A Method for Measuring the Evolution of a Topic on the Web: The Case of "Informetrics"

Judit Bar-Ilan
Department of Information Science, Bar-Ilan University, Ramat Gan, 52900, Israel.
E-mail: barilaj@mail.biu.ac.il

Bluma C. Peritz
The Hebrew University of Jerusalem, Jerusalem, 91904, Israel. E-mail: bluer@cc.huji.ac.il

The universe of information has been enriched by the creation of the World Wide Web, which has become an indispensible source for research. Since this source is growing at an enormous speed, an in-depth look of its performance to create a method for its evaluation has become necessary; however, growth is not the only process that influences the evolution of the Web. During their lifetime, Web pages may change their content and links to/from other Web pages, be duplicated or moved to a different URL, be removed from the Web either temporarily or permanently, and be temporarily inaccessible due to server and/or communication failures. To obtain a better understanding of these processes, we developed a method for tracking topics on the Web for long periods of time, without the need to employ a crawler and relying only on publicly available resources. The multiple data-collection methods used allow us to discover new pages related to the topic, to identify changes to existing pages, and to detect previously existing pages that have been removed or whose content is not relevant anymore to the specified topic. The method is demonstrated through monitoring Web pages that contain the term "informetrics" for a period of 8 years. The data-collection method also allowed us to analyze the dynamic changes in search engine coverage, illustrated here on Google—the search engine used for the longest period of time for data collection in this project.

Introduction

The World Wide Web is continuously growing at an incredible speed both in terms of its content and in terms of the number of users accessing it. The Web has become an indispensable information source. Its growth patterns are of interest for technical, theoretical, social, and economic reasons and are one of the goals of the emerging Web science (Berners-Lee, Hall, Hendler, Shadboldt, & Weitzner, 2006).

Received June 17, 2008; revised March 16, 2009; accepted March 17, 2009

© 2009 ASIS&T • Published online 14 May 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21097

This article introduces a method for studying the evolution of topics on the Web. The procedures involve the combination of two data-collection techniques: retrieving data from search engines and revisiting Web pages identified at previous data-inspection points. The combination of the two techniques allows the study of several evolution patterns: creation of new pages, removal of previously existing pages, and modification of the content and structure of existing pages.

As a specific case, we present the results of a longitudinal study that monitored the growth and changes that occurred to Web pages containing the term "informetrics" for a period of 8 years, between 1998 and 2006. This is the first study that we are aware of that tracks the evolution of a topic on the Web for such a long period of time and uses multiple data-collection methods.

Longitudinal studies that follow the development of a topic on the Web over time provide insights to understanding the changing roles of the Internet in the overall development of a topic, and to the growing importance of the World Wide Web as an information source.

Related Work

Web Growth, Dynamics, and Structure

There is ongoing interest in estimating the size of the Web (e.g., Bharat & Broder, 1998; Gulli & Signorini, 2005; Lawrence & Giles, 1998, 1999). Recent reports on the size of the Web can be found, for example, in Pandia (2007) and in de Kunder (2008). Internet World Stats (2008) publishes statistics about the number of Internet users around the world. As of March 2008, the estimated number of Internet users, based on information from various sources, was more than 1.4 billion—about 21% of the world population.

The structure of the Web can be modeled as a bowtie graph, based on the analysis of large crawls (Broder et al., 2000). A similar structure has emerged for the much smaller Chilean Web (Baeza-Yates, Castillo, & Saint-Jean, 2004).

The Web is considered to be a scale-free network, and its emergence can be explained by preferential attachment (Barabasi, Albert, & Jeong, 2000). The basic model does not take into account page or link deletions. In a slightly different model, Albert and Barabasi (2000) took into account changes to the existing link structure, by what they called "rewiring." Huberman and Adamic (1999) and Fenner, Levene, and Loizou (2005) introduced models that allow for removal of Web pages. Dorogovtsev and Mendes (2000) and Fenner, Levene, and Loizou (2006) proposed models where link deletions are allowed. The aforementioned studies took into account different aspects of the ever-changing Web, but we are not aware of any generative model that incorporates *all* of the dynamic processes that take place on the Web (i.e., appearance, disappearance, modification, and redirection). One possible reason for the lack of such a model is that growth is by far the most significant process, and even without incorporating the additional processes, the existing models provide reasonable explanations for the structure of the Web.

Longitudinal Studies

Previous longitudinal studies have monitored sites and pages for shorter periods of time, usually for several weeks or months (e.g., Adar, Teevan, Dumais, & Elsas, 2009; Cho & Garcia-Molina, 2000; Fetterly, Manasse, Najork, & Wiener, 2004; Kim & Lee, 2005; Ntoulas, Cho, & Olston, 2004; Olston & Pandey, 2008). In these shorter term studies, the datasets were huge, and the monitored pages were visited often (i.e., typically, once a week).

Only a few studies have reported findings based on several years of data collection, but even these are for a shorter length than that of the current study. Koehler (2004) monitored a fixed set of pages for 325 weeks (over 6 years). Gomes and Silva (2006) gathered data on the Portuguese national Web for a period of 3 years (eight data-inspection points), Baeza-Yates and Poblete (2003) based their results on three data-inspection points over a period of 3 years of the Chilean Web, and Toyoda and Kitsuregawa (2006) had access to the Japanese Web archive which collects data about once a year, and based their results on data from 2003 to 2004 (three data-inspection points). Ortega, Aguillo, and Prieto (2006) crawled about 1,000 sites twice, once in 1997 and once in 2004, and Dontcheva, Drucker, Salesin, and Cohen (2007) crawled 100 Web pages every day for 5 months to study structural changes. Additional studies were covered in a survey by Ke, Deng, Ng, and Lee (2006). Payne and Thelwall (2007, 2008) studied the changes that occurred to the "academic" Webs of the United Kingdom, New Zealand, and Australia for a period of 5 years (Each of these Webs were crawled once a year between 2000 and 2005.) The findings showed that the number of static pages and links in each of these academic Webs have stabilized as of 2001. Although overall about two thirds of the links were unchanged, there were large individual differences for universities.

All previous studies that we were able to locate used a single data-collection method. They either monitored a fixed dataset (e.g., Fetterly et al., 2004; Koehler, 2004), crawled in a prespecified manner a fixed number of pages from given starting points (e.g., Cho & Garcia-Molina, 2000; Kim & Lee, 2005), or attempted to download complete Web sites (e.g., Ntoulas et al., 2004; Payne & Thelwall, 2007, 2008) and/or entire national Webs (e.g., Baeza-Yates & Poblete, 2003; Gomes & Silva, 2006; Toyoda & Kitsuregawa, 2006).

