Trend Detection Through Temporal Link Analysis

Einat Amitay, David Carmel, Michael Herscovici, Ronny Lempel, and Aya Soffer
IBM Research Lab in Haifa, Haifa 31905, Israel. E-mail: {einat, carmel, miki, rlempel, ayas}@il.ibm.com

Although time has been recognized as an important dimension in the co-citation literature, to date it has not been incorporated into the analogous process of link analysis on the Web. In this paper, we discuss several aspects and uses of the time dimension in the context of Web information retrieval. We describe the ideal case—where search engines track and store temporal data for each of the pages in their repository, assigning timestamps to the hyperlinks embedded within the pages. We introduce several applications which benefit from the availability of such timestamps. To demonstrate our claims, we use a somewhat simplistic approach, which dates links by approximating the age of the page's content. We show that by using this crude measure alone it is possible to detect and expose significant events and trends. We predict that by using more robust methods for tracking modifications in the content of pages, search engines will be able to provide results that are more timely and better reflect current real-life trends than those they provide today.

Introduction

Time is a very important dimension in the cocitation literature (Garfield, 1955; Popescul et al., 2000). Time is considered one of the most important factors in detecting subjects that are obsolete and those that are emerging. Success through time is also a measure used by libraries to rank journals as part of their decision to subscribe or unsubscribe to journals. Authors of scientific papers decide where to publish their papers based on the current popularity of a journal and the recency and importance of citations made to that journal.

It has been shown that citations of journal articles behave in a consistent manner. In general, the more time passes, the less citations a paper receives (Egghe, 2001; Garfield, 1998). In fact, a journal will be considered more prominent the higher its citation half-life is (i.e., how old in years most of the papers currently cited are in the literature that were previously published in this journal). Combined with another measure called *impact factor* (the frequency with which the average article in a given journal has been cited in a particular year), libraries determine the value of a certain journal to their collection. Because the value of journals can change over time, this evaluation is carried out in many libraries on an annual or biannual basis. Furthermore, authors learn about the importance of their acceptance to a journal or the citation of their work in a certain journal based on such evaluations (Garfield, 1955, 1970, 1972).

Unlike the publications studied in co-citation analysis, pages on the Web are modified and updated with respect to real world events. For example, when a government changes in a certain country, the URL of the official government site remains the same, but its content is changed to fit the policies of the new government. This is different from the documents published in hard-copy that become obsolete, or *stale,* and are then replaced by new, *fresh* documents. Consequently, Web links exhibit a behavior that is the opposite of that found in scientific citations: the more time passes the more citations (links) a page receives (Adamic & Huberman, 2001).

Using in-degree related metrics to measure or indicate popularity and authoritativeness is a well-accepted practice on the Web, and most current Web search engines apply some form of link-analysis in their ranking process. However, both mentioned attributes of Web pages are time-dependent. Consider, for example, popularity. While there are some classic, timeless popular icons, in general, popularity is dictated by conventions, trends, and fashion that change over time. In order to measure and repesent the temporal nature of popularity on the Web, we believe that there is a need to find a way to couple hyperlinks with temporal data.

As users, we expect a ranked list of pages about a concept or theme that was displayed a year ago to be different from a ranked list about this same theme displayed today. For example, a query like "The World Trade Center" yielded completely different results before and after the terror attack on September 11, 2001. Before that date, most of the top 100 results from the major search engines pertained to the thriving business and tourism district near and within the WTC buildings. In the year following the attack, most of the top 100 documents returned relate to the event itself, from official information to personal stories and calls for aid. Recently, the same query returns documents that relate to the restoration of the site (referred to since as "Ground Zero"). This change in the composition of the top Web resources described by the term "World Trade Center" (as returned by the

Accepted January 23, 2004

 ● Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20082

search engines AltaVista and Google) reflects a shift in the meaning of the concept worldwide.

Not all concepts shift so radically, and most of the events in the world do not attract the media's attention as this one did. However, change and trends are part of our everyday life, and we claim that by tracing the dates of creation and updates of documents and hyperlinks on the Web, we gain valuable information about temporal changes, events and happenings in the real world.

In this paper, we discuss several aspects and uses of temporal data in the context of Web information retrieval (Web IR). We envision and describe the ideal case—where search engines track and store temporal data for each of the pages in their repository. We define the *timestamped link* as a method to represent the time dimension and associate it with hyperlinks. We introduce several applications into which such temporal data can be incorporated. These applications enable us to (1) measure the activity within a topical community as a function of time, and (2) adapt standard link-based ranking schemes to capture *timely authorities,* the authorities that are on the rise today and should be ranked over the resources of days past. To demonstrate our claims, we use a somewhat simplistic approach which approximates the age of the page's content by obtaining its "last modified" HTTP header field. We show that by using this simple measure alone it is possible to detect and expose significant events and trends. We predict that by using more robust methods for detecting changes to each page's content, search engines will be able to provide results that are more timely and better reflect current real-life trends than search results provided today.

The rest of this paper is organized as follows. In the next section we survey previous work related to temporal Web data. Then we go on to describe our vision of an ideal timestamped Web representation. We also present the methods we employed to approximate these timestamps for associating time with links. We proceed to define the *Timestamped Link Profile* (*TLP*) of a theme and show how to use this profile to measure changes in a community over time. We additionally describe an adaptation of link-based ranking schemes that captures timely authorities. We conclude by presenting future work direction and a short conclusion.

Related Work

Several studies have tracked the general process of page updates on the Web. Kumar et al. (1999) reported that the mean time between updates is six months, while Cho & Garcia-Molina (2000) estimated this figure at no more than four months. Recently, Fetterly et al. (2003) conducted a large-scale study in which 150 million Web pages were recrawled during 11 consecutive weeks. The authors correlated the varying availability and rate of change of these pages to multiple factors such as page length and top-level domain to which the pages belong.

Moving beyond tracking changes to individual pages, there have been studies of the evolution of Web communities. The experiments described by Kumar et al. (1999) were performed over a period of time to measure the change between two or several samples of Web data. These measurements showed that Web communities change over time. The authors identify small complete bipartite graphs that they call bipartite cores, and predict that many of these cores will evolve over time to form the centers of full-fledged Web communities. Toyoda & Kitsuregawa (2003) identify several changes that Web communities may undergo as they evolve. They define several metrics of community growth, emergence, and stability, and conduct measurements on communities extracted from four Japanese Web archives that were crawled from 1999 to 2002.

An interesting novel analysis of temporal data was presented by Kleinberg (2000) in the context of tracking e-mail streams. Kleinberg assumed a model in which e-mail discussions concerning a particular topic (an e-mail thread) alternate between a state of rapid exchanges (a burst of activity) and a state of slow exchanges. Within this model, he devised an algorithm that identifies the periods of heightened activity and can also identify a hierarchy of bursty activity. Kumar et al. (2003) applied Kleinberg's burst detection analysis to the changes in the pattern of inter-linkage between blogs (a term coined for "personal Web logs") over time. They were able identify a transitional behavior in blogs that occurred in late 2001. While both Kleinberg and Kumar et al. applied rigorous analysis to sequences of temporal data, our work is more anecdotal. We first aim to establish a correlation between real-life events and detectable bursts in temporal data related to hyperlinks. While e-mails, for example, directly discuss real-life events, hyperlinks are twice removed from reality: They merely link to the resource discussing the event or concept. Furthermore, note that both e-mail exchanges and links appearing in blogs usually have very reliable and easily extractable timestamps attached to them; In contrast, it is unclear how to best associate time with hyperlinks (an issue addressed in the next section).

