Keeping Up With The Changing Web*

Brian E. Brewington
George Cybenko
Thayer School of Engineering
Dartmouth College
Hanover, New Hampshire 03755-8000

brian.e.brewington@Dartmouth.edu
george.cybenko@Dartmouth.edu

January 29, 2000

Abstract

Our access to information today is unprecedented in history. However, information depreciates in value as it gets older, and the problem of updating information to keep it current presents new design challenges for information providers and consumers. These issues lead to novel concepts and results in the context of the World Wide Web. We quantify what it means to for search engines to be "up-to-date" and estimate how often search engines must re-index the web to keep current with it changing pages and structure.

Three weeks prior to the Soviet invasion of Czechoslovakia, Corona satellite imagery of the area showed no signs of imminent attack. By the time another round of imagery was available, it was too late to react; the invasion had already taken place. In a real sense, the information obtained by the satellite weeks earlier was no longer useful. The fact that information has a useful lifetime is well known in the intelligence community.

On the other side of the Iron Curtain, an entirely different problem existed. The East German secret police, the Stasi, was especially vigilant in monitoring its own people – it is estimated that one of every three East Germans was a Stasi operative of some sort. The "missions" were simple, mostly consisting of monitoring neighbors and other people with whom they had frequent contact. Information was gathered in copious quantities, ranging from diaries to odor samples, all neatly cataloged and stored for future use. The basements of the former Stasi headquarters are filled with reports, observations and gossip about who was and was not engaged in some sort of subversive activity. This is to say nothing of the records that were destroyed as the Stasi fled their headquarters. Overwhelmed by the sheer volume of information they held, the Stasi was probably unable to use most of the reports retained.

While information overload has been given much attention, it is increasingly exacerbated by a new challenge: the dynamic nature of most information. This is true for almost all information sources: from a newspaper to a temperature sensor to the web. When monitoring an information source, when can we say that our previous observations have gotten stale and so need to be refreshed? How can these refresh operations be scheduled to satisfy a required level of "up-to-dateness" without violating resource constraints such as limitations on how much data can be observed per time unit (such as limited by bandwidth or computing resources)?

In this article, we investigate the tradeoffs that arise when monitoring a collection of dynamic information sources. We discuss the World Wide Web in detail, estimate how fast its documents change and explore what it means for a web index to stay current. For a simple class of web monitoring systems, namely

*This research was partially supported by AFOSR grant F49620-97-1-0382, DARPA grant F30602-98-2-0107 and NSF grant CCR-9813744. Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the above agencies.

search engines, our idea of "up-to-dateness" is combined with actual measured data to estimate revisit rates for web search engines. Though much of this discussion centers on the web, our ideas and results apply to more general information monitoring problems, arising in health care, finance, surveillance and vehicle maintenance for example, as well.

1 Information as a depreciating commodity

Most information has a useful lifetime – knowledge is a depreciating commodity. A good analogy is the purchase of a car, which begins to lose value as soon as it is driven off the lot. In the information domain, we expend resources (time, money and bandwidth for example) to obtain information but that information's value typically starts to depreciate immediately, as it gets stale.

Viewed as a commodity, information has two important characteristics: first, there is the initial value of having correct information, and second, there is a rate at which that value depreciates. The value of information is a very subjective and domain specific concept. For instance, compare the value of knowing that a bus is barreling down a street towards you, versus the value of knowing which Beanie Babies are up for grabs at eBay. "Value" must be quantified in some way but it depends on the context.

The second aspect of the "information as commodity" concept is how long the information is expected to be useful and at what rate its value depreciates. When do we next have to check for oncoming traffic? How often should we monitor the eBay site? Roughly speaking, when uncertainty becomes unacceptable, it is necessary to observe again. The idea of assigning values and depreciation rates to information and then optimizing observation costs was first formally explored in [CBB+97].

Given a large number of dynamically changing sources to monitor, and only limited resources (bandwidth and storage) available to do the monitoring, a decision must be made as to what should be observed next and when.

To get a feel for the tradeoffs involved, consider the following simple example. A source of information which changes daily, say a newspaper, requires daily observation. Clearly, observing a daily-changing source consumes about 30 times the bandwidth necessary to monitor a source that only changes monthly. Now suppose we have a collection of sources, some of which change daily and some of which change monthly. Consider these alternatives: we could either (1) observe a fast source daily, or (2) observe 30 different monthly sources. Which approach is better depends on which source is more valuable to us and what the probabilities of actually finding changes at those sources are.

