

Temporal link analysis

Einat Amitay, IBM Research Lab in Haifa, MATAM, Haifa 31905, ISRAEL

David Carmel, IBM Research Lab in Haifa

Michael Herscovici, IBM Research Lab in Haifa

Ronny Lempel, Computer Science Department, Technion, Israel

Aya Soffer, IBM Research Lab in Haifa

Uri Weiss, IBM Research Lab in Haifa

Abstract

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 IR. We describe the ideal case – where search engines trace and store temporal data for each of the pages in their repository. In order to demonstrate our claims, we use a somewhat simplistic approach, which approximates the age of the page's content. We introduce several applications into which such temporal data can be incorporated. We show that even using this crude measure it is possible to detect and expose significant events and trends. We predict that by using a more robust method for detecting a page's date of last modification, search engines will be able to provide results that are more timely and better reflect current real-life trends than those they provide today

1 Introduction

Time is a very important dimension in the co-citation literature[11][17]. 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 & 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[10]. In fact, a journal will be considered more prominent the higher its citation half-life is (i.e., how old in years are most of the papers currently cited 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. Since the value of journals can change over time, this evaluation is carried out in many libraries on an annual or bi-annual 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[11][12][13].

In contrast, when plotting similar measures for citations on the Web, the reverse behaviour is exhibited: the more time passes the more citations a page receives [1]. Furthermore, 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.

There have been numerous attempts to make use of time to predict trends on the Web. For example, variation in the traffic to a set of Web pages was examined in[8]. The authors sampled the traffic by examining the log files of Web servers over a period of time. These measurements modelled the growth of the traffic to certain pages in the set in response to

changes made to the set itself. The research in [15] describes experiments performed over a period of time to measure the change between two or several samples of Web data. These measurements aimed to show that Web communities changes over time. In yet another set of experiments[3][18], it was shown that search engines return different results over short periods of time.

Although the studies mentioned above examined the time dimension in the context of the Web, their emphasis was is on the detection of the change itself and not on the temporal nature of the data studied. Furthermore, none of these studies looked into how to incorporate time into the processes that are currently used for ranking web pages, computing link-based measures of site popularity, and link analysis in general. In fact, to the best of our knowledge, the Web Information Retrieval community has never proposed such a temporal approach.

In this paper we discuss several aspects and uses of temporal data in the context of Web IR. We envision and describe the ideal case – where search engines trace and store temporal data for each of the pages in their repository. We introduce several applications into which such temporal data can be incorporated. In order 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 even using this crude measure it is possible to detect and expose significant events and trends. We predict that by using a more robust method for detecting a page’s date of last modification, search engines will be able to provide results that are more timely and better reflect current real-life trends than those they provide today.

The main contribution of this work is first and foremost in raising the issue of utilizing the time dimension in the context of link analysis. We define the *dated-inlink* and propose a method to represent the time dimension and associate it with hyperlinks. We describe how to best incorporate this representation into current practices of crawling and searching the Web. We demonstrate the benefits of incorporating this additional dimension in terms of Web IR in two applications. The first application measures the activity within a topical community as a function of time. The second application 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.

The rest of this paper is organized as follows. Section 2 describes how to associate time with links. In Section 3 we describe an ideal time-stamped Web representation, as well as the methods that we employed to approximate these timestamps using currently available information. In Section 4 we define the *Dated Inlink Profile (DIP)* of a concept 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*. Future work and conclusion are presented in Section 5. Related work is discussed as relevant throughout the paper.

2 Associating Time With Links

Using in-degree to measure or indicate popularity is a well-accepted practice and most current Web search engines apply some form of analysis on the in-links in their ranking process. However, popularity is time-dependant. There are some classic, timeless popular icons, but in general, popularity is dictated by conventions, trends, and fashion that change over time. In order to measure and represent the temporal nature of popularity, 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 that was displayed a year ago to be different from a ranked list about this same concept that is displayed today. For example, a concept like “Monica Lewinsky” yielded completely different results during and after Bill

Clinton's presidency. During Clinton's presidency, most of the top 100 results from the major search engines were related to the news item itself and the opinions and buzz it created. After President Clinton left office, most of the top 100 documents returned were (and still are) about the jokes, humour, and folklore the event created within and outside the USA. This change in the composition of the top Web resources on the Lewinsky affair (as returned by the search engines AltaVista, Google and HotBot) reflects a shift in thought in the USA and a shift in the meaning of the concept "Monica Lewinsky".