Several measures of change over time were proposed in the various studies. Bar-Yossef, Broder, Kumar, and Tomkins (2004) proposed to measure decay, and computed the measure for both currently existing pages and for previous versions of the pages accessed through the Internet Archive. Fetterly et al. (2004) based their similarity measure on "preimages," a modification of the "shingles" introduced by Broder, Glassman, Manasse, and Zweig (1997). Kim and Lee (2005) computed, among other measures, the modification rate of pages. Ntoulas et al. (2004) assessed the degree of change between different versions of the same page based on *td·idf* and word distance. Kwon, Lee, and Kim (2006) suggested measuring change based on edit distance. Toyoda and Kitsuregawa (2006) computed the "novelty measure," which assessed whether the newly identified pages at a given data-inspection point are really "new" or were simply not discovered in the previous crawls. For an extensive bibliography of temporal and evolutionary aspects of the World Wide Web, updated until 2004, the interested reader is referred to Grandi (2004).

Search Engine Dynamics

The previously mentioned studies show that the Web is extremely dynamic. The search engines add another dimension of dynamicity because of frequent changes in their databases, indexing, and ranking policies. Search engine dynamics is covered in the literature review because search engines serve as primary data-collection sources when gathering information on a specific topic. The dynamics of search engines have been observed and recorded in several studies (e.g., Bar-Ilan, 1999, 2000; Mettrop & Nieuwenhuysen, 2001; Risvik & Michelsen, 2002; Rousseau, 1999; Thelwall, 2001). Bar-Ilan (2002, 2004) introduced a set of measures to assess these dynamic changes.

Life Cycle of a Web Page

A Web page goes through different states during its existence. Some of the states are influenced only by its creator(s)/owner(s) (see States S1–S4 and partially S5) while other states are influenced by external factors. We define the following states that are relevant to manually created Web pages (excluding automatically created pages, such as pages that are created "on-the-fly" after submitting a Web form).

S1: The Web page is "born." It is created and uploaded to the Web.

DOI: 10.1002/asi

S2: the content of the Web page is changed; here, we can differentiate between:
  - S2a: The design or the structure of the page is changed (e.g., background color or placement of images).
  - S2b: The content of the Web page is modified; here, we exclude changes to the outgoing links from this page because such changes are characterized by S2c.
  - S2c: The set of outgoing links from the Web page is modified (e.g., new links are added, existing links are removed, or existing links are modified: This could be a correction of a typo, or an update in case the Web page the link points to is moved to a different URL. Links may be removed because the page the link points to ceased to exist or because the content of the page is not considered appropriate or relevant anymore. Note that depending on the design of the Web site, the owner can change the sidebar of all the pages on the site at once, and thus even dated items can have links to recently created pages. This happens quite often to archived blog posts.
  - S2d: The content of the Web page is changed, but not by the creator/owner of the Web page. Such situations are becoming more and more frequent with the advent of Web 2.0, where Web page owners invite other users to comment on the content of their page (e.g., blog postings or comments on news articles) or to provide feedback or rating. There are cases where some of the information that appears on the page is not created by the owner of the page (e.g., advertisements or other information that the owner agreed to display). These items can change without the explicit knowledge of the owner of the Web page.
- S3: The Web page is not updated anymore, but it continues to exist on the Web. Sometimes, the information on the page becomes outdated and the page is "abandoned;" that is, no one maintains it anymore (see Bar-Yossef, Broder, Kumar, & Tomkins, 2004). In other cases, the page might be "perfect," and the owner of the Web page is satisfied with the page.
- S4: The Web page "dies." It is permanently removed from the Web.
- S5: The Web page is moved to a different location. This can be a decision of the owner of the Web page or the owner of the Web site; these two are not necessarily identical. Sometimes, the old URL redirects to the new URL.
- S6: The Web page is temporarily inaccessible. This may be due to server maintenance, or server or communication failures.
- S7: The Web page is indexed by a search engine. Note that this can happen quite a while after the page was created (e.g., Thelwall, 2001).
- S8: The Web page is dropped from the index of the search engine or the search engine fails to retrieve the Web page even though the page is indexed by it (e.g., Bar-Ilan, 1999, 2000, 2002; Mettrop & Nieuwenhuysen, 2001; and the "search engine coverage" section in this article).
- S9: A link to the Web page is created.
- S10: A link to the Web page is removed.

Next, we describe a method to study the evolution of a topic on the Web, without using a crawler. A topic is defined by formulating an appropriate description of the topic. This description can be seen as a query that can be submitted to search engines. At different time points (called "data-inspection points"), Web pages about the given topic are collected with the help of search engines (i.e., Web pages in State S7), and previously located Web pages are revisited to ascertain their current state (which can be States S2–S6 or S8).

The method does not enable us to identify and differentiate between all the states defined earlier. For example, we have no way of locating a newly created page (S1) unless it is indexed by a search engine (S7), and we cannot differentiate between States S4 (page removed permanently) and S6 (page temporarily inaccessible).

If we do not examine the specific types of changes the given page has undergone over time (if any), we are not able to differentiate between a page that has not changed over time (S3) and pages that have undergone structural or content-specific changes (S2). In addition, changes may occur to Web pages, but are unnoticed by the method. Consider, for example, a Web page that was vandalized by hacker but was reverted by the page owner. If this happened between two data-inspection points, then the monitoring process will not record any change. Our inability to detect such changes has to do with the resolution of the monitoring process and has been discussed before (e.g., Brewington & Cybenko, 2000).

Although our methodology outlined next is not able to differentiate between all the aforementioned states, it provides a macroscopic view on the evolution of a set of pages on a given topic. We view this typology as a first step towards the characterization of the temporal state of a Web page.

Methods

Our aim was to study the evolution and development of a specific topic on the Web. The method is comprised of four steps:

Step 1: Defining the Topic

Search engines are utilized for data collection; thus, it is essential to delineate the topic properly with the choice of keywords. Suppose that the chosen topic is "information science." Currently (as of March 9, 2009), Google reports about 12,100,000 results; however, there are additional Web pages on the topic where the phrase "information science" does not appear explicitly, such as the query "information organization"−"information science" (i.e., Web pages with the phrase "information organization," but without the phrase "information science"). We expect the results of this query to be relevant to our topic, but the results do not explicitly contain the phrase "information science." It returns an additional 612,000 documents that are not retrieved by the previous query.

Note that if data are collected through focused crawling (Chakrabarti, van den Berg, & Dom, 1999), then examples of pages relevant to the topic are needed (Depending on the method, both positive and negative examples might be

DOI: 10.1002/asi

needed.) The quality of the focused crawl is dependent on the representativeness of the examples provided.

Step 2: Initial Data Collection

Once the topic is delineated, we can either apply focused crawling or use the major search engines as data-collection tools. Focused crawling requires considerable resources. Here, we concentrate on data gathering through the use of search engines because results retrieved by search engines are publicly available resources.

Even if we are able to define a set of keywords that covers the chosen topic (and the chosen topic only), we are faced with additional problems. We experienced some problems using Google, which is currently the most popular search engine and is one of the search engines with the widest coverage of the Web (Nielsen Online, 2008). Google currently does not allow more than 32 keywords per query. In addition, Google does not allow the submission of complex Boolean queries (e.g., conjunctions of disjunctions)—it does not recognize the use of parentheses in queries. It also is not very good in "search engine math," and seemingly, it provides only partial support for disjunctions. For example, for the query *conjunction*, it reported 70,000,000 results and for *disjunction*, 985,000 results, but for *conjunction − disjunction* 13,600,000 results (Searches were carried out on June 10, 2008; more on such inaccuracies can be found in Bar-Ilan, 2005). These inconsistencies are probably caused by partial evaluation of queries to increase search engine efficiency.