Changes in the interests and activities of the Internet masses are tracked in an attempt to observe, and sometimes predict, trends. For example, variation in the traffic to a set of Web pages was examined by Chi et al. (1998). The authors sampled the traffic by examining the log files of Web servers over a period of time. These measurements modeled the growth of the traffic to certain pages in the set in response to changes made to the set itself. "Buzz" changes, as reflected in the query streams submitted to search engines, are released (to some extent) by several of the engines[1].

Timestamping the Web

The observations presented in the previous sections indicate that it would be beneficial to associate time with links and to develop methodologies that make use of this new dimension in the context of Web search. Ideally, this involves

[1]Google: http://www.google.com/press/zeitgeist.html; Ask Jeeves: http://static.wc.ask.com/docs/about/jeevesiq.html).

having a complete revision history for Web pages: their time of creation, times of subsequent updates, and time of deletion (if the URL of the page is no longer accessible). Furthermore, the granularity of this data should allow us to infer such temporal information as the time of creation, update, or removal of every hyperlink on the page.

One way for a search engine to track temporal changes on the Web is by storing and updating temporal data for each page at crawl time. The temporal data required for an ideal coverage of the Web's time dimension and for implementing the ideas presented in this paper includes the following:

1. The dates when every page was created and last modified. Every time a page is crawled, the crawler checks its HTTP "last modified" header field (see later). If this information is not available, but the engine's repository detects that the page has changed since the last time it was crawled (e.g., by the methods presented by Broder et al., 1997), the page's date of last modification is set to the date of the crawl. In particular, this procedure updates the page's date of creation when the page is crawled for the first time.
2. The date when a page was detected as deleted. This date is set, for example, when receiving 404 codes for previously seen pages, or when a page cannot be accessed for long periods of time.
3. Dates of creation and deletion of links. In the ideal implementation, the search engine should track the additions and removals of hyperlinks in each page (e.g. using the tools described by Douglis & Ball, 1996), and tag the links with creation and deletion dates in a similar manner to that described above for the pages.

In a recent effort, Kraft and Hastor (2003) have embarked on a project that addresses many aspects of the above vision. They plan to process multiple snapshots of the Web (available through the Internet Archive[2]) to create an infrastructure for studying the evolution of links over time. The proposed infrastructure makes available first-seen and last-seen dates for every Web link, as well as a query language for accessing the time-annotated data.

While conducting this research, we did not have at our disposal any special infrastructure for tracking temporal changes to Web pages. Therefore, the manner in which we traced temporal data in this paper is a simple approximation of the mechanisms envisioned previously in this article. We relied on the "last modified" field that is sent in the HTTP header of every HTTP GET request. This information, according to the W3C's definition, is not always available nor is it reliable (W3C, 2003). Some Web servers return no data in this field for the pages they serve. Furthermore, servers of dynamic content (such as servers of news sites) often return the crawl date in the "last modified" field, regardless of the date in which the content served was actually last modified. Thus, throughout this research, we relied on the "last modified" values only when the returned date differs from the date of the crawl. On average, we found that about 40% of pages have a date associated with them, and that, depending on the percentage of news-related links, most are usable and reliable as demonstrated below.

The basic unit of temporal data that drives the applications proposed in the following section is the *timestamped link*. A timestamped link to a page p is an ordered pair (u,t), where u is an URL of a page that was last modified at time t and that links to p. We will usually be interested in timestamped links to a *theme,* where a theme will be identified as a set of pages. Formally, if P is a set of pages on a certain theme or topic, a timestamped link to P is an ordered pair (u,t) where u is a URL of a page, last updated at time t, which links to a page in P.

As Web users, we know that there is a natural process that prompts people to modify their pages when there is an event that requires new additions to (or deletions from) pages already written. There are also fossilized pages, pages that were forgotten or neglected by their authors, whose content seems frozen in time. However, when a page does undergo changes, we assume that any hyperlink found on the page (postchanges) has undergone review. We thus view each modification of the page u as a reendorsement, by its author, of the links remaining in u. The timestamped link reflects the most recent such reendorsement (recent relative to the date of acquiring the information from the HTTP header).

What Makes This Timestamp Unique?

To fully understand the rationale behind the timestamped link, it is necessary to contrast the information it represents with that represented by other types of temporal data, such as fluctuations in traffic to Web pages and changes in search engine query streams. To this end, we classify the population related to a theme into three categories: *leaders, active followers,* and *occasional passersby*. The *leaders* of a theme are those who author the authoritative Web sites pertaining to the topic, which are usually returned in the first page of search results by search engines. The *active followers* are entities (often private people) who track the theme to some extent, and who are motivated enough to actively update their own Web pages when theme-related events unfold. They do not set the agenda of the theme and thus will usually not be returned by search engines in response to theme-related queries[3]. The *occasional passersby* are people who participate in the buzz surrounding a topic: They flock to the search engines with theme-related queries, continue to visit the leading Web sites, and move on to the next fad. They do not have a persistent interest in the theme, or at least do not express such an interest by actively publishing online content.

Monitoring traffic to the theme's leading online resources, as well as tracking the rates of theme-related queries that are submitted to general search engines, primarily measures the activities of the passer-bys. Furthermore, both types of

[2]http://www.archive.org

[3]Note that active followers are different from hubs in the sense defined by Kleinberg (1999): They do not necessarily link to many of the leaders, and are certainly not dedicated to maintaining updated lists of links to leading resources. Good hub pages may, in fact, be leaders under our definition.

actions require private data (traffic and query logs) that are not readily available to the public. Activity levels of the leaders themselves, as measured by tracking changes to the content which they create, are not very informative. Leading Web sites are usually updated continuously, regardless of any major changes related to the theme in question. In contrast, the timestamped link measures the activity of the theme's followers and relies on publicly available data (through the use of search engines). We note that query logs of theme-specific vertical search engines might also reflect activity levels of the theme's followers, because those engines usually cater to theme-savvy crowds. However, as mentioned previously, query logs of such engines are usually not available.

The Timestamped Link Profile

We propose a tool that can be used to measure the activity within a topical community as a function of time. A Timestamped Link Profile (TLP) of a theme is the normalized projection of that theme's timestamped links onto the time axis. In other words, the TLP is a histogram of the age of timestamped links—it measures the relative number of timestamped links that are associated with every time interval. It can also be interpreted as the temporal distribution of a single link to an authority on the theme. Accordingly, timestamped link counts in a TLP are L_1-normalized so that the sum of all counts equals 1.

