Keeping Up with the Changing Web

Most information depreciates over time, so keeping Web pages current presents new design challenges. This article quantifies what "current" means for Web search engines and estimates how often they must reindex the Web to keep current with its changing pages and structure.

Brian E. Brewington
George Cybenko
Dartmouth College

Three weeks before 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 occurred. In a real sense, information the satellite obtained weeks earlier was no longer useful. As this example vividly illustrates, members of the intelligence community know all too well that information has a limited useful lifetime.

Elsewhere, an entirely different problem existed. The East German secret police—the Stasi—were especially vigilant in monitoring their own people. Analysts estimate that one of every three East Germans was a Stasi operative of some sort. Their missions were simple, mostly involving monitoring neighbors and other people with whom they had frequent contact. The Stasi gathered information in copious quantities, ranging from diaries to odor samples, all neatly cataloged and stored for future use. (The Stasi collected odor samples by rubbing bits of chemically treated fabric over surfaces contacted by an individual under observation. They then sealed these pieces of fabric in test tubes and cataloged them; despite being stored for long periods of time, the smell of the fabric persisted.) 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. And this doesn't even include the records that the Stasi destroyed as they fled their headquarters. Overwhelmed by the sheer volume of information they held, the Stasi were probably unable to use most of the reports they gathered.

Information overload can certainly be a problem, but another, perhaps less apparent, challenge increasingly exacerbates it: Most information—from a newspaper story to a temperature sensor measurement to a Web page—is dynamic. When monitoring an information source, when do our previous observations become stale and need refreshing? How can we schedule these refresh operations to satisfy a required level of currency without violating resource constraints (such as bandwidth or computing limitations on how much data can be observed in a given time)?

We investigate the trade-offs involved in monitoring dynamic information sources and discuss the Web in detail, estimating how fast its documents change and exploring what constitutes a "current" Web index. For a simple class of Web-monitoring systems—search engines—we combine our idea of currency with actual measured data to estimate revisit rates.

INFORMATION: A DEPRECIATING COMMODITY

Most information depreciates over time. A good analogy is purchasing a new car: As soon as we drive it off the lot, its value begins decreasing. In the information domain, we expend resources (time, money, and bandwidth) to obtain information. But that information's value typically starts to depreciate immediately, and it gets stale.

Viewed as a commodity, information has two important characteristics: its initial value and the rate at which that value depreciates.[1] The value of information is subjective and domain specific. For instance, compare the value of knowing that a bus is barreling down a street toward you with that of knowing which Beanie Babies are up for grabs at eBay. Somehow, we must quantify information's value.

The second aspect of the information-commodity concept involves determining how long the informa-

0018-9162/00/$10.00 © 2000 IEEE

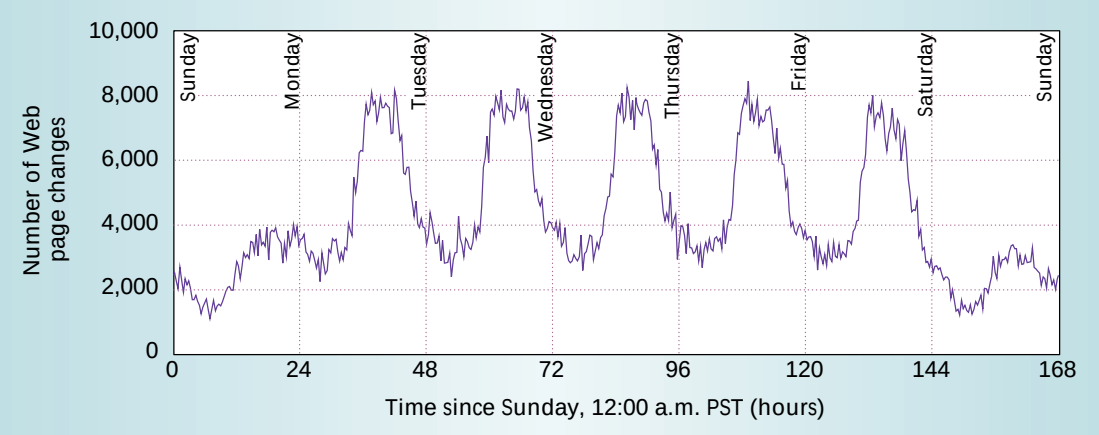

Figure 1. Histogram showing typical frequency of Web page changes in a week (24 hours/day × 7 days/week), based on last-modified times (Pacific Standard Time). Peaks in modification frequency are clearly visible during US working hours and diminish on weekends.