Thus, we cannot learn about the growth of a topic based on the numbers reported by the search engine, and in any case, we are interested in the actual Web pages, not just the number of retrieved Web pages. Here, we encountered a further problem: Search engines limit the actual number of Web pages retrieved for a query (Google does not retrieve more than 1,000 search results.) This problem for smaller queries can be overcome by what we call "chunking" (i.e., breaking up the original query into subqueries), a similar technique also advocated by Thelwall (2008). The idea is to create a nonoverlapping cover of subqueries of the original query, where each subquery returns at most 1,000 results (or the limit of the search engine if this limit is different). Currently, this can be achieved by the combination of the following:

- Including/excluding additional search terms, such as running the pair of queries *informetrics scientometrics* and *informetrics scientometrics*, when the topic we are interested in is *informetrics*. In this case, the first subquery returns pages that contain the term "informetrics" and also contain the term "scientometrics" while the second subquery returns pages that include "informetrics," but do not include "scientometrics."
- Limiting the query by site or filetype, such as informetrics site:.es (pages from Spain only) or informetrics filetype:.pdf (only pdf files) and using the same inclusion/exclusion technique as mentioned earlier. For example, if we want to chunk by including/excluding pages that contain the term "scientometrics," by including/excluding pages from Spain and by including/excluding pdf files for the topic "informetrics," we have to run eight queries:

  informetrics scientometrics site:.es filetype:.pdf
  informetrics scientometrics site:.es -filetype:.pdf
  informetrics scientometrics-site:.es filetype:.pdf
  informetrics scientometrics -site:.es -filetype:.pdf
  informetrics-scientometrics site:.es filetype:.pdf
  informetrics-scientometrics site:.es -filetype:.pdf
  informetrics-scientometrics -site:.es filetype:.pdf
  informetrics-scientometrics -site:.es -filetype:.pdf

  Note that if the search engine follows the rules of set theory, their intersection between any two of the results sets of the earlier queries should be empty; that is, each URL should appear in only one of the queries, and if all sets are of a size less than the limit of the search engine, then the union of the results sets should be equal to the set of all the URLs that the search engine indexed with the term "informetrics." In practice, this is not exactly the case because search engines use different approximation and retrieval methods to enhance the efficiency of searches, and users are rarely interested in the complete set of results. Also note that here we used the query format that is appropriate for Google as of the beginning of 2009. The query format may be different or even inapplicable with other search engines and may undergo changes with Google as well in the future.
- Limiting the query by date. Using this method, we also can partition the results set into nonintersecting subsets of smaller sizes. Such a feature was available at *Ask* during data collection using the *betweendate* modifier. At the beginning of 2009, this option no longer appeared on the advanced search page. Date-range queries can still be run at the old AltaVista site (www.altavista.com), which is now powered by Yahoo!. But note that search engines are highly dynamic; they remove and add features and behaviors all the time to improve the search experience of the public, and not necessarily to serve webometric purposes.

There are topics for which it is extremely difficult or even impossible to define a set of representative keywords, and additional methods have to be used to gather information about the extent of the topic on the Web. Poems or short stories are one such example: Pages containing poems or short stories usually do not contain these terms. In such cases, one has to apply other techniques to study the evolution of these topics.

None of the search engines provide comprehensive coverage of the Web. It has been shown (e.g., Bharat & Broder, 1998; Gulli & Signorini, 2005; Lawrence & Giles, 1998, 1999) that the overlap between search engines is small. Even though experimental data on the overlap are not recent and the exact overlap between the major search engines is unknown, it is advisable to collect data from several search engines to receive more comprehensive results.

Step 3: Follow-Up Data-Inspection Points

To study the evolution of the topic, data have to be gathered periodically. At the additional data-inspection points,

DOI: 10.1002/asi

two methods are applied. The first method is identical to the initial data-collection procedure. Its aim is to gather Web pages on the topic that are newly discovered by the search engines. These could be pages that were recently created, were created quite a while ago but were only recently discovered by the search engine, or were created some time ago but recently their content underwent changes and the page became relevant to the monitored topic. It is even possible that the page was indexed by the search engine at a previous data-inspection point and contained the search terms defining the topic, but for some reason, the search engine failed to retrieve it at the previous data-inspection points (Mettrop & Nieuwenhuysen, 2001).

The second method is to revisit URLs identified at previous data-inspection points, but not retrieved by the first method. The second method allows us to learn about changes that have occurred to previously identified URLs: These may have disappeared from the Web, may have been modified and ceased to be relevant to the topic, or the first data-collection method simply was not perfect and "missed" these pages. Search engines cannot and do not cover the whole Web (e.g., Bharat & Broder, 1998; Gulli & Signorini, 2005; Lawrence & Giles, 1998, 1999); the size of their database is limited, and they cannot retain all previously indexed Web pages if they want to cover newly added and newly discovered pages as well. Thus, they have to drop previously indexed pages in favor of newly created ones. Combining the two data-collection methods provides a more complete picture of what exists on the topic on the Web.

Step 4: Data Analysis

Our aim is to analyze the longitudinal patterns of the dataset in general and the changes in the distribution of the domains over time. First, all the collected Web pages should be checked to ascertain that they belong to the topic. Judging the *relevance* of the pages to the topic is very complex (e.g., Mizzaro, 1998; Saracevic, 1996); thus, we decided to check the *technical relevance* of the pages instead. A URL is considered *technically relevant* if the URL satisfies the query that defines the topic (Bar-Ilan, 2002, 2004). We are aware that this concept is weaker than relevance, but it can be checked automatically without the need for human judges.

Note that a page that is technically relevant might not be on the topic at all if, for example, the search term appears only on the sidebar of the retrieved page. As an interesting example, consider the search *"impact factor" journal* on Google Scholar. When submitting this query, our aim was to retrieve scholarly publications that discuss journal impact factors. The number of retrieved results is enormous (nearly 200,000 at the time we conducted the searches). A closer inspection (already the fourth result at the time we ran the query) showed that this search retrieved abstracts of articles that had nothing to do with journal impact factors, only that the publisher decided to announce the most recent impact factor of the journal in which the article was published.