Not all concepts shift so rapidly, 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 on the Web, we gain valuable information about temporal changes, events and happenings in the real world.

There is evidence that pages are updated and changed for several reasons. Several studies have tracked the general process of page updates on the Web. In [15] it was reported that this process takes on average 6 months, while in [9] it was reported that the average update rate for a page is up to an average of 4 months. We also know as Web users, that there is a natural process of change in the level of interest. Such a change causes 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, and that their content seems frozen in time.

These observations indicate that it would be beneficial to associate time to links and to develop methodologies that make use of this new dimension in the context of Web search. To facilitate the incorporation of the time dimension into our calculation of popularity through time we introduce a new term – *dated-inlink*. A dated-inlink to a page p is an ordered pair (u,t) , where u is a URL that was last modified at time t and which links to p . We will usually be interested in dated-inlinks to a *topic*, where a topic will be identified as a set of pages. Formally, if P is a set of pages on a certain topic, a dated-inlink to P is an ordered pair (u,t) where u is a URL, last updated at time t , which links to a page in P .

There are other temporal aspects of inlinks we might want to capture. For a complete taxonomy of temporal aspects of inlinks refer to Section 3, where we outline the desired features of a database storing dated-inlinks in a Web search engine.

3 Time-stamping the Web

Ideally, we would like to have a complete revision history for Web pages: their time of creation, times of subsequent updates, and time of deletion (if the URL 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 and thus enable a robust and accurate definition of dated-inlinks.

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 below). If this information is not available but the engine's repository detects that the page has changed since the last time it was crawled (for example, by the methods presented in [6]), 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, and tag creation and deletion dates to the links in a similar manner to that described above for the pages.

The manner in which we traced temporal data in this paper is an approximation of the mechanism described above. The temporal data we used was retrieved from the "last modified" field that is sent with the page's HTTP header. This information, according to the W3C's definition, is not always available nor reliable[19]. 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 from the date of crawl. Furthermore, we associate the date of modification of the links on a page with the date of modification of the page itself. This is a gross approximation but we found it suffices for our needs, as reported in the next Section.

4 Examples for the use of dated-inlinks

4.1 The Dated-Inlink Profile

We propose a tool that can be used to measure the activity within a topical community as a function of time. A Dated Inlink Profile (*DIP*) of a concept is the normalized projection of that concept's dated inlinks onto the time axis. In other words, the DIP measures the relative number of dated-inlinks that are associated with every time interval. It can also be interpreted as the temporal distribution of a single inlink over time.

Technically, a DIP for a concept C is assembled by submitting a query describing C to a search engine, which returns a set of n pages, P . We then ask for pages that link to each of the top n results returned by the engine for that query, taking the "last modified" value of those pages. Ideally, these pages and their last modification date form the set of dated inlinks for the concept C . The DIP is then plotted, by projecting the set of dated inlinks onto the time axis with some granularity. 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 inlinks for each page by querying search engines. Second, not all of the inlink URLs we collect have valid or usable "last modified" values, as explained in Section 3. This implies that our DIP is based on a sample of the inlinks to the concept C .

Figure 1 shows the aggregate DIP of approximately 5000 dated inlinks, collected for about 90 different pages gathered with about 20 topic queries. As evident from the DIP, most of the dated inlinks are fairly recent and their number drops sharply the farther we go back in time.

