populated category is the "Shopping" category which is about half the size. Three more categories—"Adult", "Regional" and "Computers"—are again half the size with many less-populated categories abounding. This distribution is remarkably similar to the power law which is so prevalent in Internet statistics. The top five categories just mentioned will be labelled in bold on the histograms summarizing the category data to remind the reader that the other category populations are significantly smaller.

Figure 5 summarizes the average number of days between changes for pages of various categories. Again, as this does not resemble a uniform distribution, we are provided with more evidence that our hypothesis is acceptable. The fact that "News" pages seem to change the quickest should come as no surprise as most news sites change their pages many times daily. We would expect that the average number of days between changes for the "News" category would be much lower than that observed but we are somewhat limited by the under-estimation factor in our survey which we discussed in the previous section and the fact that the Internet Archive must often wait weeks and sometimes months before getting a new "snapshot" of a page. In other words, the finest degree of resolution of our study

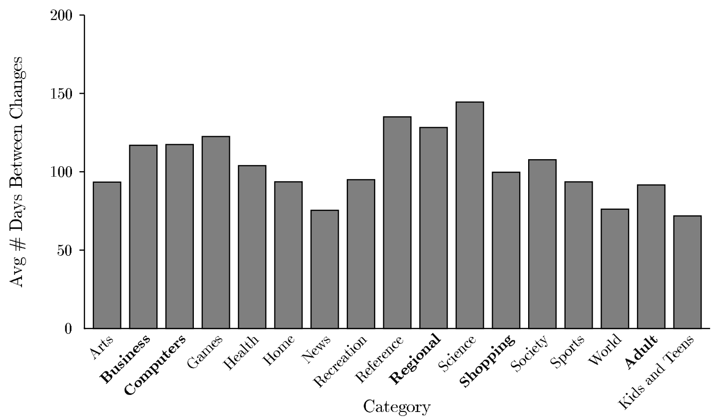

Figure 5: Average time (in days) between changes for each category.

is in the order of weeks.

Not surprisingly pages in the "Reference" category are some of the slowest changing as most reference sites, once created, tend not to change much at all since the information that they are referencing is often static (e.g. dictionaries, survey results, etc.). To varying degrees the remainder of the graph in Figure 5 can be argued as fitting common beliefs about the Internet.

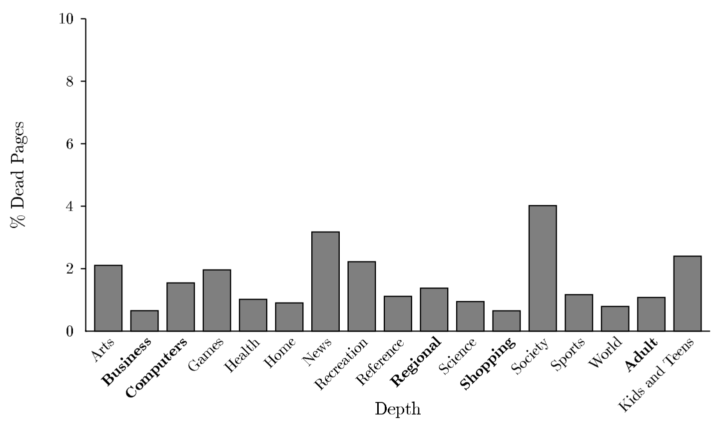

Figure 6: Percentage of dead pages in each category.

Again it is not surprising that "News" pages, the ones that change the most often, are also the most likely to die (see Figure 6 which shows the percentage of dead pages for each category). Intuitively, most news sites must frequently purge articles that are a few weeks old in order to maintain a large enough resource pool for newer articles. On the other hand, "Business" related pages seem to have a relatively low chance of dying. This mirrors the previous observations of the .com domain and, given the natural association of the "Business" category to the commercial sector, may be explained in a similar manner.