These kinds of problems arise and are solved in our personal lives daily. How do you focus your attention while driving? What sections of the newspaper do you read first? Applications of information monitoring and scheduling arise in health care, weather forecasting, competitive business analysis, marketing and military intelligence.

2 Monitoring the World Wide Web

The World Wide Web is an excellent testbed for studying information monitoring. The web consists of a huge collection of decentralized web pages that are modified at random times by their maintainers. Search engines strive to keep track of the ever changing web by finding, indexing and re-indexing pages. How should observation resources be invested to keep users happy?

Solving the problem requires knowing the distribution of rates of change for documents on the web. A large sample of web page data (from a service of ours, the Informant[1]) gives us a starting point for exploring rates of change. There are two parts to the service; it monitors specific URLs for changes, and also runs standing user queries against one of four search engines[2] at user-specified intervals. Any of three events trigger notification of a user by email: (1) a monitored URL changes, (2) new results appear in the top results returned by a search engine in response to a standing query, or (3) any of the current top search results shows a change. A change, for purposes of this discussion, is any alteration of the web page, no matter how minor.

[1] http://informant.dartmouth.edu

[2] AltaVista, Excite, Infoseek, and Lycos

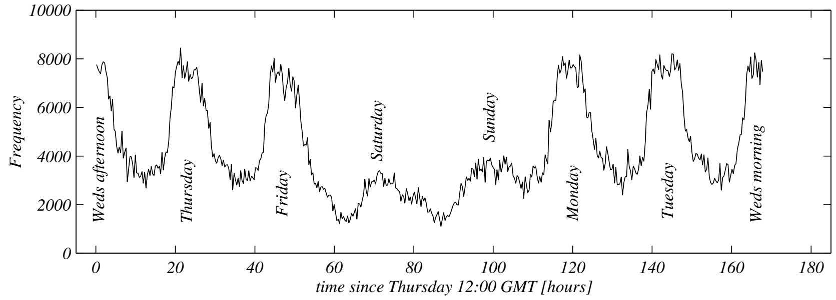

Figure 1: **Histogram: Last-modified times (GMT), mod 24×7 hours.** Peaks in modification frequency are clearly visible during US working hours, and diminish on weekends. Any assumption of stationarity in page alteration probability will break down at this scale.

Beginning in March 1999, we began storing HTML page summary information for all downloads. This has involved processing nearly 200 gigabytes of HTML data (about 100,000 web pages per day). The archived information includes the last-modified time stamp (if given), the time of observation (using the remote server's time stamp if possible), and stylistic information (content length, number of images, tables, links and similar data). The Informant selects and monitors these web pages in a very specific way. For details about the sampling and bias aspects of the data, we refer the reader to our longer technical report [BC99].

Our data has turned up some interesting artifacts about how the web changes. Last-modified timestamps, returned for about 65% of the documents observed, show that most web pages are modified during the span of US working hours (between around 8 AM and 8 PM, Eastern time). This is shown in Figure 1 using a histogram of where modification times fall during the week. The nonuniformity of the distribution means that page changes are not, strictly speaking, "memoryless" or stationary such as Poisson process would be. Another difficulty which arises from estimating page change rates using server timestamps is that this restricts the sample to a less dynamic part of the web. Often, timestamps are not given for documents which are not intended to be cached. Our estimates of rates of change, based on pages which do provide the time at which they were last modified, are therefore lower than the true change rates.

We can formalize web page dynamics, or the dynamics of any information source for that matter, by modeling the changes that occur as a renewal process [Pap84]. A good example and analogy is a system of replacement parts. Imagine a light fixture into which we place a lightbulb. Whenever that bulb burns out, it is replaced immediately. The time between lightbulb failures is the "lifetime" of a bulb. At a specific instant, the time since the present lifetime began is be the "age" of the bulb. The analogy to information source changes is that a lifetime is the time between changes (where change is arbitrarily but unambiguously defined). The age is the time between a given instant and the most recent change prior to that instant. We diagram these concepts in Figure 2.