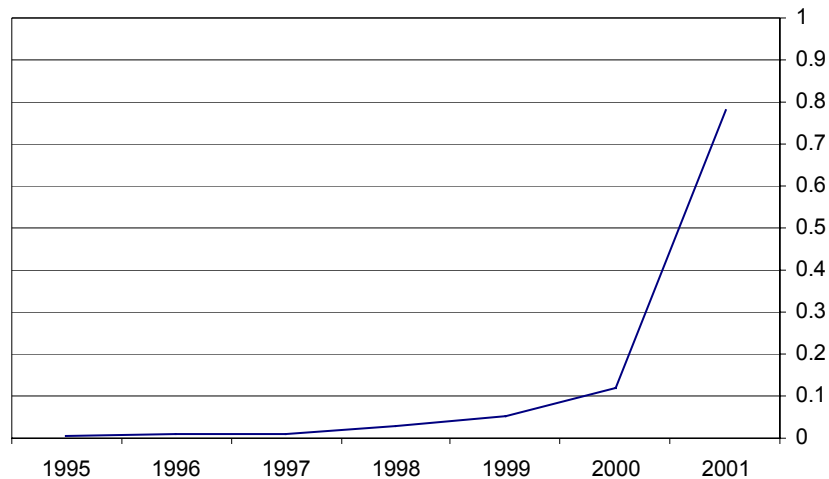


Figure 1 – Aggregate DIP in years of ~5000 dated-inlinks

A DIP of a concept can be used to discover abnormal changes in the activities within the concept’s virtual community. Like an electrocardiogram record of the heart, which aids physicians searching for abnormal peaks, the DIP of a concept can be used to detect abnormal activity patterns relating to the concept. Major deviations from the typical DIP can provide hints to major events relating to the concept. Figure 2 shows the DIP of the concept “Boris Yeltsin”. The DIP clearly shows abnormal activity in 1999, the year of Yeltsin’s resignation.

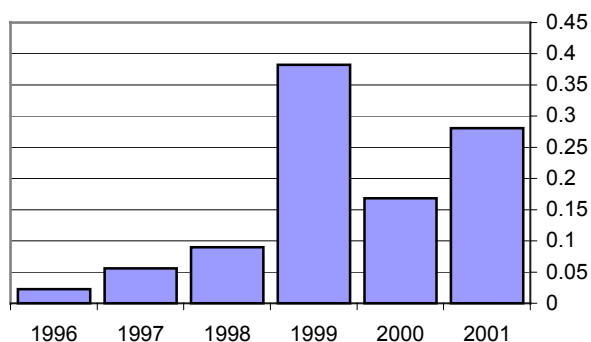


Figure 2 – DIP of the concept “Boris Yeltsin”

We note that there are other methods for tracking the activities that pertain to a certain concept over time. These include monitoring traffic to the concept’s main online resources, and tracking the rates of queries that describe the concept that are submitted to general and concept specific search engines. Both of these methods require private data, while DIPs rely on publicly available data.

4.2 Comparing activity levels in communities of related concepts

Another use of the DIP is in comparing between changes in the activity levels in communities of related concepts. Figure 3 shows the DIPs of Web sites of six popular file-sharing applications. The graph clearly shows that five of the file-sharing applications’ sites exhibit an ordinary DIP, while bearshare.com exhibits an abnormal DIP: the fraction of its dated inlinks in the late months of 2001 drops with respect to the levels of August 2001. This could be due

to the fact that the BearShare client of Gnutella created a lot of hype around the beginning of 2001, but was reported to be an unstable application by its users somewhere in early/mid-2001¹. In September 2001 it had a new version released, but (as indicated in another experiment described later in this paper) this release did not prompt enough activity as to lift BearShare’s DIP in October 2001. In many of the user forums discussing BearShare’s performance, the application is “accused” of using SpyWare², downloading or exporting information about the users without notifying the users first. Such behaviour in the file-sharing user community is not greatly appreciated. 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.