Technically, a TLP for a theme is assembled by submitting a query (or several queries) describing the theme to a search engine (or several engines). We then take the union of the top-*n* results returned by the engine(s) for the query(ies) and denote the resulting set of pages by *P*. Next, we ask for pages that link to each of the pages in *P*, taking the "last modified" value of those pages. Let *Q* denote the pages linking to the pages of *P*, and let *L* denote the multiset of timestamped links from the pages of *Q* to the pages of *P* (*L* is a multiset because each page in *Q* may link to multiple pages of *P*). We also decide on a date range of interest and on the number of (equal-sized) intervals by which the date range should be partitioned. The TLP is then plotted by associating each timestamped link of the multiset *L* with the interval that contains the last update time of the page containing L. Note that some links may fall outside the required date range, and so they do not participate in the TLP. Finally, we normalize the TLP by dividing the number of timestamped links associated with each interval by the total number of timestamped links (associated with any interval).

There are, however, several caveats to this approach. First, we cannot collect all pages that link to a specific page of *P*. We can only access several hundred such links for each page by querying search engines. Second, not all of the URLs we collect have valid or usable "last modified" values, as explained earlier in this section. This implies that our TLP is based on a sample of the links to the theme's leading resources.

Figure 1 shows the aggregate TLP of about 40 unrelated themes collected in late 2001 and mid-2003. The total number of timestamped links for those themes is approximately 5,000 and 8,000, respectively (different themes are used for the two collections). The similarities between the two plots are quite clear. Because the aggregated themes in each figure are unrelated, we can consider their individual TLPs to be independent, and so aggregating independent temporal distributions results in an average, or typical, distribution. We will therefore refer to the plots shown in Figure 1 as the *baseline* TLP, and will compare TLPs of individual themes to the baseline TLP to identify abnormalities.

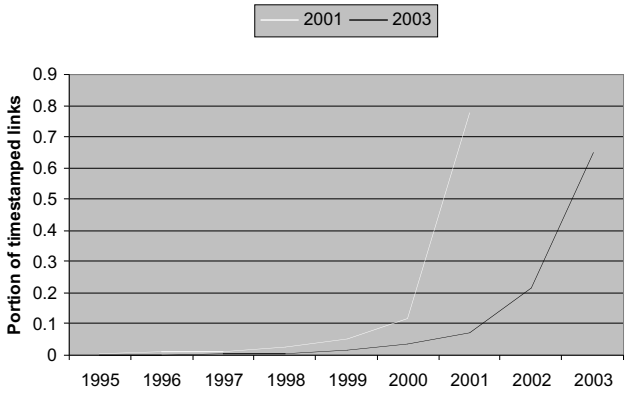

FIG. 1. The "baseline" TLPs in two different points in time, a result of aggregating TLPs of ~20 unrelated topics comprising of ~5,000 timestamped links collected in late 2001, and aggregating TLPs of ~20 unrelated topics comprising ~8,000 timestamped links collected in mid-2003.

An immediate observation is that most of the timestamped links are fairly recent; i.e., most pages that link to the resources of the themes have been modified within a year of the time we collected the links, and their numbers drop sharply the farther we go back in time. This is not surprising: As noted in the introduction, the average Web page is updated several times a year, and new pages are constantly being added to the ever expanding Web. In the next section, we show several instances where significant deviation from this typical distribution is indicative of some real-life singularity concerning the theme or concept in question.

As with any histogram of data, TLPs become meaningful when the total number of timestamped links is at least an order of magnitude larger than the number of time intervals in the profile. Therefore, a theme must either have many authoritative pages with several incoming links or few authoritative pages with many incoming links, in order for its TLP to significantly reflect real-life trends or events.

Applications Using Timestamped Links

This section highlights several potential applications of timestamped links and TLPs. We begin by exemplifying how singular real-life events related to a theme can "scar" that theme's TLP, leaving marks that deviate considerably from the baseline TLP. Then we compare TLPs of related themes, e.g. homepages of companies/products offering competing services and pages related to the same term. By

examining the differences between their TLPs, one can deduce comparative knowledge on which entities are gaining popularity and interest, and which are losing ground. We then proceed to examine the TLP of a specific theme at two different points in time, rather than examining TLPs of related themes at the same point in time. Continuing the same line of thought, we then introduce the *dynamic* TLP, a technique by which TLPs are collected on a regular basis, aggregated visually and analyzed in order to trace interest levels surrounding a theme over time. Finally, we incorporate timestamped links within existing link-based topic distillation algorithms, in an attempt to identify *timely authorities*—authorities that are on the rise today, as opposed to prominent authorities of the past.

The applications we foresee display timely results by way of graphs and ranked lists. This approach is motivated by Boyd (1999), who surveyed the way people perceive time series for analyzing large amounts of data. She concluded that there are three main reasons why a digested form via a graph or a ranked list may be easier to view such information: (1) The time series of the scale found online is usually too large in size for humans to grasp it. (2) Display is usually constrained by space, and therefore it is more useful to have a graph or ranked list of items from which people can extract information at a glance. (3) People usually prefer a digested form of time series rather than just having the numbers.

Abnormal TLPs and Their Relation to Real-Life Events

A TLP of a specific theme can be used to discover abnormal changes in the activities within the theme's virtual community. Major deviations from the baseline TLP can provide hints to major events relating to the theme. Figure 2 shows the TLP of the theme "Year 2000 Bug" (Y2K). The TLP clearly shows abnormal activity in 1999, the year that generated most hype and malevolent prophecies about the effect of the Y2K bug on the human race.

Almost a third of the approximately 750 timestamped links we sampled remained frozen in time, somewhere in 1999. Four years later, one out of every three pages that describe and trace the Y2K bug events has remained untouched and probably forgotten and neglected since 1999. Such fossilization of public interest as reflected on the Web is the underlying basis and strength of timestamping links.

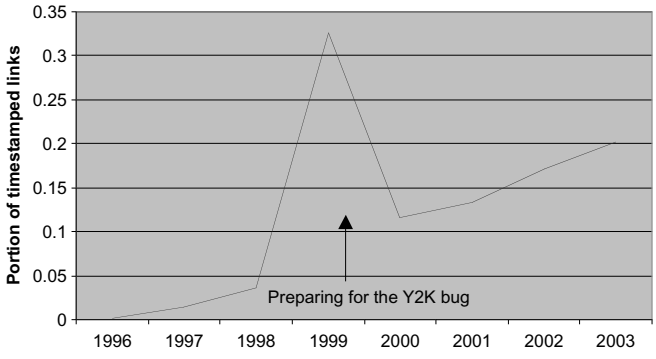

FIG. 2. TLP for the theme "Year 2000 Bug" (~750 timestamped links).