Clearly, lifetimes and ages are related to one another; the age distribution of web documents roughly indicates the aggregate rate of change of the population. We can estimate the probability distribution of document age from our observations as shown in Figure 3. This data roughly matches that found by [DFKM97] for low-popularity pages. From the figure, it is clear that the web is very young: one in five web pages in our data sample is younger than twelve days, and one in four is younger than twenty days. What is less clear is why. It is difficult to say what portion of this trend is due to a preponderance of dynamic pages, and what is due to relatively static pages that haven't been online very long. Unfortunately, since web servers do not generally provide the time when the document first came online, there is no reliable way to use a single observation to guess whether a particular page is highly dynamic or just newly arrived. Without a growth model to distinguish between the two, it is difficult to calculate a distribution of change rates from the age data.

However, since the data include multiple observations of documents, we can use observed lifetimes to estimate change rates. As in Figure 2, the lifetime is measured by the difference in successive timestamps. The observed lifetime probability density and cumulative distribution functions (PDF and CDF) are shown

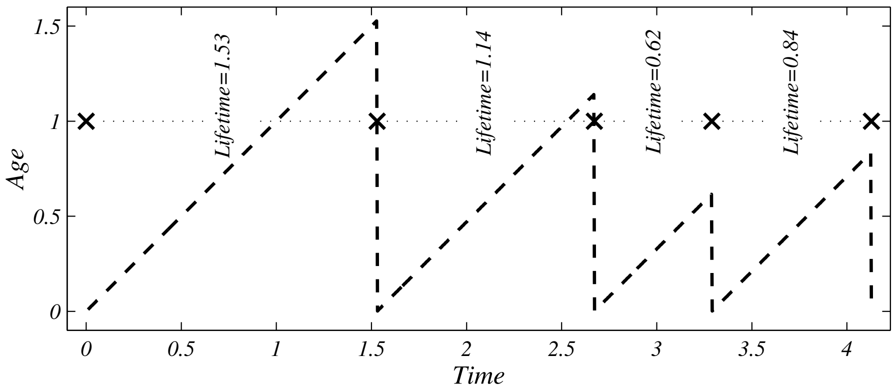

Figure 2: **Lifetimes vs. ages.** A single lifetime is represented by the time separating changes, which are denoted by $\times$'s in the graph. For each lifetime, the age (shown as a dashed line) increases linearly from 0 to the lifetime, then resets to 0 as the next lifetime begins.