A more formal definition of technical relevance and of related terms follows:

- A URL u is *technically relevant* at time t ($trel_t(u) = 1$), if the document residing at u at time t satisfies the query; otherwise, $trel_t(u) = 0$. All documents are checked at only the specific *data check points*; these are the initial data-inspection point and all the follow-up data-inspection points.
- A URL u is *technically relevant* during the period of time starting with $t1$ and ending with $t2$ [$trel_{t1\sim t2}(u) = 1$] if it was *technically relevant* at each time t, $t1 \leq t \leq t2$, that it was revisited. Note that it is impossible to detect whether the specific page changed several times between the data check points, at which times it might not have satisfied the query or even might have been removed temporarily.
- A URL u is *intermittent* during the period of time starting with $t1$ and ending with $t2$ (denoted $int_{t1\sim t2}(u) = 1$) if $trel_{t1}(\mathrm{u}) = 1$ and $trel_{t2}(u) = 1$ but there was a time t, $t1 < t < t2$, such that $trel_t(u) = 0$ (i.e., it either did not satisfy the query or was inaccessible at time t).
- A URL u has *disappeared* at time $t2$ (denoted d_{t2}) if $trel_t(u) = 1$ at all times t prior to $t2$, but for all data check points t, $t \geq t2$, $trel_t(u) = 0$. Note that it is possible that a URL defined as disappeared based on the available data will become intermittent if the data monitoring continues for a longer time.

Case Study

Data Collection

The experiment started in January 1998. In the first stage (until June 1998), data were collected from the then-major search engines (AltaVista, Excite, Hotbot, InfoSeek, Lycos, and Northern Light) by running the query *informetrics OR informetric.* Originally, we intended to run the query *informetrics* only, but because of Northern Light's automatic stemming, the query was extended. Data were collected once a month, and changes between the data collected in consecutive data-inspection points were observed (see details in Bar-Ilan & Peritz, 1999). In June 1998, 866 URLs were identified through the collective effort of the previously mentioned search engines. The query was chosen because we were looking for information on the scientific field of *informetrics*—quantitative analysis of documents in all forms; however, as can be expected, on the Web, informetrics has additional meanings as well (e.g., names of companies). Note that here we only demonstrate the data-collection method outlined earlier; we are well aware that the query is not sufficient for collecting all the pages belonging to the scientific field of informetrics.

Search results fluctuated considerably between the data-inspection points (see Bar-Ilan, 1999); thus, when rerunning the experiment in June 1999, an additional data-collection method was employed besides querying the search engines. The URLs that satisfied the query in June 1998 were revisited in 1999 even if they were not located by the search engines in 1999. No data were collected in 2000 and in 2001; however, in retrospect, this has not been a shortcoming of the

DOI: 10.1002/asi

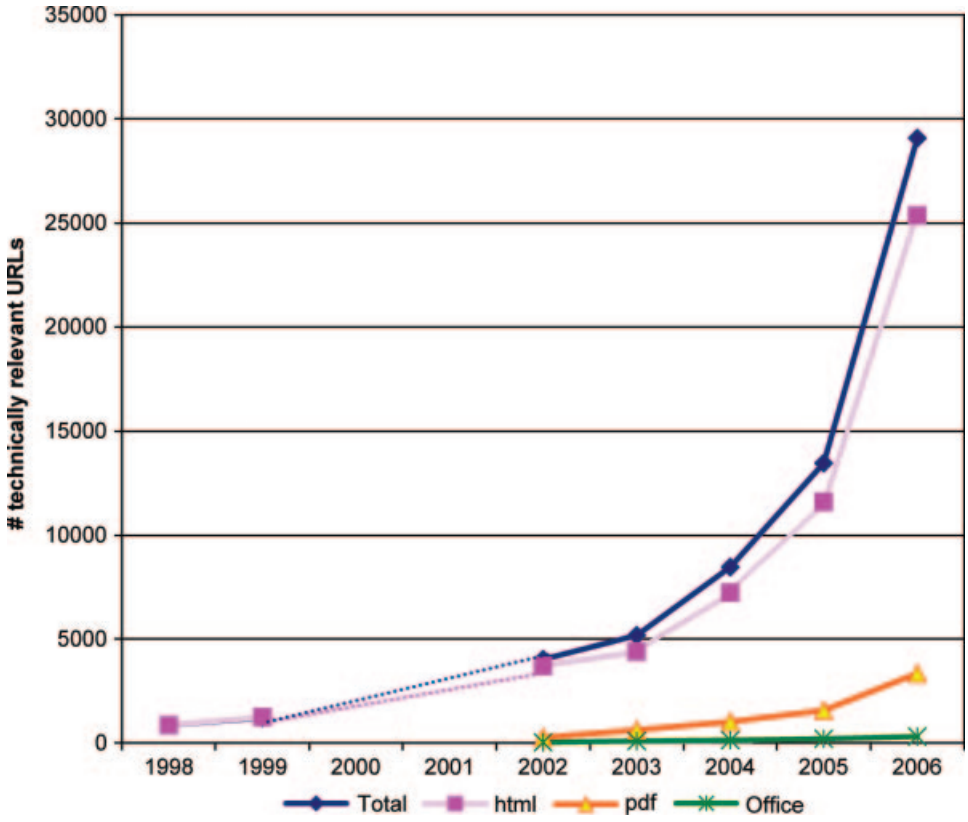

FIG. 1. The overall growth of the topic over the years as reflected by the number of *technically relevant* URLs identified at each of the data-inspection points.

research since the growth and change patterns can be easily interpolated for the missing data-inspection points (see Figure 1).

In June 1999, 2002, 2003, 2004, 2005, and 2006, two separate data-collection procedures were employed:

- Submitting the query *informetrics OR informetric* to the largest search engines at the time (Multiple search engines were used to increase the number of documents that satisfy the query.)
  - In 1999, the same search engines were used as in 1998; namely AltaVista, Excite, Hotbot, InfoSeek, Lycos, and Northern Light.
  - In 2002 and 2003, AllTheWeb, AltaVista, Google, HotBot, Teoma, and Wisenut were employed. By 2002, search engines started to retrieve non-HTML pages as well (pdf, ps, doc, etc.). We reported the findings for these data-inspection points in Bar-Ilan and Peritz (2004).
  - In 2004, we queried AllTheWeb, AltaVista, Gigablast, Google, Hotbot, Teoma, Yahoo!, and Wisenut. Note that in June 2004, AllTheWeb and AltaVista still retrieved slightly different results from the then-newly launched Yahoo! search engine; Hotbot served a different set of results as well.
  - In 2005 and 2006, we queried Exalead, Google, MSN, Teoma (Ask), and Yahoo!.

  Although the initial dataset was rather small (<900 URLs), enormous growth was witnessed during the years, and in 2006, the search engines retrieved 24,272 different URLs (An additional 4,642 URLs were located through the "revisit" process in 2006.)

  Search engines limit the number of displayed results for a query (Limitations as of June 2006 were 1,000 for Google and Yahoo!, 2,000 for Exalead, 250 for MSN, and 200 for Teoma.) To try to overcome these limitations, we used "chunking," as explained earlier. In one case, we included/excluded 22 additional terms to break down the query results into small-enough chunks.

  The whole set of searches on all search engines were run within 1 to 2 hr to minimize the effect of time on the search results. For each year, the searches were carried out in June. The URLs were extracted from the search results pages, and duplicates (usually the same URL retrieved by several search engines) were eliminated. The URLs were compared as text strings; thus, for example, informetrics.com and www.informetrics.com were considered two different URLs.