tion remains useful and at what rate its value depreciates. When should we check for oncoming traffic again? How often should we monitor the eBay site? Roughly speaking, when uncertainty becomes unacceptable, we need to observe again. With many dynamically changing sources to monitor and only limited monitoring resources (bandwidth and storage), we must decide what to observe next and when.

To better understand the trade-offs, consider the following simple example. An information source that changes daily—say, a newspaper—requires daily observation. Observing a daily-changing source consumes about 30 times the bandwidth necessary to monitor a source that changes only monthly. Now suppose we have a collection of sources—some changing daily, others monthly—and we can observe only one source per day. We have two alternatives: Observe a fast source daily or observe 30 different monthly sources (one each day). Our choice will depend on each source's value to us, along with the probability of finding changes at that source.

We face such problems daily in our personal lives. 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 healthcare, weather forecasting, competitive business analysis, marketing, and military intelligence. The problem of staying current with resources on the Web is perhaps the simplest application area to study as a prototype.

MONITORING THE WEB

The Web—a huge collection of decentralized Web pages modified at random times—is an excellent testbed for studying information monitoring. Search engines strive to keep track of the ever-changing Web by finding, indexing, and reindexing pages. How should we invest observation resources to keep users happy?

Solving this problem requires knowing the distribution of Web-document change rates. A large sample of Web page data from a service of ours, the Informant (http://informant.dartmouth.edu), gives us a starting point for exploring change rates. The Informant accomplishes two tasks: It monitors specific URLs for changes, and it runs standing user queries against one of four search engines (AltaVista, Excite, Infoseek, or Lycos) at user-specified intervals. The Informant notifies the user (by e-mail) if a monitored URL changes, new results appear in the top results returned by a search engine in response to a standing query, or 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.

Beginning in March 1999, we began storing Hypertext Markup Language page summary information for all downloads. This 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.[2]

Our data reveals some interesting artifacts about how the Web changes. Last-modified time stamps, returned for about 65 percent of the documents observed, show that most Web pages are modified during US working hours (5 a.m. to 5 p.m., Pacific Standard Time). This is clear in Figure 1, a histogram of where modification times fall during the week. The distribution's lack of uniformity means page changes are not, strictly speaking, memoryless or stationary, as a Poisson process would be. (Poisson processes, popular in queuing theory, are memoryless: The probability of an event in any short time interval is independent of the time since the last event.) Another

problem with estimating page change rates using server time stamps is that doing so restricts the sample to a less dynamic part of the Web. Documents that change frequently and aren't intended for caching typically do not include time stamps. Therefore, because we base our change rate estimates on pages that have time stamps (which provide the times those pages were last modified), these estimates are lower than the true change rates.