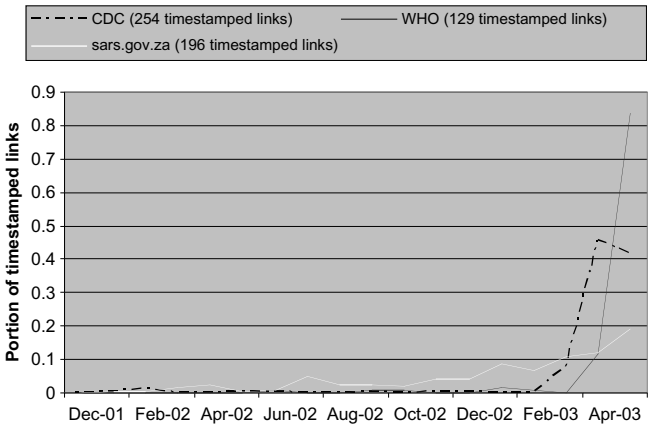

FIG. 3. Comparison between the CDC and WHO SARS sites and sars.gov.za.

Comparing Activity Levels in Communities of Related Themes

Another use of the TLP is to compare changes in the activity levels in communities of related themes (related either semantically or syntactically). The term *SARS,* for example, relates to many themes. It is the acronym for the South African Revenue Service, as well as the acronym for Samples of Anonyms Records from the UK Census of the Office for National Statistics, and more recently, since March 2003, it is also the acronym for the ill-reputed Severe Acute Respiratory Syndrome epidemic.

Because submitting the query "SARS" to a search engine can bring any of the above answers as a perfectly good and relevant answer, we decided to compare the TLPs of three sites, all named SARS: The site of the South African Revenue Service[4], the US Center for Disease Control (CDC) site about the Severe Acute Respiratory Syndrome[5], and the World Health Organization (WHO) site about the Severe Acute Respiratory Syndrome[6].

As shown in Figure 3, the buzz created by the very recent events is easily detected by the TLP. While the South African Revenue Service is an official and established institution with many links and distinguished references, the very young disease name is by far more interesting within the general Web community. Both the CDC and WHO sites attract tremendous public interest by having new timestamped links pointing to them daily (by way of either updating links or creating new links to either or both sites). Such intense interest will probably subside as the disease is contained, and linking pages will eventually fossilize once it is eradicated.

[4]South Africa's SARS office: http://www.sars.gov.za
[5]CDC site: www.cdc.gov/ncidod/sars/
[6]WHO site: http://www.who.int/csr/sars/en/

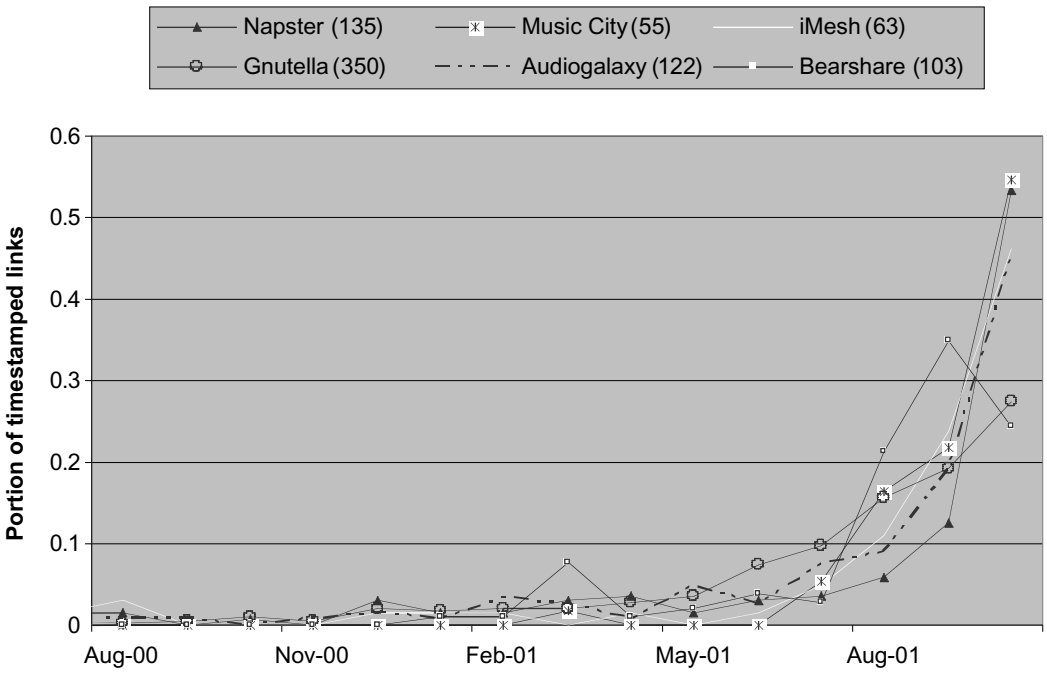

FIG. 4. TLPs of several file-sharing applications.

Figure 4 shows another example for comparing several related themes. The graphs shown are TLPs of Web sites of six popular file-sharing applications collected in October 2001. The graphs clearly demonstrate that five of the file-sharing applications' sites exhibit an ordinary TLP, while bearshare.com exhibits an abnormal TLP: the fraction of its timestamped links in the late months of 2001 drops with respect to the levels of August 2001. This could be caused by the fact that the BearShare client of Gnutella created a lot of hype around the beginning of 2001, but it was reported by its users in early and mid-2001[7] to be an unstable application carrying SpyWare[8]. The reported developments may have slowed the growth of the BearShare community and caused many of its users to stop updating BearShare links on their pages. Since then, BearShare has exhibited normal temporal patterns and has returned to attract the public interest as demonstrated by the other file-sharing applications.

Comparing a Theme to Itself in Two Different Points in Time

Figure 5 plots two TLPs for the theme "Nicole Kidman"—one taken in late 2001 and the other in mid-2003. Both TLPs are overlaid on the same graph, using a time resolution of months counted back from the day of the crawl. Because the actress Nicole Kidman has only recently won several awards for her role in the movie *The Hours,* including an Academy Award, it is not surprising that her virtual Web persona attracts more attention today than it did in 2001.

The strength of the simple TLP approach is that anyone can track his or her temporal reputation using this method without requiring heavyweight processing, logging, or paying a search engine to reveal users' behavior. For example, a business (with a Web site) that has launched an ad campaign can track the level of new timestamped links daily and determine the effectiveness of its advertisement. When combined with other sources of information such as traffic and query log, TLPs offer a more focused view on the community of content makers and information providers on the Web.

Tracing Themes Over Time

A TLP measures the most recent time where followers of a theme edited their Web pages. It can thus reveal recent significant events related to the theme, events that prompted many editorial activities. The ability of the TLP to detect the most recent activity of each theme follower becomes a limitation when tracing themes and communities over time. Over the course of time, themes and communities experience many significant events, which in turn cause many peaks of editorial activity levels. Each such peak hides (or visually *flattens*) previous activities, as our current analysis is restricted to the last modification date of each link-containing page. Consequently, TLPs cannot look back in time beyond the most recent peak. Therefore, when tracing themes over time, TLPs should be collected periodically, with vibrant themes demanding frequent mining of TLPs.