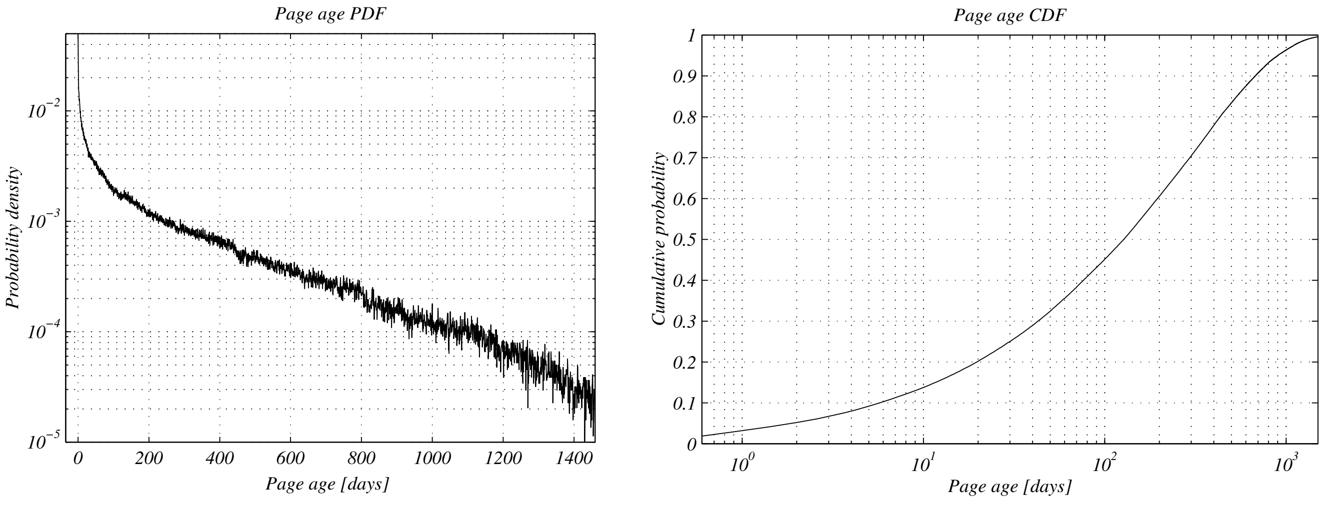

Figure 3: **Distribution of web page ages.** Here we show estimates of the probability density function (PDF) and cumulative distribution function (CDF) of web page age. On the left, we estimate the PDF using a rescaled histogram of web page ages, using one age observation per page. On the right, the corresponding CDF is formed by integrating the estimate of the PDF.

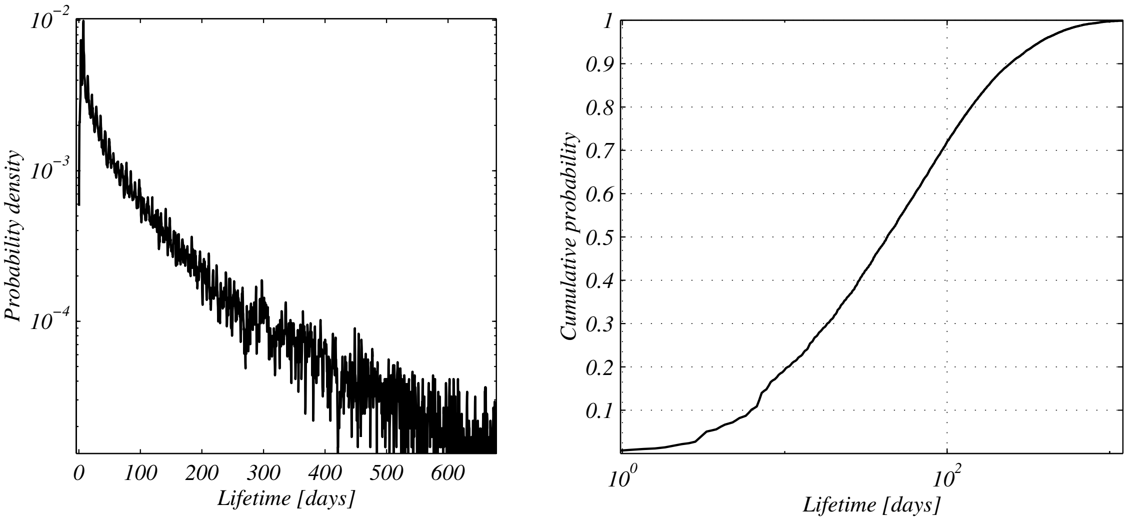

Figure 4: **PDF and CDF of observed web page lifetimes.** On the left, a rescaled histogram approximates the PDF of observed web page lifetimes, or differences in successive modification timestamps. On the right, we show the corresponding CDF. Note that these distributions are heavily influenced by the timespan of observation of single pages and also by the sample rate.

together in Figure 4. Inferring change rates from these observed lifetimes presents difficulties for pages that change either very quickly or very slowly. First, when a change is observed, there is no way to know whether this is the only change that happened since the last observation was made–this is essentially an aliasing problem. On the other end of the speed spectrum, changes that take a long time to happen are inherently less likely to be observed if a page is only monitored for a short time span. It is necessary to correct for both effects to find the underlying distribution of change rates.

Three simplifying assumptions help the estimation. First, assume that pages change according to independent Poisson processes (which we already know is an approximation by virtue of the nonuniform distribution of the times at which pages change), each characterized by an event rate λ. Some fast-changing pages change on a more periodic schedule, and some (no more than 4%) change on every observation, so this is not a perfect model. Still, Poisson (memoryless) processes[3] do a good job of modeling most of the pages. Second, we assume a parametric form (a Weibull distribution[4]) for the distribution of *mean* lifetimes for these Poisson processes. Third, we assume that the time for which a page is observed by the Informant is independent of its rate of change. This makes it possible to adjust for the limited timespan of observation (watching a source for a short time biases against changes that take a while to happen).