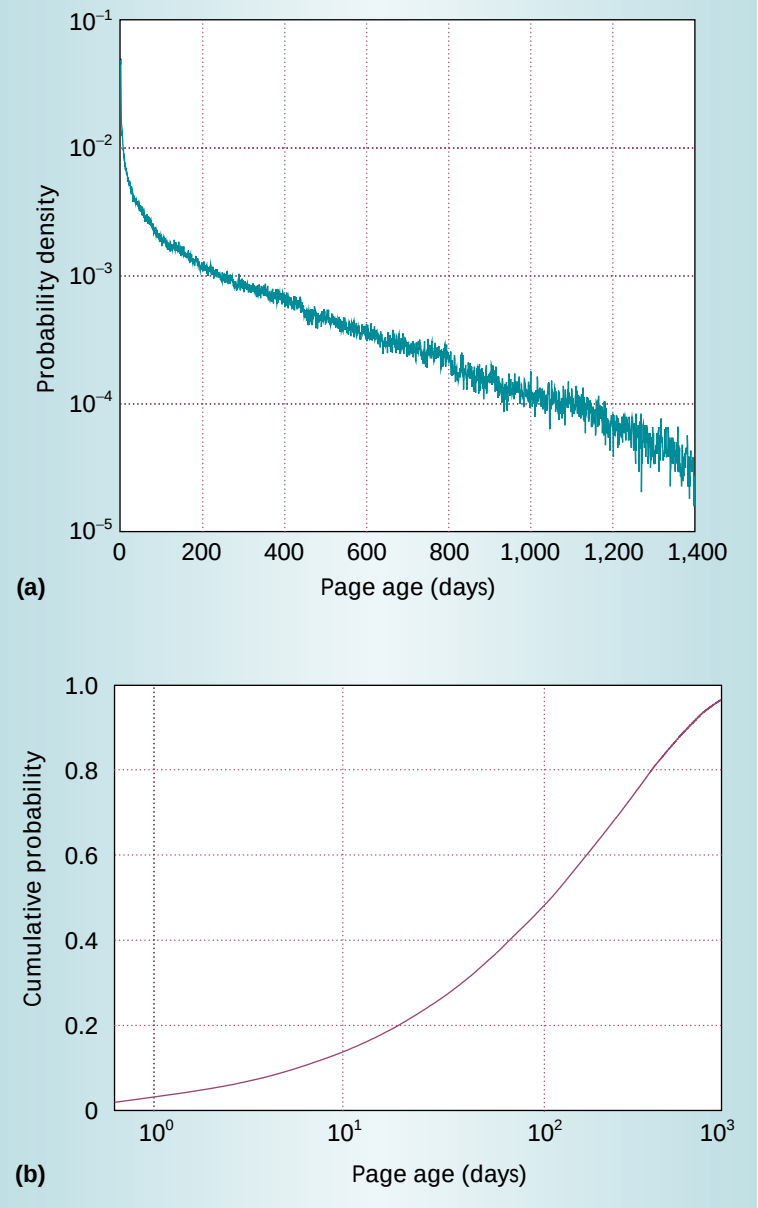

Figure 2. Estimated distribution of Web page ages. We estimate the probability density function (PDF) using (a) a rescaled histogram of Web page ages, one age observation per page, and form the corresponding (b) cumulative distribution function (CDF) by integrating the estimate of the PDF.

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.[3] A good analogy is replacing a light bulb. Whenever the bulb burns out, we replace it. The time between light bulb failures is the bulb's *lifetime*. At a specific instant, the time since the present lifetime began is the bulb's *age*.

Clearly, lifetime and age are related; the age distribution of Web documents roughly indicates the population's aggregate change rate. We can estimate the probability distribution of document age from our observations, as shown in Figure 2. (This data roughly matches that found by Fred Douglis and his colleagues for low-popularity pages.[4]) As the figure shows, the Web is very young: One in five Web pages in our data sample is younger than 12 days, and one in four is younger than 20 days. What is less clear is why. It is difficult to say which portion of this trend is due to a preponderance of dynamic pages, and which is due to relatively static pages that haven't been online very long. Unfortunately, Web servers do not usually provide the time when the document first came online. Hence, 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, calculating a change rate distribution from the age data is difficult.

However, because the data includes multiple observations of documents, we can use observed lifetimes to estimate change rates. We measure the lifetime as the difference in successive time stamps; Figure 3 shows the observed lifetime probability density function (PDF) and cumulative distribution function (CDF). Inferring change rates from these observed lifetimes presents difficulties for pages that change either very quickly or very slowly. For pages that change very quickly, there is no way to know whether an observed change is the only change since the last observation (this is essentially an aliasing problem). Pages that change very slowly also cause difficulties, because we are inherently less likely to observe their changes if we monitor the page for a short time. To find the underlying change-rate distribution, we must correct for both effects.

Three simple assumptions help our estimation. First, we 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 event rate λ. Some fast-changing pages change on a more periodic schedule, and some (no more than 4 percent) change on every observation, so this is not a perfect model. Still, Poisson (memoryless) processes adequately model most pages.