  All the documents residing at the identified URLs were downloaded to our local computer within 0 to 2 days of the searches to minimize the effect of time elapsed between the search and download times on the possible changes that the documents undergo over time. A second attempt was made to download inaccessible URLs. Finally, the entire set of HTML documents was tested for the presence of the string *informetric*.
- All pages that contained either the term *informetrics* or the term *informetric* (i.e., satisfied the query) at least the first time that they were identified by the search process were revisited at each of the later data-inspection points.

DOI: 10.1002/asi

TABLE 1. Number of technically relevant (trel) URLs identified at the data check points.

	Total trel URLs (% of total)	Trel HTML or text documents	Trel pdf documents	Trel MS Office documents	Trel postscript documents	Trel xml documents
1998	866 (2.4%)	866	0	0	0	0
1999	1,249 (3.3%)	1,249	0	0	0	0
2002	4,034 (11.1%)	3,705	272	31	26	0
2003	5,176 (14.3%)	4,399	625	92	60	0
2004	8,454 (23.3%)	7,225	1,027	140	62	0
2005	13,454 (37.1%)	11,594	1,577	210	73	0
2006	28,914 (80.2%)	25,358	3,349	310	63	18
Total unique URLs during whole period	36,282	31,999	3,839	360	84	18

TABLE 2. Longitudinal patterns of HTML documents.

First identified in		1999	2002	2003	2004	2005	2006
1998 (total 866)	trel	648	291	242	216	176	156
	intermittent			1	3	7	9
	inaccessible/disappeared	183	495	551	575	615	629
	term not in document	35	80	71	72	68	72
1999 (total 601)	trel		219	166	147	129	112
	intermittent			1	0	2	1
	inaccessible/disappeared		321	390	408	432	447
	term not in document		61	44	46	38	41
2002 (total 3,196)	trel			2,440	2,025	1,555	1,351
	intermittent				46	59	134
	inaccessible/disappeared			574	850	1,206	1,326
	term not in document			182	275	376	385
2003 (total 1,549)	trel				1,209	881	708
	intermittent					30	54
	inaccessible/disappeared				228	471	588
	term not in document				112	167	199
2004 (total 3,580)	trel					2,318	1,838
	intermittent						353
	inaccessible/disappeared					767	1,006
	term not in document					495	383
2005 (total 6,438)	trel						4,873
	intermittent						
	inaccessible/disappeared						915
	term not in document						650

As discussed earlier, the combination of the two methods allowed us both to follow the "fate" of previously identified pages and to enrich the collection of pages with newly retrieved pages from the search engines.

Results

Growth

During the 8-year period under study, 36,300 different URLs were identified that satisfied the query at least the first time they were located. Table 1 and Figure 1 depict the overall growth of the topic over the years as reflected by the number of *technically relevant* URLs identified at each of the data-inspection points. The growth over the years is considerable; 80.2% of the total unique URLs were identified during the whole period and satisfied the query at the last data check point (2006) while only 2.3% of the total were discovered by the search engines in 1998. When analyzing the data, we have to take into account two processes: the growth of the Web as a whole and changes in coverage of the search engines.

Longitudinal Patterns

Overall, there was a 33-fold growth in the number of technically relevant documents identified between 1998 and 2006. Even if we consider HTML documents only, we observe a nearly 30-fold growth. Addition/creation of new Web documents is not the only process that takes place on the Web. Documents may continue to exist, but cease to be technically relevant, and they may become temporarily or permanently inaccessible (i.e., the server or the document has been removed from the Web). Table 2 provides details

DOI: 10.1002/asi

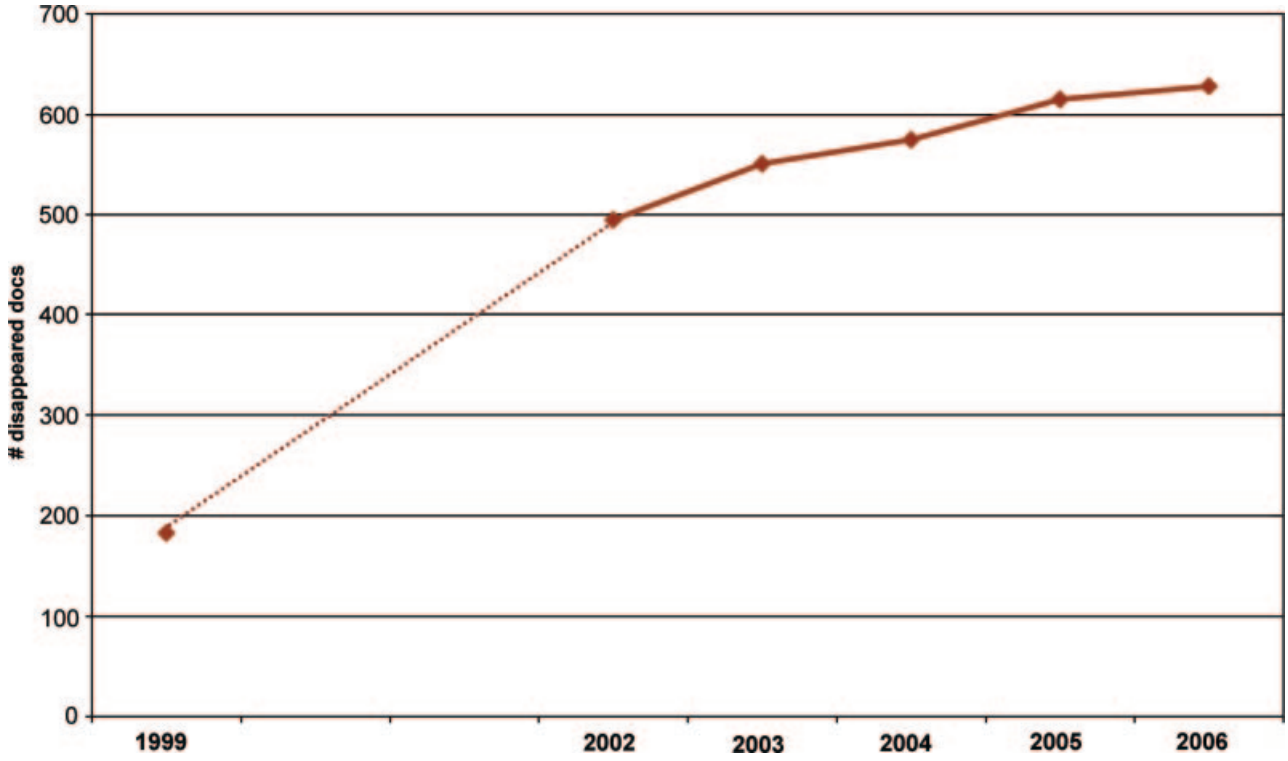

FIG. 2. The rate of disappearance over time.

about the longitudinal patterns of the URLs identified during 1998 to 2005. We see that even after 8 years, 165 of the initial 866 URLs still exist and still satisfy the query. Intermittence is quite negligible; that is, once a document is removed from the Web or ceases to contain the search terms, it rarely becomes technically relevant again. Some of the intermittence can be explained by the inaccessibility of the server at the specific data check point. Note that we tried to access all the documents that returned some error code for a second time, within 2 to 3 days of the data-inspection point. The data in Table 2 indicate that the most significant process besides the creation of new pages is the removal of existing ones from the Web. It seems that "younger" pages disappear at a faster rate than "older" ones. A possible explanation is that older pages become forgotten and abandoned. It can be seen from Figure 2, based on the URLs first identified in 1998, that the rate of disappearance slows down over time (for further analysis of the data, see Bar-Ilan & Peritz, 2009).