This section describes a technique, called *dynamic* TLP, in which we overlay a series of disjoint TLPs of the same

[7]http://www.securiteam.com/windowsntfocus/5SP0P2K40U.html

[8]http://news.cnet.com/news/0–1005–200–5921593.html, http://www.poppies.org/forum/DCForumID25/21.html

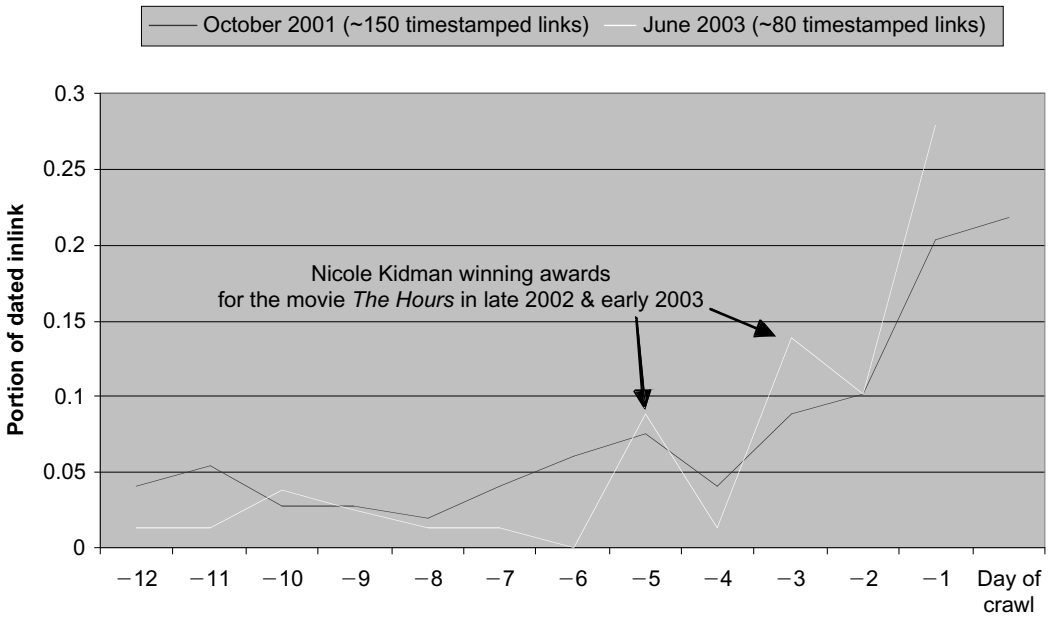

FIG. 5. Comparison of the Nicole Kidman buzz in late 2001 and mid-2003 (looking back one year).

theme onto a single plot. By *disjoint,* we mean that each timestamped link was plotted only once, even if it was encountered multiple times. Note however, that because timestamped links are ordered pairs of the form (*url,time*), a certain URL may participate in many of the plotted timestamped links if we encounter many revisions of it while collecting the data. To summarize, a timestamped link (*u,t*) will be plotted in one of the following two cases:

1. The URL *u* was not part of any timestamped link previously encountered. Such cases obviously occur while collecting the first TLP of the series, but they may also occur while collecting subsequent TLPs: Recall that we rely on search engines to collect the timestamped links. Over time, the engines change both the top-*n* results for our queries and also the URLs they return as linking to the results.
2. The URL *u* has already participated in timestamped links encountered earlier, but it has since been modified. Therefore, the timestamped link it forms now differs from the ones previously encountered.

Figure 6 displays an overlaid series of disjoint TLPs for the theme "Bin Laden," collected in daily increments between September 12, 2001, and October 11, 2001. It traces the TLPs for the top 10 results returned for the queries "Ussama/Usama/Ossama/Osama Bin Laden" for the period sampled.

Overlaying the TLPs allowed us to detect the surge in activity that the September 11, 2001, events created on the Web. Had we simply plotted the TLP in October, the activity levels of September and October would have been more balanced. This follows from the fact that many of the pages that pointed to the theme and that were updated in September (and were thus plotted in our series) were subsequently updated in October 2001 as the events in the United States and Afghanistan developed. Had we plotted just the last TLP of the series, their September activity would have been flattened and suppressed by the more recent October 2001 activity.

Figure 7 shows a (simple) TLP taken on March 11, 2002, exactly six months after the attack. The graph demonstrates well the phenomenon we tried to avoid by introducing the dynamic TLP. The editing activity produced in October 2001 and the following events that were related directly to the "Bin Laden" theme overshadowed the severity of the September 11 events as these are reflected in a simple TLP.

Figure 8 shows a subset of Figure 6, tracing only timestamped links to the URL of the PBS *FrontLine* documentary about Bin Laden released in April 1999. The same pattern is exhibited: A sudden surge in interest appears in September 2001, followed by a decrease in October.

In Figure 9 we show a TLP taken in early June 2003 the same PBS *FrontLine* site. In this plot it is obvious that some-

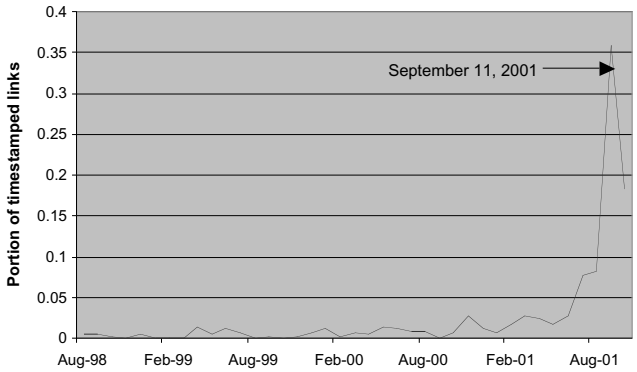

FIG. 6. A dynamic TLP of the theme "Osama/Ossama/Ussama/Usama Bin Laden" (~450 timestamped links), taken in daily increments over a period of one month (September 12, 2001 to October 11, 2001).

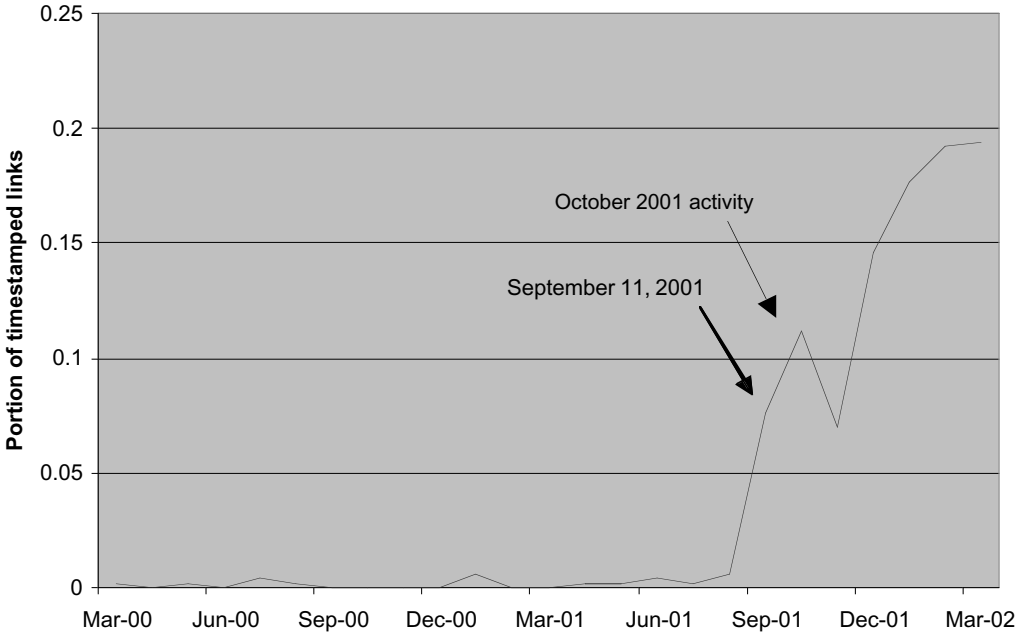

FIG. 7. Six months after September 11, 2001, on March 11, 2002, we again collected a TLP of the theme "Bin Laden" (~500 timestamped links).