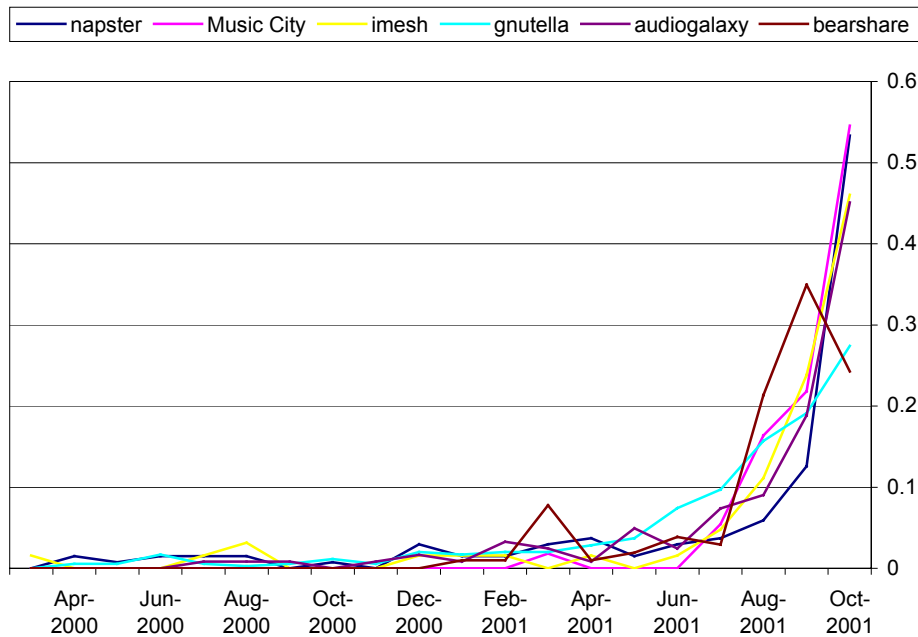


Figure 3 – Comparison of the normalised distribution of dated-inlinks to main Web sites of file sharing applications

4.3 Tracing concepts over time

A limitation of the DIP is its inability to detect abnormal temporal changes in the activity around a concept’s very recent past (e.g. a week before). In order to overcome this limitation we overlaid a series of *disjoint* DIPs of the same concept onto a single plot. By “disjoint” we mean that each dated-inlink was plotted only once, even if it was encountered multiple times. Note however, that since dated-inlinks are ordered pairs of the form $(url, time)$, a certain URL may participate in many of the plotted dated-inlinks if we encounter many revisions of it while collecting the data. To summarize, a dated-inlink (u, t) will be plotted in one of the following two cases:

¹ <http://www.bearshare.com/news.htm>,
<http://download.cnet.com/downloads/0-1896420-601-7217078.html>
² <http://news.cnet.com/news/0-1005-200-5921593.html>,
<http://www.poppies.org/forum/DCForumID25/21.html>

- 1) The URL u was not part of any dated-inlink previously encountered. Such cases obviously occur while collecting the first DIP of the series, but they may also occur while collecting subsequent DIPs: recall that we rely on search engines to collect the dated-inlinks. 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 a dated-inlink encountered earlier, but it has since been modified. Therefore, the dated-inlink it forms now differs from the one previously encountered.

Figure 4 and Figure 5 display a series of disjoint DIPs for the concept “Bin Laden”, collected in daily increments between September 12, 2001 and October 11, 2001. Figure 4 shows the overall picture, tracing the DIPs for the top 10 results returned for the queries Ussama/Usama/Ossama/Osama Bin Laden for the period sampled. Figure 5 shows a subset of Figure 4, tracing only dated-inlinks to the URL of the PBS documentary about Bin Laden released in April 1999.

Overlaying the DIPs allowed us to detect the surge in activity that the September 11th events created on the Web. Had we simply plotted the DIP of October 11th, 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 concept and that were updated in September (and were thus plotted in our series), were subsequently updated in October as the events in the USA and Afghanistan developed. Had we plotted just the last DIP of the series, their September activity would have been “hidden” and replaced by their more recent October activity.