Shortcomings of Using Only a Single Data-Collection Method

Until this point, we have not differentiated between the URLs identified through the two data-collection methods: extensive search and revisiting of previously identified URLs. In this section, we demonstrate the necessity of multiple data-collection methods, even if we are not interested in the "fate" of the pages that are not part of the monitored topic anymore (i.e., disappeared, or still exist but their content has changed). One also may question the wisdom of using multiple search engines instead of using just the most comprehensive one.

The search engine scenery undergoes frequent changes; thus, there is no search engine that was used for data collection at all seven data-inspection points. Google was the most frequently utilized search engine; thus, we show the results of using Google for collecting HTML pages on the topic between 2002 and 2006.

As noted earlier, special care has been taken to try to retrieve all the documents the search engine reported having in its database for the given query. There are two obstacles. First, search engines omit results that they consider to be similar to the ones already displayed. This problem is rather easy to overcome either by clicking on the appropriate link on the end of the short list or by adding "&filter = 0" to the end of the search URL (Here, we are discussing Google only.) Second, there is a limit on the number of displayed search results (1,000 for Google) regardless of the number of results reported. To overcome this problem, the query has to be broken into a set of subqueries in such a way that the number of reported results for each subquery is less than 1,000 and the set of subqueries covers the original query. We employed this technique; however, it is known that Google is "a little weak on 'search engine math'" (Bar-Ilan, 2005), and it is quite possible that the number of URLs obtained by the subquery method is less than the total.

Data were collected from Google from 2002 onwards (five data-inspection points); in 1999, it was not among the largest search engines, and no data were collected in 2000 and 2001.

DOI: 10.1002/asi

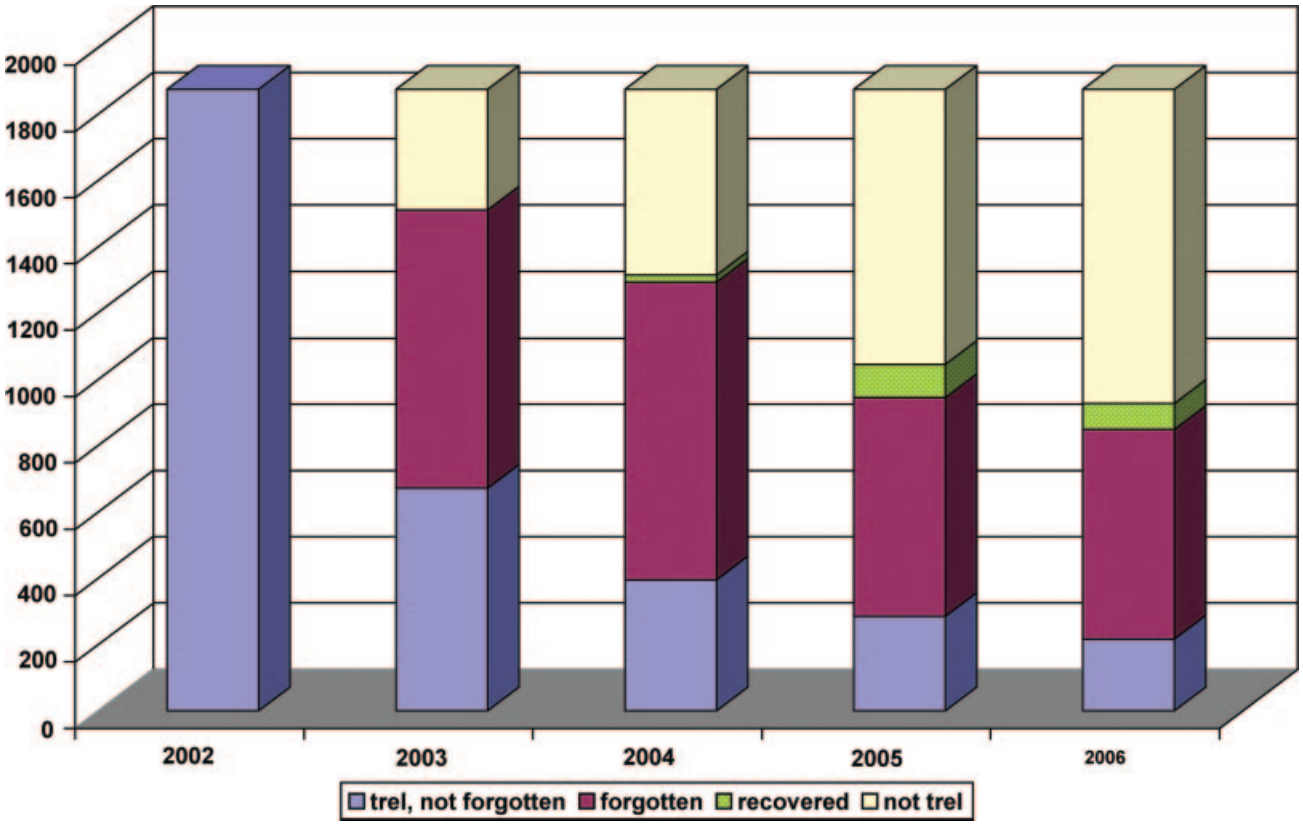

FIG. 3. The extent of the dynamic changes that Google has undergone; these processes are especially visible for the URLs first retrieved by Google in 2002.

Google is the search engine that has been employed for the largest number of times for data collection for this study.

In addition to the measures introduced earlier, for a search engine, we can compute additional measures (see Bar-Ilan, 2002, 2004) that reflect on the performance of the search engine over time.

- A URL u is *forgotten* at time t if it was retrieved by the search engine at time $t1 < t$, $trel_t(u) = 1$ (i.e., it exists and satisfies the query at time t), but it was not retrieved by the search engine at time t.
- A URL u is *recovered* at time $t2$ if it was forgotten at time $t < t2$ but was retrieved by the search engine at time $t2$ and $trel_{t2}(u) = 1$. Note that not all URLs can be recovered at a later time; even though at time t it was *technically relevant* but was not retrieved by the search engine, it is possible that a later time $t2$, it ceases to exist altogether or ceases to contain the search term (i.e., $trel_{t2}(u) = 0$).