thing happened in late 2001 that relates to the theme "Bin Laden"; however, there are two other "probable events" that the graph traces (one in late 2002 and another in early 2003). To test the probable relation between real-world events and their manifestation in the TLP plot, we traced the behavior of the TLP to a related site, the BBC online overview of Bin Laden. This site is one of the most popular sources of information about Bin Laden online. This TLP exhibits a very similar pattern: a rise in interest in September 2001, followed by another short rise in late 2002, and a very recent rise and fall in early 2003. Juxtaposing the two TLPs along the same timeline, the events picked up by the two different TLPs are now made obvious. The first rise in interest in the theme "Bin Laden" occurred during September and October 2001. Then, a year later, the community commemorated the 2001 events and brought them up again. The last and most recent event detected is probably the War in Iraq, which relates to the global war on terror and the hunt for Bin Laden.

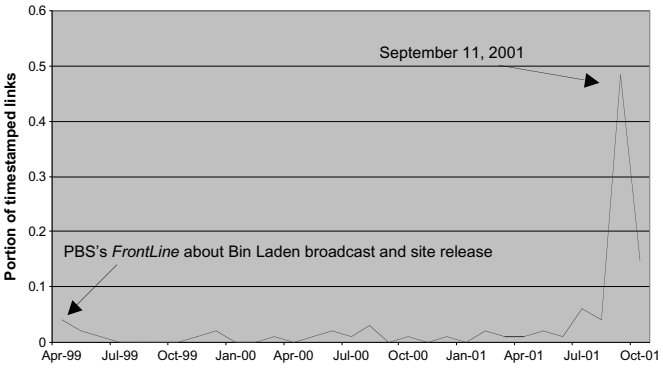

FIG. 8. Dynamic TLP to the site of a PBS program about Bin Laden aired in April 1999 (http://www.pbs.org/wgbh/pages/frontline/shows/binladen). The accumulation was taken in daily increments over a period of one month (September 12, 2001 to October 11, 2001) (~100 timestamped links).

Note that the gravity of the impact of the September 2001 events is weakened by this graph, while the more recent events seem to show evidence of much more impact on the world. This is so because the simple TLP only shows the most recent change. Had we maintained a crawler for sampling timestamped links from the time of the September 11 2001, attack, the ratio between recent interest and historical interest might have been preserved.

From Authorities to Timely Authorities

Search engines usually attempt to retrieve the resources that are currently considered the most authoritative for the submitted queries. However, as far as we know, the commercially available search engines make no special effort to distinguish today's authoritative pages from yesterday's authorities. In particular, previously published link-based ranking schemes do not attribute extra weights to links in recently updated pages: Links originating in pages that were last updated in 1996 are deemed as significant as links originating in fresh pages. Among our contributions is an adaptation of link-based ranking schemes that captures *timely authorities,* the authorities that are on the rise today and should be ranked over the resources of days past.

For each query, we start by assembling a subgraph of the Web as described by Kleinberg (1999). The system first submits the query to a set of search engines (in our experiments, queries were submitted to Google and AltaVista). The union of the top-*X* results from all engines constitute the root set

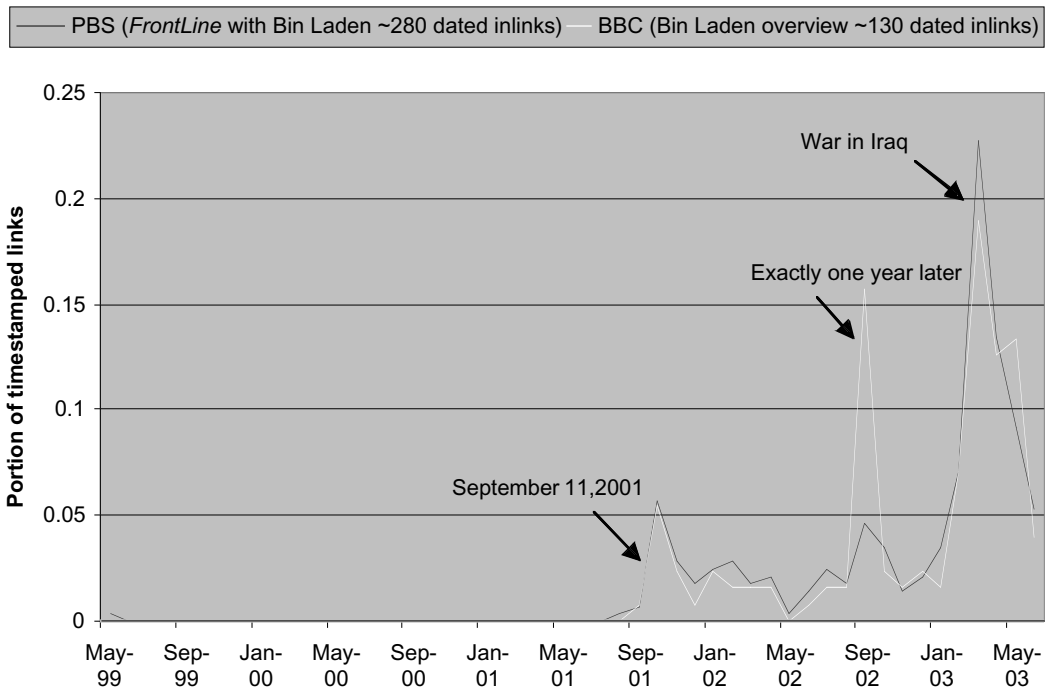

FIG. 9. A juxtaposition of a TLP for the PBS program *FrontLine* about Bin Laden (~280 timestamped links) and TLP for a BBC overview site about Bin Laden (~280 timestamped links) both taken in June 2003, demonstrating how events in the real world affect different sites related to the theme "Bin Laden" in a very similar way.

of pages (we used $X = 30$). The collection of root pages is expanded into a larger set of pages S by following the hyperlinks linking to and emanating from the root pages. The resulting set of pages represents a directed subgraph of the World Wide Web (WWW), whose nodes are the pages in S and whose edges are the hyperlinks which interconnect them.