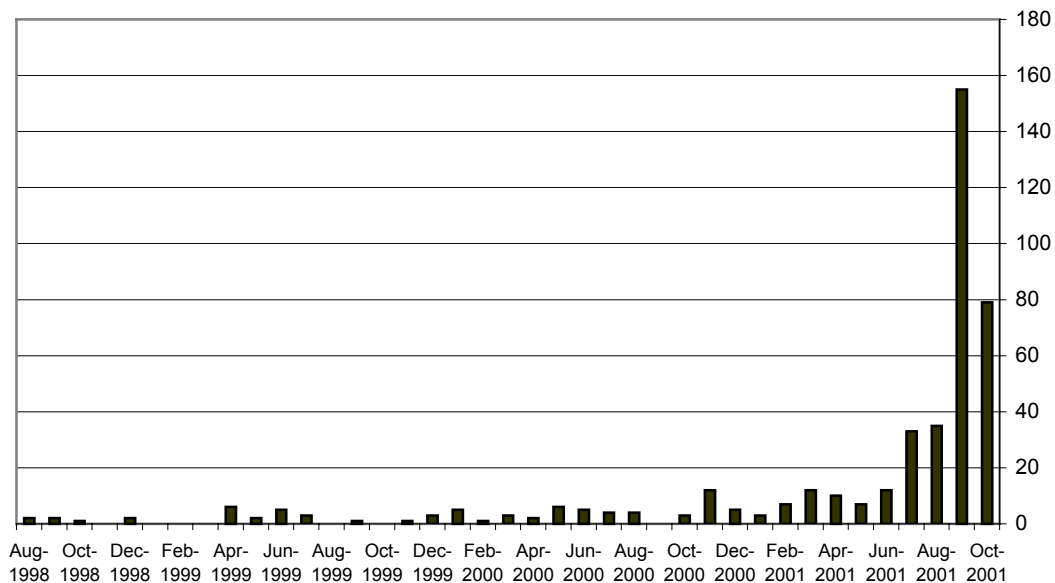


Figure 4 – A combination of a static and a dynamic DIP of the concept “Osama/Ossama/Ussama/Usama Bin Laden”, taken in daily increments over a period of one month (September 12, 2001 – October 11, 2001)

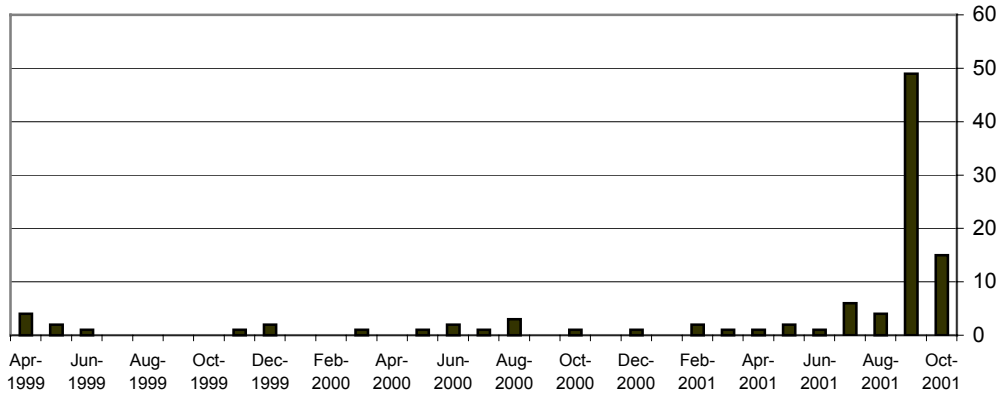


Figure 5 – Accumulative DIP to the site of a PBS program about Bin Laden aired in April 1999 (<http://www.pbs.org/wgbh/pages/frontline/shows/binladen>). The accumulation was taken in daily increments over a period of one month (September 12, 2001 – October 11, 2001)

4.4 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.