Enforcing these assumptions, we find the parameters that do the best job of producing the observed lifetime distribution. The optimization yields values that correspond to the distribution of mean change times shown in Figure 5. A quick look at the figure shows that the median value of the mean lifetime $\bar{t}$ is around 117 days, the fastest-changing quartile has $\bar{t} < 62$ days, and the slowest-changing quartile has $\bar{t} > 190$ days. Taking this as a prior distribution of mean change times makes estimation of change rates for individual pages more efficient, which could conceivably assist search engines in operating more efficiently and identifying documents that may never change. More on our estimation methods can be found in [BC99].

[3] Poisson processes, popular in queuing theory, have the property of being "memoryless": the probability of an event in any short time interval is independent of the time since the last event.

[4] Weibull distributions are two-parameter (one for scale, one for shape) generalizations of exponential distributions. See [MR94] for more.

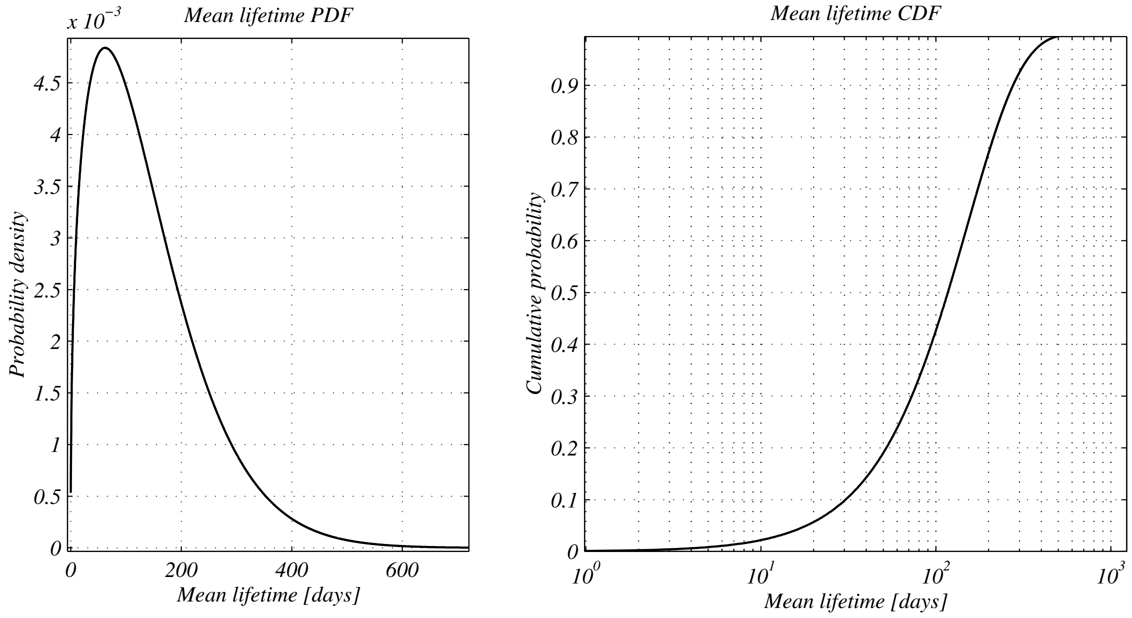

Figure 5: **Estimated PDF and CDF of mean web page lifetime $\bar{t}$.** Parameter estimation implies these distributions of mean lifetimes $\bar{t}$ for the documents observed by the Informant. Note that these *mean values* (or inverse change rates) are to be distinguished from the distribution of *observed* lifetimes shown in Figure 4. The average is around 138 days, the maximum of the PDF occurs at 62 days, and the median is 117 days.

3 The meaning of "up-to-date"

Our data and the above analysis gives us some idea of how fast web pages are changing. But what does it mean to say that a web search engine, for example, is current or "up-to-date"? We can't expect search engines to be completely current on the whole web all the time, knowing what new pages are being created instantaneously. So being realistic about being "up-to-date" or "current" requires us to relax time and certainty in some way. This is the motivation behind the concept of (α, β)-currency introduced below.

First of all, we define a web page entry in a search engine to be β-current if the web page has not changed since the last time it was checked by the search engine and β time units ago. In this concept, β is a time relaxation or a grace period. We don't expect a search engine to have instantaneous knowledge. β-currency is graphically depicted in Figure 6.