Second, we assume that the distributions of these Poisson processes' mean lifetimes are in parametric form

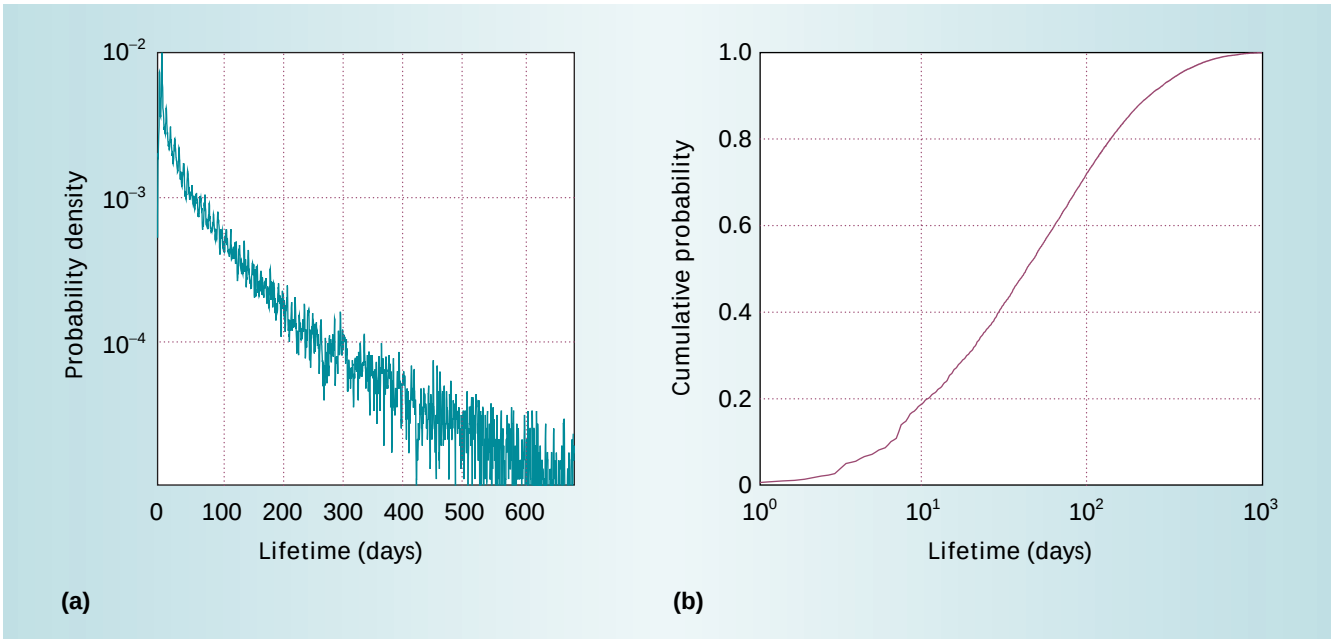

Figure 3. Observed distribution of Web page lifetimes: (a) a rescaled histogram approximates the PDF of observed Web page lifetimes or differences in successive modification time stamps and (b) the corresponding CDF. Both the observation time span for single pages and the sample rate heavily influence these distributions.