In previous work [2], we have used a combination of HITS[14] and SALSA[16] for finding the most authoritative Web sites for a given domain. Following [7], links are weighted according to the anchor text that is associated with them and its similarity to the query. In order to add the time dimension to the process of assigning authority scores to Web pages, we modified the manner in which weights are assigned to the links: links from fresh pages (pages that were updated recently) are assigned higher weights than links emanating from stale pages. In what follows, we present experiments of our modification for three 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 by listing the number of authorities that hold top-*n* positions in both lists.

	intersection@5	intersection@10	intersection@15	intersection@20
File sharing	3	3	9	11
Harry potter	3	6	11	12
Xbox	4	7	10	13

Table 1 – Quantitative differences between lists of authorities and lists of timely authorities

We note that introducing similar time-based link weights into PageRank [5], while beyond our capabilities, is intriguing and should be pursued by search engines that possess the required Web-wide connectivity information.

Table 1 shows results obtained with our analysis for the query “file sharing”. As we learned from our previous experiments, Bearshare, which received the 5th 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 due to the fact that it no longer provides a downloadable application, a fact that is well known to the majority of users of file sharing applications. Many of the new file sharing applications that entered the timely authorities list are applications that are emerging within the community of heavy users. Since the file sharing user community is a very active one, its timely authorities will probably change within relatively short periods of time.

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)

Table 2 – Comparison between “file sharing” authorities computed without dated-inlinks and “file sharing” authorities computed with dated-inlinks

Table 3 shows results obtained with our analysis for the query “Harry Potter”. At the time of writing this paper the first “Harry Potter” movie is about to be released by the Warner Brothers film production company. This release is attracting a lot of attention and buzz in the virtual world. Fan sites are competing to report and display images from the pre-released 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 currently redirects 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 2nd. We predict that at least for a while the site will become the most timely authoritative site for the query “Harry Potter”.

Many of the sites that comprise the top 20 timely authorities are fan sites with recent & heavy activity. Hpgalleries.com for example, provides daily reports on the movie’s pre-release reviews (the movie was screened in Britain for several hundred young viewers). The site also allows its virtual community of users to send their own views on and expectations from the next two books which are about to be released in early 2002.

jkrowling.com is the site of the authoress’s agent, where the reader can find an up-to-date list of all the representative of the agent and the current translations of the book all over the world.

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)

Table 3 - Comparison between “Harry Potter” authorities computed without dated-inlinks and “Harry Potter” authorities computed with dated-inlinks

Table 4 shows our results for the query “xBox”. xBox is a Microsoft computer game. In the timely authorities list the young and fresh sites like xboxattic.com dominated the ranking and displaced the official xbox.com site to the 4th place. teamxbox.com lost 4 places in the timely authorities list since it was inactive for a month, displaying a message that it will be back in a week’s time. While writing these lines teamxbox.com returned to normal activity and would probably regain its place the next time we calculate its timely authoritativeness score. Hackers’ sites, like cheatmaster.com, are very popular and timely. Such sites are favoured among xBox users who are looking for ways to work around problems and obstacles in the software they use.

Basic authorities	Timely authorities
1. www.xbox.com	1. www.xboxattic.com (5)
2. www.xboxaddict.com	2. www.xboxfaction.com (3)
3. www.xboxfaction.com	3. www.xboxaddict.com (2)
4. www.teamxbox.com	4. www.xbox.com (1)
5. www.xboxattic.com	5. www.xboxactive.com (6)
6. www.xboxactive.com	6. xbox.ign.com (10)
7. www.xboxfootball.com	7. www.xsnake.com (13)
8. www.xbox.com/sobe	8. www.teamxbox.com (4)
9. www.burstnet.com/ads/ad9224a-map.cgi/ns	9. www.xboxwired.com (18)
10. xbox.ign.com	10. www.globalxbox.com (14)
11. www.planetxbox.com	11. www.cheatmasters.co.uk/xbox/xbox_cheats.html (19)
12. www.123xbox.com	12. www.operationmsxbox.com (10)
13. www.xsnake.com	13. www.xboxgaming.com (new)
14. www.globalxbox.com	14. www.ultimateresourcesite.com/xbox/main.htm (new)
15. www.xboxcenter.com	15. www.xboxvillage.com (new)
16. www.funxbox.com	16. www.123xbox.com (12)
17. www.xbox-domain.com	17. www.absolutexbox.com (new)
18. www.xboxwired.com	18. www.xboxmovies.com (new)
19. http://www.cheatmasters.co.uk/xbox/xbox_heats.html	19. http://www.howstuffworks.com/xbox.htm (new)
20. http://www.operationmsxbox.com	20. http://xbox.gamezone.com (new)