Having assembled the subgraph to be analyzed, we proceed to assign weights to its hyperlinks. The rankings of many link-analyzing algorithms can be biased toward favorable pages by assigning different weights to links in the collection. For example, links that refer to predefined pages of high quality may have their weights boosted (Aridor et al., 2000; Chakrabarti et al., 1999). In query-specific searches, links referring to pages with matching textual content, or links with anchor text that appear in the query, are often awarded high weights (Bharat & Henzinger, 1998; Chakrabarti et al., 1998). Among the algorithms that are affected by such techniques are HITS (Kleinberg, 1999) and SALSA (Lempel & Moran, 2000). Specifically, we assign weights to links based on two parameters:

1. Following Chakrabarti et al. (1998), links are weighted according to the similarity between the anchor text which is associated with them and the query.
2. To add the time dimension to the analysis, we further alter the weights associated with the links by adding a bonus as follows. For each link, let d denote the time difference (in days) between the current date and the timestamp associated with the link. In our implementation we adjusted the weights as follows:
   - If ($d = 0$ or $d \leq 1$ year): bonus $= 0$
   - If (if $d > 0$ and $d < 1$ week): bonus $= 1.5$
   - If ($d \geq 1$ week and $d < 1$ month): bonus $= 1.0$
   - If ($d \geq 1$ month and $d < 6$ months): bonus $= 0.5$
   - If ($d \geq 6$ months and $d < 1$ year): bonus $= 0.25$

   Basically, links from fresh pages (pages that were updated recently) were assigned higher weights than links emanating from stale pages. This is the only algorithmic difference between the calculation of "basic" authorities and the calculation of "timely" authorities.

We then analyzed the link structure of the subgraph S in the manner described by Aridor et al. (2000), assigning each candidate page with a hub score and an authority score. These scores are computed by summing the scores produced by the HITS and SALSA algorithms.

In what follows, we present experiments of our modification for two queries. In each experiment, two lists of 20 authorities are shown. The list on the left was produced without considering temporal data, while the list on the right took into account the temporal data. Furthermore, the parentheses next to every URL on the right list indicate the rank of that URL in the left column (or "new" when the URL did not appear in the left column). Table 1 summarizes the differences between the two ranked lists in every experiment conducted in 2001 by listing the number of authorities that hold top-n positions in both lists.

We note that introducing similar time-based link weights into PageRank (Brin & Page, 1998), while beyond our capabilities, is intriguing and should be pursued by search

TABLE 1. Quantitative differences between lists of authorities and lists of timely authorities.

Query	Intersection@5	Intersection@10	Intersection@15	Intersection@20
File sharing	3	3	9	11
Harry Potter	3	6	11	12

TABLE 2. Comparison between "file sharing" authorities computed without timestamped links and "file sharing" authorities computed with timestamped links.

Basic authorities	Timely authorities
1. www.imesh.com	1. www.imesh.com (1)
2. www.napster.com	2. www.kazaa.com (7)
3. www.filetopia.com	3. www.riffshare.com (4)
4. www.riffshare.com	4. www.filetopia.com (3)
5. www.bearshare.com	5. www.filerogue.com (17)
6. gnutella.wego.com	6. www.filefreedom.com (new)
7. www.kazaa.com	7. www.neo-modus.com (11)
8. www.musiccity.com	8. www.swaptor.com (13)
9. www.filenavigator.com	9. www.songspy.com (15)
10. www.napigator.com	10. www.splooge.com (new)
11. www.neo-modus.com	11. www.musiccity.com (8)
12. www.aimster.com	12. www.audiogalaxy.com (20)
13. www.swaptor.com	13. www.carracho.com (new)
14. www.downloadcommunity.com	14. www.bigredh.com (new)
15. www.songspy.com	15. www.mojonation.net (19)
16. opennap.sourceforge.net	16. www.downloadcommunity.com (14)
17. www.filerogue.com	17. espra.net (new)
18. konspire.sourceforge.net	18. konspire.sourceforge.net
19. www.mojonation.net	19. www.peergenius.com (new)
20. www.audiogalaxy.com	20. www.grokster.com (new)

engines that posses the required Web-wide connectivity information.

Table 2 shows results from 2001 obtained with our analysis for the query "file sharing." As we learned from our previous 2001 experiments, BearShare, which received the fifth place in the basic analysis, lost its position in the top 20 results when temporal data was incorporated into the analysis. Napster was also dropped from the timely authorities list, probably because of the fact that at the time of the experiment in late 2001 it no longer provided a downloadable application, a fact that was well known to the majority of users of file-sharing applications. Many of the new file-sharing applications that entered the timely authorities list were applications that were emerging within the community of heavy users. Because the file-sharing user community was and still is a very active one, its timely authorities probably change within relatively short periods of time.

Table 3 shows results obtained with our analysis for the query "Harry Potter." At the time this experiment was conducted in 2001, the first "Harry Potter" movie was about to be released by the Warner Brothers film production company. This release attracted a lot of attention and buzz in the virtual world. Fan sites were competing to report and display images

TABLE 3. Comparison between "Harry Potter" authorities computed without timestamped links and "Harry Potter" authorities computed with timetamped links.

Basic authorities	Timely authorities
1. www.hpfactsandfun.com	1. www.hpfactsandfun.com (1)
2. www.scholastic.com/harrypotter	2. www.harrypotter.com (12)
3. www.mugglesforharrypotter.org	3. www.mugglesforharrypotter.org (3)
4. www.geocities.com/EnchantedForest/Mountain/5101	4. www.geocities.com/EnchantedForest/Mountain/5101 (4)
5. www.geocities.com/~no-quarter/potter	5. www.hpgalleries.com (new)
6. www.kidsreads.com/harrypotter	6. www.scholastic.com/harrypotter (2)
7. www.harrypotterfans.net	7. www.geocities.com/~no-quarter/potter (5)
8. hosted.ukoln.ac.uk/stories/stories/rowling/potter	8. www.jkrowling.com (new)
9. www.hpnetwork.f2s.com	9. www.hpnetwork.f2s.com (9)
10. www.i2k.com/~svderark/lexicon	10. www.scholastic.com/harrypotter/home.asp (new)
11. www.harrypotter.ws	11. www.the-leaky-cauldron.org (new)
12. www.harrypotter.com	12. www.i2k.com/~svderark/lexicon (10)
13. www.harrypotterrealm.com	13. hosted.ukoln.ac.uk/stories/stories/rowling/potter (8)
14. www.mindspring.com/~gwil/wizwords.html	14. www.kidsreads.com/harrypotter (6)
15. www.whoisharrypotter.com	15. www.harrypotterfans.net (7)
16. www.homestead.com/hogwarts_33/harrypotter.html	16. www.redmailorder.com/potter/portkey (new)
17. www.geocities.com/the_dilapidated_one	17. hpgalleries.community.everyone.net/commun_v3/scripts/directory.pl (new)
18. thedursleys.homestead.com/home.html	18. www.hpgalleries.com/moviegallery3.htm (new)
19. www.harrypottermania.main-page.com	19. www.harrypotter.ws (11)
20. www.harrypotterguide.ic24.net	20. www.homestead.com/harrypotterbymegz/home.html (new)

from the prereleased movie, and to provide background information about the people involved in making that movie. HarryPotter.com is the URL of the movie's official Web site (it then redirected the reader to harrypotter.warnerbros.com). This is probably the reason for the site's jump in the timely authorities list from the 12th place to the second. We then predicted that the site will become the most timely authoritative site for the query "Harry Potter."