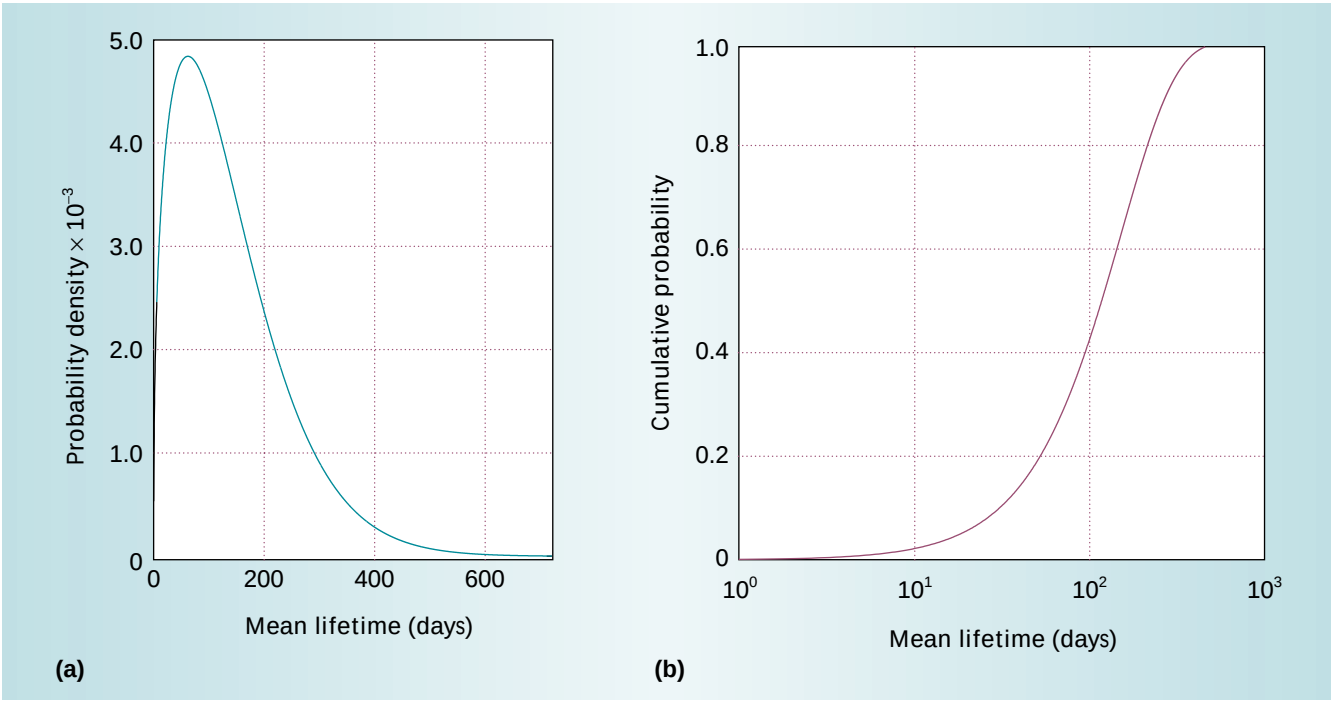

Figure 4. Estimated distribution of mean Web page lifetimes $\bar{t}$: (a) PDF and (b) CDF. Parameter estimation yields these distributions of mean lifetimes $\bar{t}$ for the documents that the Informant observes. We must distinguish these mean values *(or inverse change rates) from the* observed *lifetimes shown in Figure 3. For the PDF, the average is around 138 days; the maximum occurs at 62 days, and the median is 117 days.*

(Weibull distributions: two-parameter—one for scale, one for shape—generalizations of exponential distributions).[5]

Third, we assume that the time for which the Informant observes a page is independent of that page's change rate. This lets us adjust for the biases that watching a source for a short time introduces when changes in that source take a while to occur.

Enforcing these assumptions, we find the parameters that best estimate the observed lifetime distribution. The optimization yields values that correspond to the distribution of mean change times shown in Figure 4. 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 $\bar{t}$ less than 62 days, and the slowest-changing quartile has $\bar{t}$ greater than 190 days. Using this estimate of the lifetime distribution as an a priori distribution of mean change times makes estimating change rates for individual pages more efficient. Search engines can then operate more efficiently and identify documents that may never change.[2]

WHAT DOES "CURRENT" MEAN?

Our data and the above analysis give us an idea of how fast Web pages are changing. But what constitutes a "current" Web search engine? We can't expect search engines to be completely current on the entire Web all the time, instantly identifying every new page created. We need to relax the time and certainty requirements somehow.

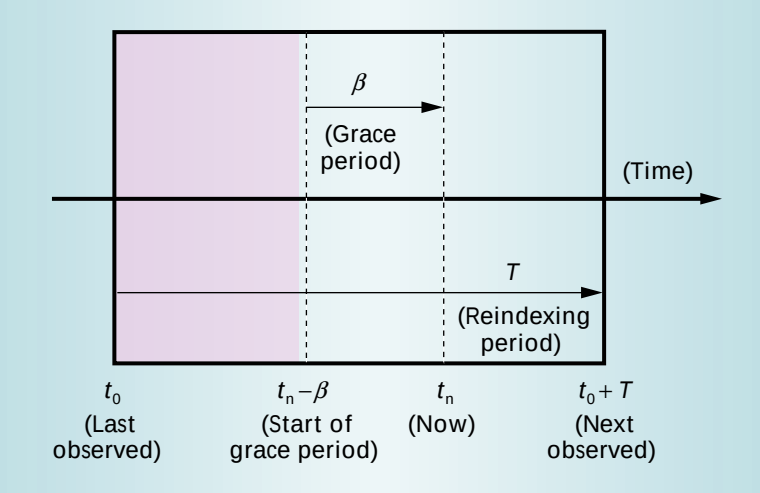

Figure 5. Definition of β-currency. For our knowledge of a page to be β-current, no change can occur during (the shaded region in the figure) the time between our last observation and the beginning of the grace period. However, the page is still β-current even when changes occurring during the grace period (up to β time units before the present) go unnoticed.