This is a natural concept, one that we implicitly use all the time. For example, a morning daily newspaper would be "12 hours"-current when you read it in the morning. The news in the paper would be current up to the grace period of 12 hours, the time required to write a news story, edit, print and distributed the paper. We don't expect news that occurred one hour ago to be in the newspaper.

Since web pages can change at random times, there is no way to guarantee that a search engine is, for example, 1 week current without requiring a search engine to download and index the whole web every week. This is where the random distribution of web page change times plays a role. Some pages change frequently and should be checked often, some pages rarely change and can be checked much less frequently. Checking quickly changes pages more often than slowly changing pages makes the problem feasible in the sense that the whole web does not have to be checked at some unreasonably small time interval. On the other hand, we lose the guarantee that every page monitored by the search engine will be β-current for some β.

We therefore introduce a probability, α, that the search engine is β-current with respect to a random web page. This is precisely the concept of (α, β)-currency. We say that a search engine is (α, β)-current if the probability that a randomly chosen web page has a β-current entry is at least α. A (0.9, 1 week)-current search engine would have a probability of 0.9 that a randomly chosen page is properly indexed as of one

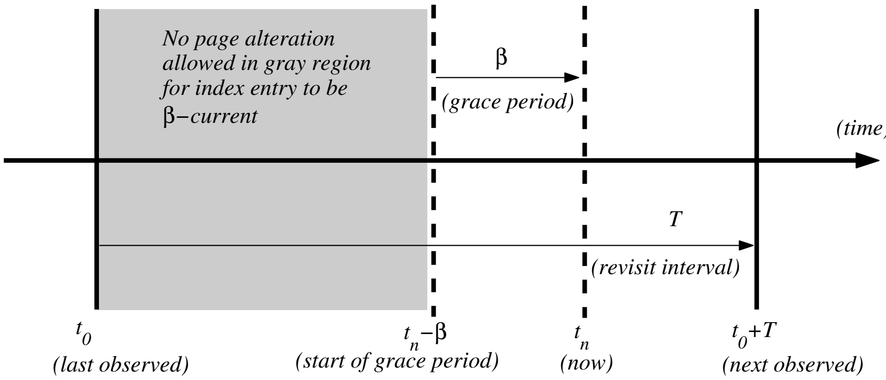

Figure 6: **Definition of "β-current".** This diagram shows what is meant when we say that an index entry is current with respect to a grace period, β. In order to be β-current, no modification can go unobserved up to β time units before the present.

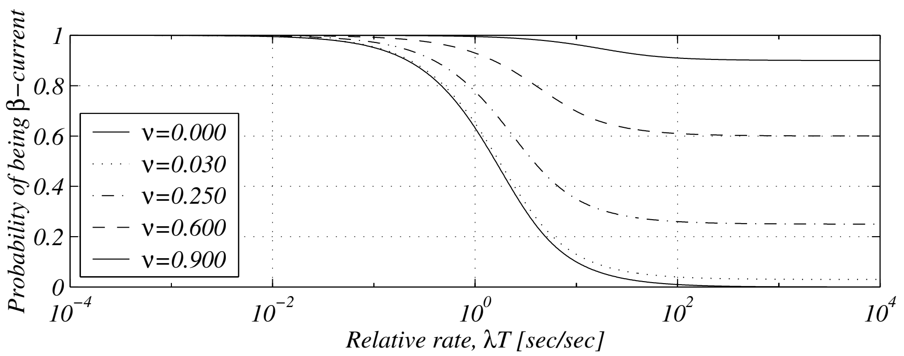

Figure 7: **Probability of β-currency vs. relative rate for a single source.** For a single Poisson source having change rate λ, re-indexed with period T, we plot the expected value of the probability α as a function of relative rate $z = \lambda T$ [dimensionless] and grace period percentage $\nu = \beta/T$ during which unobserved changes are forgiven.

week ago.

4 How fast must observers work?

Applying the concept of (α, β)-currency, we can use our data on the rate of change of web documents in order to estimate how quickly a search engine needs to work to maintain a given level of (α, β)-currency. The most naive possible re-indexing strategy, and also the simplest, is to re-observe every information source periodically, so that each document is seen every T time units (on average). Figure 7 shows how the probability α varies as a function of the (dimensionless) relative re-indexing rate, λT, and the grace period fraction, $\nu = \beta/T$. Note that as the revisit interval, λT, grows the probability, α, approaches the fraction of time $\nu = \beta/T$ during which unobserved changes are allowed. For large λT, an observation becomes worthless almost immediately because pages are changing much faster than the re-indexing interval and so on average only the fraction of all re-indexes falling with the grace period, namely β/T will be β-current. Conversely, many extra observations are performed when λT is small so that α approaches one as λT approaches zero.