4.3 Page Depth Data

Although certainly not as dramatic as the corresponding graphs for the previous groups the graph of average time between changes for pages of varying depths (Figure 7) is arguably not uniform and therefore provides even more support for our hypothesis. It indicates that there is a slight propensity for pages to change less quickly the deeper they are in a site's hierarchy.

The histogram in Figure 8 shows the percentage of dead pages at each page depth. It is interesting

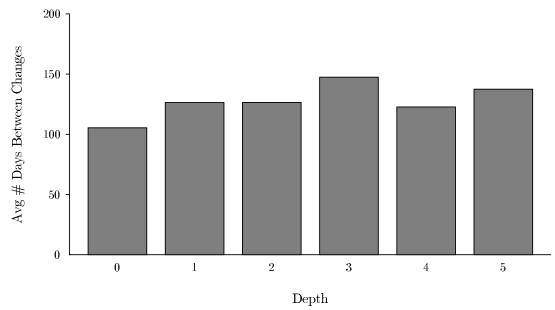

Figure 7: Average time (in days) between changes for each page depth.

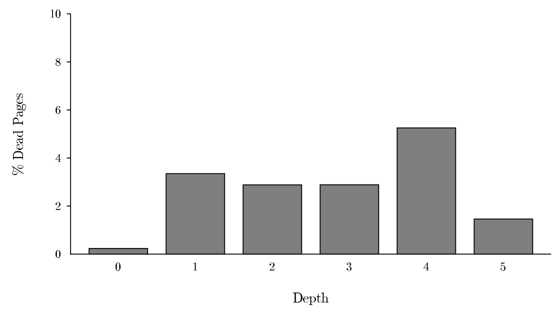

Figure 8: Percentage of dead pages at each page depth.

in that it's distribution seems to indicate that the pages most likely to die are those at *medium* depth (as opposed to surface pages or deeply buried pages). One could make the hypothesis that the deeper the depth of a page at an organization, the more likely it is that that page is used exclusively by the organization. If this correlation holds then the distribution in Figure 8 perfectly matches the results of Pitkow and Pirolli [33] on survival analysis (described in Section 1). It would be worthwhile to test this hypothesis in an attempt to increase what is known about page-death rates.

5 Sampling the Web

Due to the vast number of pages found on the web it is infeasible to look at the entire web in order to determine how the rate of change of various categories of web page differ. We therefore needed to analyze a random sample of web pages and generalize our results to the population of all web pages. As it turns out, getting such a sample is not trivial. Several methods have been used but no standardized method has emerged. Each of the known possibilities is considered below, followed by a justification of our particular choice of sampling technique.

5.1 Random IP Address Testing

Lawrence and Giles [25] proposed randomly choosing an IP address and checking if the ports which typically responded to HTTP requests were active. If so they crawled all the pages accessible from the main page and randomly selected one of them.

This method works exceptionally well for gathering population statistics about the web and for random sampling of servers but it is ill-suited in uniformly sampling pages. Since the number of pages available at a server ranges dramatically (from a few to several million) there are problems in efficiently implementing this method in such a way as to find random pages uniformly.

Additionally, this method suffers as it cannot sample from pages that cannot be crawled to from a server's main page, nor can it sample from servers that may only be accessed by non-standard ports.

This last disadvantage could be overcome through the use of a port scanning but such activity is overly invasive and may be ethically questionable.

5.2 Query-Based Sampling

Query-based sampling (as used by Lawrence and Giles [24] and Bharat and Broder [4]) involves querying a search engine then randomly choosing one of the pages returned by the query. It is quicker than the previous method as it requires no crawling (only interface to a search engine) and doesn't require the finding of an active server. However, the population of possible pages is limited to those that are in the search engines database. Although many search engines claim to index the entire web there are some claims (see for example Murray and Moore [27] or the Online Computer Library Center [25]) that this is far from the case (this disparity in coverage estimates remains an issue to be resolved). Additionally, such samples may contain query bias and ranking bias.