Therefore, we define a Web page entry in a search engine to be *β-current* if the Web page has not changed between the last time the search engine checked it and β time units ago. In this context, β is a grace period. We don't expect a search engine to have instantaneous knowledge. Figure 5 illustrates β currency.

We implicitly use this concept 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, edit, print, and distribute the paper. We don't expect news that occurred one hour ago to be in the newspaper.

Web pages can change at random times. Therefore, we cannot guarantee that a search engine is, for example, one week current unless it downloads and indexes the entire 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; others rarely change and can be checked much less frequently. Checking quickly changing pages more often than slowly changing pages makes the problem feasible because we don't need to check the entire Web at an 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 the probability α that the search engine is β-current with respect to a random Web page. Then we can say a search engine is (α; β)-current if the probability of a randomly chosen Web

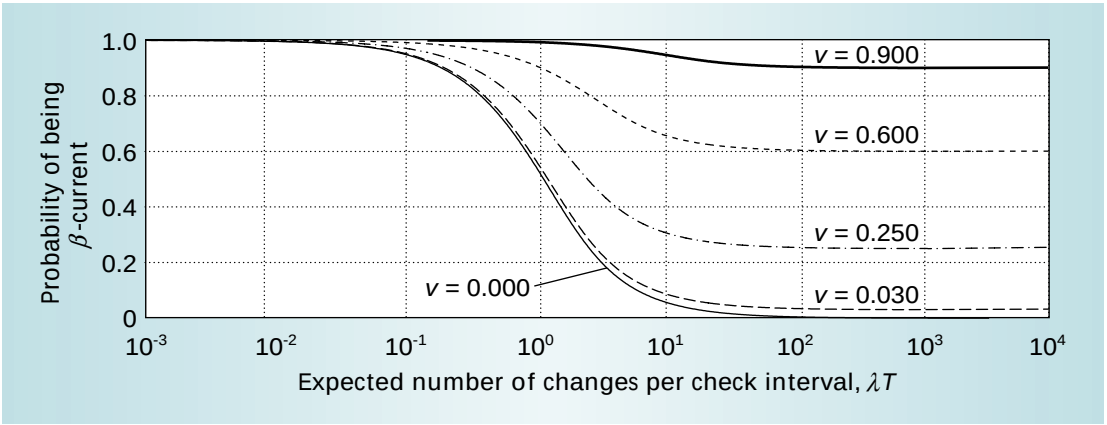

Figure 6. Probability of β-currency versus expected number of changes per check interval, for a single source. For a single Poisson-process source having change rate λ, reindexed with period T, we plot the expected value of probability α as a function of expected number of changes per check interval z = λT and grace-period fraction v = β/T (during which unobserved changes are forgiven).

page having a β-current entry is at least α. For example, the probability that a (0.9; 1 week)-current search engine would properly index a randomly chosen page given a grace period of one week would be 0.9.

HOW FAST MUST OBSERVERS WORK?

Applying the concept of (α; β)-currency, we can use our data on Web document change rates to estimate how quickly a search engine must work to maintain a given level of (α; β)-currency. The most naive reindexing strategy, and the simplest, is to reobserve every information source periodically so that the search engine visits each document every T time units on average. Figure 6 shows how probability α varies as a function of (dimensionless) relative reindexing time λT and grace period fraction $v = \beta/T$. As the relative reindexing time λT grows, probability α approaches fraction of time $v = \beta/T$ when unobserved changes are allowed. For large λT, an observation becomes worthless almost immediately, because pages are changing far more quickly than the reindexing interval. Therefore, on average, only the fraction β/T of all reindexes falling within the grace period will be β-current. Conversely, the Informant performs many extra observations when λT is small, so α approaches 1 as λT approaches 0.