The currency of a specific re-indexing strategy (used for example by a search engine) can be evaluated by averaging the probability of being β-current (as plotted in Figure 7) over the observed distribution of

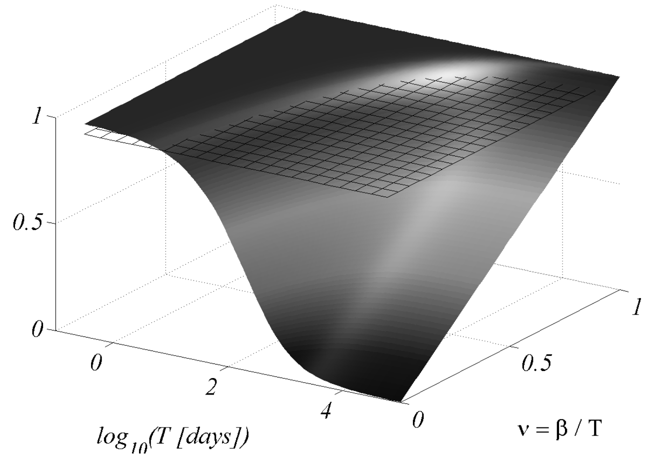

Figure 8: **Probability α for an entire collection as a function of $\nu = \beta/T$ and T:** The probability surface α for a collection of sources with a rate distribution diagrammed in Figure 5. The plane at $\alpha = 0.95$ intersects the surface in a level set, which is plotted in Figure 9 (with β values used instead of percentages ν).

web page change rates. The probability of a search engine being β-current is a function of the re-indexing period, T, and the grace period, β,[5] giving a probability α for each pair (T, β). Using our estimates of mean web page change times, a general performance surface in which α is a function of β and T can be derived. Figure 8 depicts the surface for our empirical data. Passing a plane through the surface gives a level set of all possible pairs of re-indexing period T and grace period fraction ν having a certain probability of a randomly-chosen source being β-current. Clearly, other re-indexing schemes exist where T is not constant but is a function of λ; see [CLW97] for a good discussion of such schemes.

Using the level set plotted in Figure 9, a (0.95, 1-day)-current web search engine needs a re-indexing period of 8.5 days. For (0.95, 1-week)-currency, that re-indexing period becomes 18 days. Notice that these figures do not depend upon the number of documents in an index, so a re-indexing period defines a set of pairs (α, β), regardless of changes in the size of the index. Alternatively, total bandwidth requirements to maintain a given level of currency can be estimated for a uniform index of a given size. By "uniform" we mean that no documents are given any sort of preference; all are re-indexed at the same rate. For example, an (0.95, 1-day) index of the entire web, using the estimate of 800 million pages from [LG99], and an average page size of 12 kilobytes[BC99], would require a total bandwidth of (approximately)

$$\frac{800 \times 10^6 \text{ pages}}{8.5 \text{ days}} \times \frac{12 \text{ kilobytes}}{1 \text{ page}} = \frac{104 \text{ Mbits}}{\text{sec}} \text{ for (0.95, 1-day) currency of index of the entire web.}$$

A more modest index, closer to those actually in use, might have 150 million documents at (0.95, 1-week) currency, requiring a bandwidth of around

$$\frac{150 \times 10^6 \text{ pages}}{18 \text{ days}} \times \frac{12 \text{ kilobytes}}{1 \text{ page}} = \frac{9.4 \text{ Mbits}}{\text{sec}} \text{ for (0.95, 1-week) currency of index of around 1/5 of the web.}$$

5 Summary

New communication technology allows access to huge amounts of information. The problem of how to focus our attention will grow dramatically in the next millennium, especially as we increasingly recognize that information changes, grows, and becomes obsolete. This problem is already acute for web search engines that must periodically re-index the web in order to stay current.

[5] We used the same grace period for all pages; this is not necessary.