Search engines are not and cannot be stable (e.g., Bar-Ilan, 2004, 2005); thus, they introduce an additional level of dynamicity into the already-dynamic Web. The search engine's database is of limited size and grows more slowly than the Web; thus, pages indexed once inevitably have to be dropped in favor of newly discovered pages.

The extent of the dynamic changes that Google has undergone is considerable; these processes are especially visible for the URLs first retrieved by Google in 2002, as can be seen in Figure 3. Note that the number of "forgotten" pages sometimes decreases over the years because some of the URLs not picked up by the search engine cease to be technically relevant to the query. Overall, we see a monotonic decrease in the number of *trel* pages retrieved. This is a result of two processes: (a) Some of the original *trel* pages cease to exist or cease to be technically relevant, and (b) there are dynamic changes in the lists of URLs covered by the search engine. Had we used only Google (or any other single search engine) to collect data, we thus would not have observed some of the growth of the topic. Even when using multiple search engines to collect data, some pages still may remain forgotten; thus, supplementing the search procedure with revisiting previously identified URLs is essential. In addition, one must be reminded that some, or perhaps many, technically relevant pages remain undiscovered despite the use of multiple search engines since the search engines do not cover the whole Web (e.g., Bharat & Broder, 1998; Lawrence & Giles, 1998, 1999).

Conclusions

This article describes a general methodology for studying the evolution of a topic on the Web. To our knowledge, our extensive case study is the first longitudinal one (8 years) on the evolution of a topic on the Web that observes the "birth"

DOI: 10.1002/asi

of new pages and the "decay" and/or the "modification" of existing ones. Actually, we are not aware of additional studies that have monitored topics over time. Most other studies have monitored either a fixed set of Web pages or a set of Web sites. When monitoring a fixed set of Web pages (e.g., Koehler, 2004), growth is ignored, and only decay and changes can be observed. When monitoring Web sites, one learns only about the growth and changes within the monitored sites. These are, of course, important observations as well, especially when studying academic Webs (Payne & Thelwall, 2007, 2008) or the Web pages of a whole country (e.g., Baeza-Yates & Poblete, 2003). Monitoring a topic enables us to learn about the different evolutionary processes over the *whole* Web.

The results show that models of the evolving Web (e.g., Barabasi et al., 2000) have to take into account not only growth but disappearance and modification as well. The models that do not take into account disappearance and modification work reasonably well because at this stage of the Web development, the major process is the addition of new pages. But to understand Web evolution, there is need for more sophisticated models.

The growth of the specific topic on the Web is astonishing. If we compare the growth in the number of scientific publications that contain the words "informetrics" or "informetric" and were published before 1999 (corresponding to the first data-inspection point) with the number of publications published before 2007 (corresponding to the last data-inspection point) as reported by the Thomson Reuters *Web of Science*, we observe a 2.5-fold growth as compared with the 33-fold growth on the Web.

We also have shown that when carrying out longitudinal studies, one cannot rely on a single search engine, even if it is the largest one, because of the dynamic changes in the content of the search engine's database. We advocate using multiple search engines; but even when following this advice, one must take into account the limitations of collecting data through search engines. Search engines do not and cannot cover the whole Web, not even the whole "indexable" Web (loosely defined by Lawrence & Giles, 1999, p. 99 as "the Web that the engines do consider indexing") and definitely do not cover the so-called "invisible Web" (UC Berkeley, 2008).

The second method of revisiting previously identified URLs allows ascertaining the "fate" of previously identified Web pages belonging to the topic, whether they were simply dropped from the search engines' indexes because of space limitations or they ceased to exist or simply ceased to belong to the specific topic. This method is essential in cases where we do not only want to characterize growth, but also other processes such as decay and modification as well, and it helps to overcome some of the complexities related to using only search engines only for data collection. A possible technique to increase the coverage even further is to employ focused crawling starting from the pages discovered by the search engines, but this requires more computational resources.

Although longitudinal studies are not easy to conduct, they are the only ones to provide us with a clear and comprehensive picture of the development of a field; they are needed and strongly recommended. It is our belief that the methods applied in this study can be applied in various settings to discover coverage, growth, decay, and other special and unforeseen characteristics of the Web. Such additional studies will help in evaluating the typology of the temporal states of Web pages introduced in this article.

Acknowledgment

A preliminary version of this article was presented at the 1st International Workshop on Understanding Web Evolution (WebEvolve2008), April 22, 2008, in Beijing, China.

References

Adar, E., Teevan, J., Dumais, S.T., & Elsas, J.L. (2009). The Web changes everything: Understanding the dynamics of Web content. In Proceedings of the Second ACM International Conference on Web Search and Data Mining (pp. 282–291), Barcelona, Spain. New York: ACM Press.

Albert, R., & Barabasi, A.L. (2000). Topology of evolving networks: Local events and universality. Physical Review Letters, 85(24), 5234–5237.

Baeza-Yates, R., Castillo, C., & Saint-Jean, F. (2004). Web dynamics, structure and page quality. In M. Levene & A. Poulovassilis (Eds.), Web dynamics (pp. 93–112). Berlin, Germany: Springer.

Baeza-Yates, R., & Poblete, B. (2003). Evolution of the Chilean Web structure composition. In Proceedings of the First Latin American Web Congress (pp. 11–13). Washington, DC: IEEE CS Press.

Barabasi, A.L., Albert, R., & Jeong, H. (2000). Scale-free characteristics of random networks: The topology of the World Wide Web. Physica A, 281, 69–77.

Bar-Ilan, J. (1999). Search engine results over time: A case study on search engine stability. Cybermetrics, 2/3(1), Paper No. 1. Retrieved June 10, 2008, from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html

Bar-Ilan, J. (2000). Evaluating the stability of the search tools HotBot and Snap: A case study. Online Information Review, 24(6), 439–449.

Bar-Ilan, J. (2002). Methods for measuring search engine performance over time. Journal of the American Society for Information Science and Technology, 54(3), 308–319.

Bar-Ilan, J. (2004). Search engine ability to cope with the changing Web. In M. Levene & A. Poulovassilis (Eds.), Web dynamics (pp. 195–218). Berlin, Germany: Springer.

Bar-Ilan, J. (2005). Expectation versus reality—Search engine features needed for Web research at mid-2005. Cybermetrics, 9, Paper No. 2. Retrieved June 10, 2008, from http://www.cindoc.csic.es/cybermetrics/articles/v9i1p2.html

Bar-Ilan, J., & Peritz, B.C. (1999). The life-span of a specific topic on the Web—The case of "informetrics:" A quantitative analysis. Scientometrics, 46, 371–382.

Bar-Ilan, J., & Peritz, B.C. (2004). Evolution, continuity and disappearance of documents on a specific topic on the Web—A longitudinal study of "informetrics." Journal of the American Society for Information Science and Technology, 56, 980–990.

Bar-Ilan, J., & Peritz, B.C. (2009). The lifespan of "informetrics" on the Web: An eight year study (1998–2006). Scientometrics, 79(1), 7–25.