We can evaluate the currency of a specific reindexing strategy (for example, a search engine's) by averaging the probability of being β-current (as plotted in Figure 6) over the observed distribution of Web page rates. This probability is a function of reindexing period T and grace period β (we used the same grace period for all pages, but this is not necessary), giving probability α for each pair (T, β). Using our estimates of mean Web page change times, we can derive a general performance surface where the z-coordinate, α, is a function of β and T.

Figure 7 depicts the surface for our empirical data. Passing a horizontal plane through the surface gives a level set (the set of all values on the surface) of all possible pairs of reindexing period T and grace period fraction v having a given z-coordinate α (the probability of a randomly chosen source being β-current). In some reindexing schemes, T is not constant but is a function of λ.[6]

Using the level set plotted in Figure 8, a (0.95; 1-day)-current Web search engine needs a reindexing period of 8.5 days. For (0.95; 1-week)-currency, that reindexing period becomes 18 days. These numbers do not depend on the number of documents in an index, so a reindexing period defines a set of pairs (α; β), regardless of changes in index size. Alternatively, we can estimate total bandwidth requirements to maintain a given level of currency for a uniform index of a given size. By "uniform" we mean not giving any particular documents preference; the search engine reindexes all documents at the same rate. For example, given approximately 800 million total Web pages[7] with an average page size of 12 Kbytes,[2] a (0.95; 1-day) index of the entire Web would require a total bandwidth of approximately

$$\frac{800 \times 10^6 \text{ pages}}{8.5 \text{ days}} \times \frac{12 \text{ Kbytes}}{1 \text{ page}} = \frac{104 \text{ Mbits}}{\text{second}}$$

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

$$\frac{150 \times 10^6 \text{ pages}}{18 \text{ days}} \times \frac{12 \text{ Kbytes}}{1 \text{ page}} = \frac{9.4 \text{ Mbits}}{\text{second}}$$

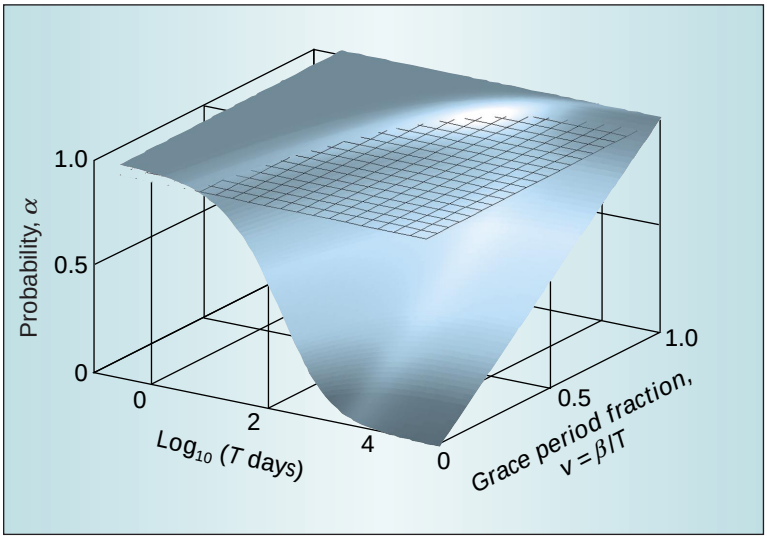

Figure 7. Probability α for an entire collection as a function of $v = \beta/T$ and T. This probability surface is for a collection of sources, with the rate distribution shown in Figure 4. The plane at $\alpha = 0.95$ intersects the surface in a level set (see Figure 8).