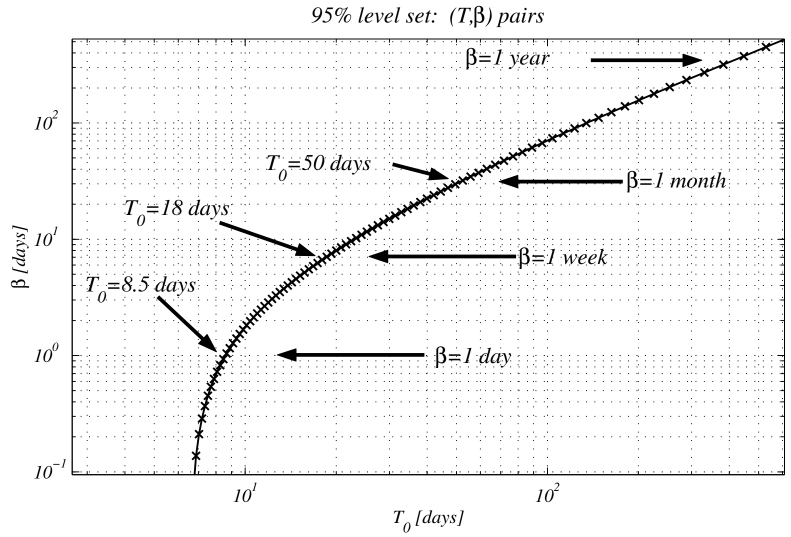

Figure 9: **Relating β and T: α=95% level set.** This is the level set of pairs (T, β) which for probability $\alpha = 0.95$ of being β-current. Regardless of the size of the collection, this data can be used to estimate how current an engine is when the indexing period T takes on a value (in days) along the horizontal axis. As T becomes large, the relative check rate is too slow, and β approaches $0.95T$.

This article barely scratches the surface of what needs to be done and what can be done. Our models are elementary and our data collection strategies are biased. Nonetheless, several useful concepts have been developed here and a preliminary analysis of what it means for a web search engine to be current has been derived.

Future work in this area must explore the role of weighting on web pages – namely some pages are more popular or important than others and that should be taken into account. Secondly, our re-indexing scheme is based on a single revisit period for all pages. If the re-indexing period is allowed to vary with the page, the optimal re-indexing strategy becomes a complex optimization problem that must be solved numerically, using all the empirical data available for individual page change rates.

No only are some pages more important than others, some page changes are more important than others. This adds a whole new dimension to the problem of estimating re-indexing rates, one in which re-indexing depends on the type of change expected to be found. Our data allows such modeling but we have yet to undertake the required modeling and analysis.

The ideas presented here are generic to the problem of monitoring many changing sources of information subject to constraints in monitoring resources. We believe that other areas of application include health care, surveillance, investing and finance.

References

[BC99] Brian Brewington and George Cybenko. How fast is the web changing? Preprint; available from http://actcomm.dartmouth.edu/papers/ as either brewington:webchange.ps.gz or brewington:webchange.pdf, 1999.

[CBB+97] George V. Cybenko, Aditya Bhasin, Brian Brewington, Robert Gray, Katsuhiro Moizumi, and Kartik Raghavan. The Shannon Machine: A system for networked communication and computation. Available via anonymous ftp://witness.dartmouth.edu/pub/shannon.ps, 1997.

[CLW97] E. G. Coffman, Z. Liu, and R. R. Weber. Optimal robot scheduling for web search engines. *Journal of Scheduling*, 1997. Available at http://www.inria.fr/mistral/personnel/Zhen.Liu/Papers/.

[DFKM97] Fred Douglis, Anja Feldmann, Balachander Krishnamurthy, and Jeffrey Mogul. Rate of change and other metrics: A live study of the world wide web. In *Proceedings of the USENIX Symposium on Internetworking Technologies and Systems*, December 1997. Available from http://www.research.att.com/~anja/feldmann/papers.html.

[LG99] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 1999.

[MR94] Douglas C. Montgomery and George C. Runger. *Applied Statistics and Probability for Engineers.* John Wiley and Sons, Inc., 1994.

[Pap84] Athanasios Papoulis. *Probability, Random Variables and Stochastic Processes.* McGraw-Hill, 2nd edition, 1984.