As well, query bias occurs because in order to get the initial list a query must be presented to the search engine. Naturally, any pages returned will be biased towards that query and/or its source. This bias can be towards specific words that appear in the query or more generally towards the language that those words are from. Even if words are randomly selected from different languages, the probability with which a word from a particular language appears will detract from the uniformity of the sample. Even more generally, since a query is initially required and the majority of search engines index on text there will be bias towards pages with more textual content.

Ranking bias occurs because most search engines have a maximum number of web pages which they will return to a particular query. If such a "cut-off" occurs then the sample will be affected by the way in which the search engine ranks pages. If the queries are modified in such a way as to reduce the ranking bias then the query bias will correspondingly be increased.

5.3 Random Walking

Random walking (see for example Henziger et al. [19], Bar-Yossef et al. [2] or Rusmevichientong et al. [35]) involves starting at a web page chosen randomly from a relatively small start set and then following links randomly. With some modification of the weights assigned to each link this method yields a provably uniform distribution over time. This method has been analyzed extensively from a graph-theoretic point of view which is a definitely advantageous. There are some minor problems which arise in the choosing of the start set. Additionally, the method can become computationally intensive as the entire walk must be stored in order to simulate the web as an undirected graph (which it clearly is not). The biggest problem, unfortunately, is that this method only works with a well-connected area of the web and it is hypothesized that only 25% of the web forms such a strongly connected area [7].

5.4 List Sampling

List sampling is actually a broader designation of sampling which involves the creation of a list of possible URLs using a variety of different methods (possibly including those just described). This kind of sampling has been used by a variety of researchers, but with no accepted method of list creation (see for example Arasu et al. [1]). When a random web page is required it is selected randomly from the list.

Although the creation of a good list could take a long time the actual sampling (needing only a simple look-up) is very quick. Additionally, some such lists are publically available (such as the one at the Open Directory Project [31]).

In our opinion, this method is the best method if care is taken to ensure that the list is continually updated in an unbiased a manner as possible. It would be tremendously beneficial if someone were to take up the challenge of creating and maintaining a good list of URLs for the use of the web-research community.

5.5 Conclusions and Future Work

The next step in this line of research should be the creation of a sample database of pages and the refreshing of this collection with a Uniform Refresh Model, an Adaptive Refresh Model with uniform initial refresh rates and an Adaptive Refresh Model with refresh rates predicted as outlined previously. The resulting observation and comparison of each method may provide useful insight into the problem of collection refreshing.

A random sample of web pages was broken down by a variety of different attributes—top-level domain, category and page depth. For each attribute, the most important histograms from the experiment are those which delineate the average time between page changes. These times provide reasonable approximations of page-change frequency. Since none of the corresponding distributions are uniform, they consequently constitute evidence that page-change frequency is in each case dependent on the observed attribute The information thus gained can be used to fine tune the initial refresh rate of the adaptive page refresh scheduler.

References

[1] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. *ACM Transactions on Internet Technology*, 1(1):2–43, august 2001.

[2] Z. BarYossef, A. Berg, S. Chien, J. Fakcharoenphol, and D. Weitz. Approximating aggregate queries about web pages via random walks. In *Proceedings of the 26th International Conference on Very Large Data Bases*, 2000.

[3] M. K. Bergman. The deep web: Surfacing hidden value. white paper, Bright Planet, 2000.

[4] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. In *Proceedings of 7th International WWW Conference*, 1998.

[5] K. Bharat and A. Broder. Mirror, mirror, on the web: A study of host pairs with replicated content. In *Proceedings of the Eight International World-Wide Web Conference*, May 1999.

[6] B. E. Brewington and G. Cybenko. Keeping up with the changing Web. *Computer*, 33(5):52–58, 2000.

[7] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. tomkins, and J. Wiener. Graph structure in the web: experiments and models. In *Proceedings of 9th International WWW Conference*, 2000.