Many of the sites that composed the top 20 timely authorities are fan sites with recent and heavy activity. For example, the freshest timely authority to enter the list, hpgaleries.com, provided daily reports on the movie's prerelease reviews (the movie was screened in Britain for several hundred young viewers before the official release). The site also allowed its virtual community of users to send their own views on and expectations from the next two books which were about to be released in early 2002.

The second highest entrance to the timely authorities list, jkrowling.com, is the site of the author's agent, where the reader was able to find an up-to-date list of all the representative of the agent and the current translations of the book all over the world.

Concluding Remarks and Future Work

Time, which has been recognized as an important dimension in the co-citation literature, has not yet been incorporated into the analogous process of link analysis on the Web. In this paper, we introduced several aspects of using the time dimension in the context of Web IR. We have demonstrated our claims through a simple and easily implemented approach, which approximates the age of the page's content. We have suggested more robust procedures for tracking temporal data, suited for search engines, which continuously crawl the Web and maintain a repository of the discovered resources. Search engines may incorporate the collected temporal data into static link-based rankings of pages (such as PageRank; see Brin & Page, 1998).

An issue that should be pursued in the future is the refinement of the differences between the type of information collected (a) by analyzing TLPs, (b) by tracking topical search engine queries, and (c) by tracking traffic to qualitative theme-related Web pages. Another area that merits research is the application of burst analysis algorithms to the data behind the TLP, namely to the stream of timestamped links.

Acknowledgments

We would like to thank Uri Weiss for his help in conducting the experiments in 2001.

References

Adamic, L., & Huberman, B.A. (2001). The Web's hidden order. Communications of the ACM, 44(9), 55–59.

Aridor, Y., Carmel, D., Lempel, R., Soffer, A., & Maarek, Y.S. (2000, July 7–9). Knowledge agents on the Web. Proceedings of the Fourth International Workshop on Cooperative Information Agents CIA 2000, Boston, MA. Also in M. Klush & L. Kerschberg (Eds.), Lecture notes in artificial intelligence 1860 (pp. 15–26). Springer.

Bharat, K., & Henzinger, M. (1998). Improved algorithms for topic distillation in a hyperlinked environment. In W. Bruce Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson, & J. Zobel (Ed.), Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98) (pp. 104–111). Melbourne, Australia: ACM.

Boyd, S. (1999). A signal processing approach to generating natural language reports from time series. Unpublished doctoral dissertation, Macquarie University, NSW, Australia.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual {Web} search engine. WWW7/Computer Networks & ISDN, 30(1–7), 107–117.

Broder, A.Z., Glassman, S.C., Manasse, M.S., & Zweig, G. (1997). Syntactic clustering of the Web. WWW6/Computer Networks & ISDN 29(8–13), 1157–1166.

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S.R., Raghavan, P., et al. (1999). Mining the link structure of the World Wide Web. IEEE Computer, 32(8), 60–67.

Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., & Kleinberg, J.M. (1998). Automatic resource compilation by analyzing hyperlink structure and associated text. WWW7/Computer Networks & ISDN, 30(1–7), 65–74.

Chi, E.H., Pitkow, J., Mackinlay, J., Pirolli, P., Gossweiler, R., & Card, S.K. (1998, April 18–23). Visualizing the evolution of Web ecologies. Proceedings of ACM Conference on Human Factors and Computing Systems, Los Angeles, CA. (CHI '98) (pp. 400–407).

Cho, J., & Garcia-Molina, H. (2000, May 16–18). Synchronizing a database to Improve Freshness. In W. Chen, J.F. Naughton, & P.A. Bernstein (Eds.), Proceedings of 2000 ACM International Conference on Management of Data (SIGMOD) (pp. 117–128), Dallas, TX. ACM 2000, SIGMOD Record, 29(2), June 2000.

Douglis, F., & Ball, T. (1996, January 22–26). Tracking and viewing changes on the Web. Proceedings of USENIX Annual Technical Conference (pp. 165–176), San Diego, CA, USENIX Association.

Egghe, L. (2001). A noninformetric analysis of the relationship between citation age and journal productivity. Journal of the American Society for Information Science and Technology, 52(5), 371–377.

Fetterly, D., Manasse, M., Najork, M., & Wiener, J. (2003, May 20–24). A large-scale study of the evolution of Web pages. Proceedings of the 12th International World Wide Web Conference (WWW2003) (pp. 669–678), Budapest, Hungary, ACM.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. Science, 122(3159), 108–111. Retrieved July 25, 2004, from http://www.garfield.library.upenn.edu/papers/science_v122 (3159)p108y1955.html

Garfield, E. (1970). Citation indexing for studying science. Nature, 227, 669–671.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. Science, 178, 471–479.

Garfield, E. (1998). The use of journal impact factors and citation analysis for evaluation of science. Unpublished presentation at Cell Separation, Hematology and Journal Citation Analysis Mini Symposium in tribute to Arne Bøyum, Rikshospitalet, Oslo, April 17, 1998. Retrieved July 25, 2004, from http://www.garfield.library.upenn.edu/papers/eval_of_science_oslo.html

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5), 604–632.

Kleinberg, J.M. (2002, July 23–26). Bursty and hierarchical structure in streams. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002) (pp. 91–101), Edmonton, Alberta, Canada, ACM.

Kraft, R., Hastor, E., & Stata, R. (2003, June). TimeLinks: Exploring the evolving link structure of the Web. Proceedings of the Second Workshop on Algorithms and Models for the Web-Graph, Budapest, Hungary.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2003, May 20–24). On the Bursty evolution of blogspace. Proceedings of the 12th International

World Wide Web Conference (WWW2003) (pp. 568–576), Budapest, Hungary, ACM.

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the Web for emerging cyber-communities. Proceedings of the Eighth International World Wide Web Conference, Computer Networks & ISDN, 31(11–16), 1481–1493.

Lempel, R., & Moran, S. (2000). The stochastic approach for link-structure analysis (SALSA) and the TKC effect. Proceedings of the Ninth International World Wide Web Conference, Computer Networks & ISDN, 33(1–6), 387–401.

Popescul, A., Flake, G.W., Lawrence, S., Ungar, L.H., & Giles, C.L. (2000, May 22–24). Clustering and identifying temporal trends in document databases. Proceedings of IEEE Advances in Digital Libraries 2000 (ADL 2000) (pp. 173–182), Washington, DC.

Toyoda, M., & Kitsuregawa, M. (2003, August 26–30). Extracting evolution of Web communities from a series of Web archives. Proceedings of the 14th ACM Conference on Hypertext and Hypermedia (HYPERTEXT 2003) (pp. 28–37), Nottingham, UK, ACM.

W3C, Hypertext Transfer Protocol. (2003). HTTP/1.1, Section 14: Header Field Definitions. http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html