Table 4 - Comparison between “xBox” authorities computed without dated-inlinks and “xBox” authorities computed with dated-inlinks

5 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. This paper introduced several aspects and uses of the time dimension in the context of Web IR. We have demonstrated our claims using 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. The engines may incorporate the collected temporal data into static link-based rankings of URLs (such as PageRank).

An issue that should be pursued in the future is the difference between the type of information collected (a) by analysing DIPs, (b) by tracking topical search engine queries, and (c) by tracking traffic to qualitative concept-related Web page. We remarked on the availability of the resources required for each of these measurement types. However, in addition to the technicalities of each approach, we feel that these approaches mine different aspects of the communities’ activities.

References:

- [1] Adamic L., Huberman B.A. (2001). The Web’s Hidden Order. *Communications of the ACM (CACM)*, 44(9):55-59.
- [2] Aridor Y., Carmel D., Lempel R., Soffer A., Maarek Y.S. (2000). Knowledge Agents on the Web. *CIA 2000*, pp. 15-26.
- [3] Bar-Ilan J. (1999). Search Engine Results over Time - A Case Study on Search Engine Stability. *CyberMetrics: International Journal of Scientometrics, Informetrics and Bibliometrics*, Vol. 2/3 (1998/9). Available online: <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html>
- [4] Bharat K., Mihaila G.A. (2001). When Experts Agree: Using Non-Affiliated Experts to Rank Popular Topics. In *Proceedings of WWW10*, pp. 597-602.

- [5] Brin S. Page L. (1998). The anatomy of a large-scale hypertextual {Web} search engine. WWW7/ Computer Networks & ISDN, 30(1-7):107-117
- [6] 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.
- [7] 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
- [8] Chi E.H., Pitkow J., Mackinlay J., Pirolli P., Gossweiler R, Card S.K. (1998). Visualizing the Evolution of Web Ecologies. In Proceedings of ACM CHI '98, pp. 400-407.
- [9] Cho J., Garcia-Molina H. (2000). Synchronizing a database to Improve Freshness. In Proceedings of 2000 ACM International Conference on Management of Data (SIGMOD), pp. 117-128.
- [10] 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 (JASIST), 52(5):371:377.
- [11] Garfield E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*, 122(3159):108-111. Available online: [http://www.garfield.library.upenn.edu/papers/science_v122\(3159\)p108y1955.html](http://www.garfield.library.upenn.edu/papers/science_v122(3159)p108y1955.html)
- [12] Garfield E. (1970) Citation indexing for studying science. *Nature*, 227:669-671.
- [13] Garfield E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178:471-479.
- [14] Kleinberg J.M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604-632.
- [15] Kumar R., Raghavan P., Rajagopalan S., Tomkins A. (1999). Trawling the web for emerging cyber-communities. WWW8 / Computer Networks & ISDN, 31(11-16):1481-1493.
- [16] Lempel R., Moran S. (2000). The stochastic approach for link-structure analysis (SALSA) and the TKC effect. WWW9 / Computer Networks & ISDN, 33(1-6): 387-401.
- [17] Popescul A, Flake G.W., Lawrence S., Ungar L.H., Giles C.L. (2000). Clustering and Identifying Temporal Trends in Document Databases. In Proceedings of IEEE Advances in Digital Libraries, ADL 2000, pp. 173–182.
- [18] Rousseau R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *CyberMetrics: International Journal of Scientometrics, Informetrics and Bibliometrics*, Vol. 2/3 (1998/9). Available online: <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html>
- [19] W3C, Hypertext Transfer Protocol -- HTTP/1.1, Section 14: Header Field Definitions. <http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html>