[8] J. Cho and H. Garcia-Molina. Estimating frequency of change. Technical report, Stanford University, 2000.

[9] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *Proceedings of the Twenty-sixth International Conference on Very Large Databases (VLDB)*, September 2000.

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

[11] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1–7):161–172, 1998.

[12] F. Douglis, T. Ball, Y.-F. Chen, and E. Koutsofios. The AT&t internet difference engine: Tracking and viewing changes on the web. *World Wide Web*, 1(1):27–44, 1998.

[13] F. Douglis, A. Feldmann, B. Krishnamurthy, and J. Mogul. Rate of change and other metrics: a live study of the world-wide web. In *USENIX Symposium on internetworking technologies and systems*, December 1997.

[14] J. Edwards. Private communication, 2002.

[15] J. Edwards, K. S. McCurley, and J. A. Tomlin. An adaptive model for optimizing performance of an incremental web crawler. In *Tenth International World Wide Web Conference (WWW10)*, pages 106–113, 2001.

[16] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. A large-scale study of the evolution of web pages. In *Twelfth International World Wide Web Conference (WWW12)*, pages 669–678, 2003.

[17] D. Florescu, A. Y. Levy, and A. O. Mendelzon. Database techniques for the world-wide web: A survey. *SIGMOD Record*, 27(3):59–74, 1998.

[18] Google. http://www.google.com.

[19] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform url sampling. In *Proceedings of 9th International WWW Conference*, pages 295–308, 2000.

[20] Inktomi Corp. Web surpasses one billion documents, January 2000. http://www.inktomi.com/news/-press/2000/billion.html.

[21] Internet Archive. Wayback machine. http://www.archive.org.

[22] Internet Software Consortium. Internet domain survey, January 2002. http://www.isc.org/ds/WWW-200201/index.html.

[23] C. Labovitz, A. Ahuja, and M. Bailey. Shining light on dark address space, 2001. http://www.nanog.org/mtg-0110/ppt/malan.

[24] S. Lawrence and C. L. Giles. Searching the world wide web. *Science*, 280:98–100, April 1998.

[25] S. Lawrence and C. L. Giles. Acessibility of information on the web. *Nature*, 400:107–109, July 1999.

[26] F. Menczer, G. Pant, and P. Srinivasan. Topic driven crawlers: Machine learning issues. submitted, University of Iowa, May 2002.

[27] B. H. Murray and A. Moore. Sizing the internet. white paper, Cyveillance, Inc., July 2000. http://www.cyveillance.com.

[28] M. Najork and J. L. Wiener. Breadth-first search crawling yields high-quality pages. In *Ten International World Wide Web Conference (WWW10)*, pages 114–118, 2001.

[29] Netcraft Inc. Netcraft web server survey. http://www.netcraft.com/survey.

[30] Online Computer Library Center, Inc. Size and growth. http://wcp.oclc.org/stats/size.html.

[31] Open Directory Project. http://dmoz.org/about.html.

[32] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bring order to the web. Technical report, Stanford University, January 1998.

[33] J. Pitkow and P. Pirolli. Life, death and lawfulness on the electronic frontier. In *Proceedings of ACM Conference on Computer-Human Interaction CHI 97*, March 1997.

[34] K. M. Risvik and R. Michelsen. Search engines and web dynamics.

[35] P. Rusmevichientong, D. M. Pennock, S. Lawrence, and C. L. Giles. Methods for sampling pages uniformly from the world wide web. In *AAAI Fall Symposium on Using Uncertainty Within Computation*, pages 121–128, 2001.

[36] R. H. Warren and D. Wilkinson. Choosing initial page-refresh rates for web-page collections updated with an adaptive refresh model, personal communication, 2002.

[37] R. H. Zakon. Hobbes' internet timeline v5.6. http://www.zakon.org/robert/internet/timeline.