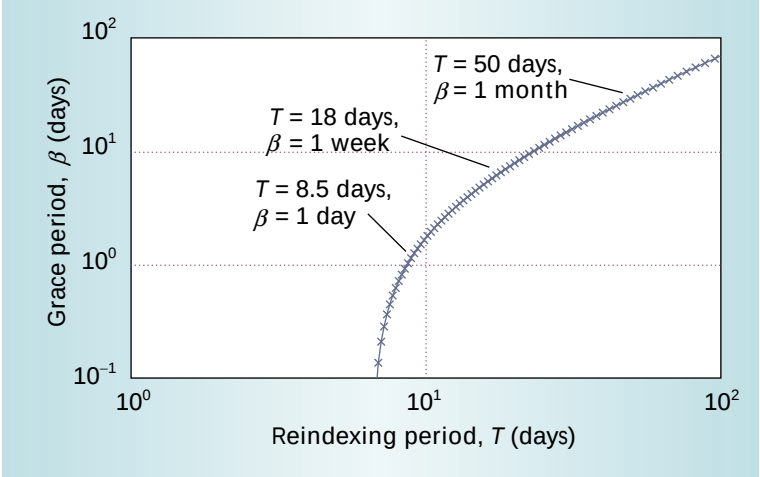

Figure 8. Level set of pairs (T, β) having a probability $\alpha = 0.95$ of being β-current. Regardless of the collection size, this data can help us estimate how current an engine is when the indexing period T has a particular value (in days) along the horizontal axis. As T becomes large, the expected number of changes per check interval becomes too low, and β approaches 0.95T. In other words, when observing very slowly, the only way to obtain 0.95 probability of β-currency is to forgive changes almost 95 percent of the time.

New communication technology allows access to huge amounts of information. The importance of focused information searches will grow dramatically in the next decade, especially as information changes, grows, and becomes obsolete. This problem is already acute for Web search engines that must periodically reindex the Web to stay current.

This article has barely scratched the surface of what needs to and can be done. Our models are elementary, and our data collection strategies are biased. Future research must explore the role of weighting Web pages—some pages are more popular or "important" than others, and we need to account for this consideration. Moreover, we base our reindexing scheme on a single revisit period for all pages. If the reindexing period varies with the page, the optimal reindexing strategy becomes a complex optimization problem that we must solve numerically, using all the empirical data available for individual-page change rates.

Not only are some pages more important than others, but some page *changes* are more important than others. This adds a whole new dimension to the problem of estimating reindexing rates, one in which reindexing depends on the type of change expected. Our data allows such modeling, but we have yet to undertake this modeling and analysis.

Although this discussion focused on the Web, our ideas and results apply to information-monitoring problems in many other areas, such as healthcare, finance, surveillance, and vehicle maintenance. ✸

Acknowledgments

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 these agencies.

References

1. G.V. Cybenko et al., "The Shannon Machine: A System for Networked Communication and Computation," *Proc. 9th Yale Workshop Adaptive and Learning Systems*, Center for Systems Science, Dunham Laboratory, Yale Univ., New Haven, Conn., 1996.
2. B. Brewington and G. Cybenko, "How Fast Is the Web Changing?" *Proc. 9th Int'l World Wide Web Conf.*, Elsevier Science, Amsterdam, 2000 (to be published).
3. A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 2nd ed., McGraw-Hill, New York, 1984.
4. F. Douglis et al., "Rate of Change and Other Metrics: A Live Study of the World Wide Web," *Proc. Usenix Symp. Internet Technologies and Systems*, Usenix, Lake Forest, Calif., 1997; http://www.Usenix.org/publications/library/proceedings/usits97/douglis_rate.html.
5. D.C. Montgomery and G.C. Runger, *Applied Statistics and Probability for Engineers*, John Wiley & Sons, New York, 1994.
6. E.G. Coffman, Z. Liu, and R. Weber, "Optimal Robot Scheduling for Web Search Engines," *J. Scheduling*, Vol. 1, 1998, pp. 15-29.
7. S. Lawrence and C.L. Giles, "Accessibility of Information on the Web," *Nature*, Vol. 400, 1999, pp. 107-109; http://www.wwwmetrics.com/.

Brian E. Brewington is conducting doctoral research at the Thayer School of Engineering at Dartmouth College. His research interests include distributed information retrieval and signal processing. Brewington has a BS in engineering and applied science from the California Institute of Technology, Pasadena. Contact him at brian.brewington@ dartmouth.edu.

George Cybenko is the Dorothy and Walter Gramm Professor of Engineering at Dartmouth College. His research interests include distributed information systems, mobile code, and security. Cybenko has a BSc in mathematics from the University of Toronto and a PhD in mathematics from Princeton University. He is a Fellow of the IEEE. Contact him at gvc@dartmouth.edu.