Bar-Yossef, Z., Broder, A.Z., Kumar, R., & Tomkins, A. (2004). Sic transit gloria telae: Towards an understanding of the web's decay. In Proceedings of the 13th International World Wide Web Conference (pp. 328–337). New York: ACM Press.

Berners-Lee, T., Hall, W., Hendler, J., Shadboldt, N., & Weitzner, D.J. (2006). Creating a science of the Web. Science, 313, 769–771.

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. In Proceedings of the

DOI: 10.1002/asi

Seventh International World Wide Web Conference. Computer Networks and ISDN Systems, 30, 379–388.

Brewington, B.E., & Cybenko, G. (2000). Keeping up with the changing Web. Computer, 33(5), 52–58.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph structure in the Web. Computer Networks, 33, 309–320.

Broder, A.Z., Glassman, S., Manasse, M., & Zweig, G. (1997). Syntactic clustering of the Web. Computer Networks and ISDN Systems, 29, 1157–1166.

Chakrabarti, S., van den Berg, M., & Dom, B. (1999). Focused crawling: A new approach to topic-specific Web resource discovery. Computer Networks, 31, 1623–1640.

Cho, J., & Garcia-Molina, H. (2000). The evolution of the Web and implications for an incremental crawler. In Proceedings of 26th International Conference on Very Large Databases (pp. 200–210). San Francisco: Morgan Kaufmann.

de Kunder, M. (2008). The size of the World Wide Web. Retrieved June 9, 2008, from http://www.worldwidewebsize.com/

Dontcheva, M., Drucker, S.M., Salesin, D., & Cohen, F.M. (2007). Changes in webpage structure over time. UW CSE Technical Report No. 20070402. Retrieved April 28, 2009, from http://www.cs.washington.edu/homes/mirad/research/pubs/TR2007-04-02.pdf

Dorogovtsev, S.N., & Mendes, J.F.F. (2000). Scaling behaviour of developing and decaying networks. Europhysics Letters, 52, 33–39.

Fenner, T., Levene, M., & Loizou, G. (2005). A stochastic evolutionary model exhibiting power-law behaviour with an exponential cutoff. Physica A, 355, 641–656.

Fenner, T., Levene, M., & Loizou, G. (2006). A stochastic model for the evolution of the Web allowing link deletion. ACM Transactions on Information Technology, 6, 117–130.

Fetterly, D., Manasse, M., Najork, M., & Wiener, J.L. (2004). A large scale study of the evolution of Web pages. Software—Practice and Experience, 34, 213–237.

Gomes, D., & Silva, M.J. (2006). Modeling information persistence on the Web. In Proceedings of the Sixth International Conference on Web Engineering (pp. 193–200). New York: ACM Press.

Grandi, F. (2004). An annotated bibliography on the temporal and evolution aspects in the World Wide Web. Retrieved October 14, 2008, from http://www-db.deis.unibo.it/~fgrandi/TWbib/

Gulli, A., & Signorini, A. (2005). The indexable Web is more than 11.5 billion pages. In Proceedings of 14th International World Wide Conference (pp. 902–903). New York: ACM Press.

Huberman, B.A., & Adamic, L. (1999). Growth dynamics of the World Wide Web. Nature, 401, 131.

Internet World Stats. (2008). World Internet users. Retrieved June 9, 2008, from http://www.internetworldstats.com/stats.htm

Ke, Y., Deng, L., Ng, W., & Lee, D.L. (2006). Web dynamics and their ramifications for the development of Web search engines. Computer Networks, 50, 1430–1447.

Kim, S.J., & Lee, S.H. (2005). An empirical study on the change of Web pages. In Proceedings of APWeb 2005, Lecture Notes in Computer Science, 3399, 632–642.

Koehler, W. (2004). A longitudinal study of Web pages continued: A report after six years. Information Research, 9(2), Paper No. 174. Retrieved June 9, 2008, from http://InformationR.net/ir/9-2/paper174.html

Kwon, S.Y., Lee, S.H., & Kim, S.J. (2006). A precise metric for measuring how much Web pages change. Lecture Notes in Computer Science, 3882, 557–571.

Lawrence, S., & Giles, C.L. (1998). Searching the World Wide Web. Science, 280, 98–100.

Lawrence, S., & Giles, C.L. (1999). Accessibility of information on the web. Nature, 400, 107–109.

Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines—Fluctuations in document accessibility. Journal of Documentation, 57(5), 623–651.

Mizzaro, S. (1998). How many relevances in information retrieval? Interacting With Computers, 10, 305–322. Retrieved June 10, 2008, from http://www.dimi.uniud.it/mizzaro/research/papers/IwC.pdf

Nielsen Online. (2008). Nielsen Online announces April U.S. search share rankings. Retrieved June 10, 2008, from http://www.nielsen-netratings.com/pr/pr_080519.pdf

Ntoulas, A., Cho, J., & Olston, C. (2004). What's new on the Web? The evolution of the Web from a search engine perspective. In Proceedings of the 13th International World Wide Web Conference (pp. 1–12). New York: ACM Press.

Olston, C., & Pandey, S. (2008). Recrawl scheduling based on information longevity. In Proceedings of the 17th International World Wide Web Conference (pp. 437–446). New York: ACM Press.

Ortega, J.L., Aguillo, I., & Prieto, J. (2006). A longitudinal study of content and elements in scientific Web environment. Journal of Information Science, 32, 344–351.

Pandia. (2007). The size of the World Wide Web. Retrieved June 9, 2008, from http://www.pandia.com/sew/383-web-size.html

Payne, N., & Thelwall, M. (2007). A longitudinal study of academic webs: Growth and stabilisation. Scientometrics, 71(3), 523–539.

Payne, N., & Thelwall, M. (2008). Longitudinal trends in academic web links. Journal of Information Science, 34(1), 3–14.

Risvik, K.M., & Michelsen, R. (2002). Search engines and Web dynamics. Computer Networks, 39, 289–302.

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and Northern Light. Cybermetrics, 2/3(1), Paper No. 2. Retrieved June 10, 2008, from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Saracevic, T. (1996). Relevance reconsidered. In Proceedings of the Second Conference on Conceptions of Library and Information Science (pp. 201–218), Copenhagen, Denmark. Retrieved June 10, 2008, from http://www.scils.rutgers.edu/~tefko/CoLIS2_1996.doc

Thelwall, M. (2001). The responsiveness of search engine indexes. Cybermetrics, 5(1), Paper No. 1. Retrieved June 10, 2008, from http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html

Thelwall, M. (2008). Extracting accurate and complete results from search engines: Case study Windows Live. Journal of the American Society for Information Science and Technology, 59, 38–50.

Toyoda, M., & Kitsuregawa, M. (2006). What's really new on the Web? Identifying new pages from a series of unstable web snapshots. In Proceedings of the 15th International World Wide Web Conference (pp. 233–241). New York: ACM Press.

UC Berkeley. (2008). Invisible Web: What it is, why it exists, how to find it and its inherent ambiguity. Retrieved October 14, 2008, from http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvisibleWeb.html

DOI: 10.1002/asi