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
972-3-5318351
barilaj@mail.biu.ac.il

Bluma C. Peritz
The Hebrew University of Jerusalem
Jerusalem, 91904, Israel
972-2-5638934
bluer@cc.huji.ac.il

ABSTRACT

The World Wide Web is growing at an enormous speed, and has become an indispensable source for information and research. New pages are being added to the Web, but there are additional processes as well: pages are moved or removed and/or their content changes. In order to obtain a better understanding of these processes, we developed a method for tracking topics on the Web for long periods of time. We use multiple data collection methods that 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 their 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 eight years. The data collection method also allowed us to analyze the dynamic changes in search engine coverage; here we illustrate these changes on Google, the search engine used for the longest period of time for data collection in this project.

Categories and Subject Descriptors

H.3.3 **[Information Storage and Retrieval]** Information Search and Retrieval, H.3.5 **[Information Storage and Retrieval]** Online Information Systems – *Web-based services*

General Terms

Experimentation, Measurement

Keywords

Web evolution, longitudinal patterns, growth, decay, intermittence

1. 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 [13].

This research 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 collection points. The combination of the two techniques allowed us to study several evolution patterns: creation of new pages, removal of previously existing ones and modification of the content and the 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 eight 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 indicate how the World Wide Web has become a major information source during a relatively short period of time. Such studies also help in understanding the changing roles of the Internet in the overall development of a topic.

2. RELATED WORK

2.1 Web growth, dynamics and structure

There has been great interest in estimating the size of the Web [14, 24, 30, 31] even though it is now quite clear that the “indexable Web” – a notion introduced by Lawrence and Giles [30] (“the Web that the engines do consider indexing” p.99) is not definable. Based on large crawls of the Web, Broder et al. [14] modeled the Web as a bow-tie graph. A similar structure emerged from the much smaller Chilean Web [3].

The Web seems to be a scale-free network, and its emergence can be explained by preferential attachment [4]. The basic model does not take into account page or link deletions. In a slightly different model, Albert and Barabasi [1] considered changes to the existing link structure, by what they called “rewiring”. Huberman and Adamic [25] and Fenner et al. [20] introduced models that allow for removal of Web pages. Fenner et al. [21] and Dorogotsev and Mendes [19] proposed models where link deletions are allowed. The above-mentioned 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 (appearance, disappearance, modification and redirection).

2.2 Longitudinal studies

Previous longitudinal studies monitored sites and pages for shorter periods of time, usually for several weeks or months (e.g. [18, 22, 27, and 34]). However in these shorter term studies, the data sets were huge and the monitored pages were visited more often (typically once a week).

WWW 2008, April 21--25, 2008, Beijing, China.

There are only a few studies that report findings based on several years of data collection, but even these are for shorter length than the current study. Koehler [28] monitored a fixed set of pages for 325 weeks (over six years). Gomes and Silva [23] had data on the Portuguese national web for a period of three years (8 data collection points), Baeza-Yates and Poblete [2] based their results on three data collection points over a period of three years of the Chilean Web, and Toyoda and Kitsuregawa [41] had access to the Japanese Web archive which collects data about once a year, and based their results on data from 2003-4 (three data collection points). Ortega et al. [34] crawled about a thousand sites twice, once in 1997 and once in 2004. Additional studies are covered in the survey by Ke et al. [26].

All previous studies that we were able to locate used a single data collection method. They either monitored a fixed data set (e.g., [22, 28]) or crawled in a pre-specified manner a fixed number of pages from given starting points (e.g. [18, 27]), or attempts were made to download complete websites (e.g. [34]) and/or entire national Webs (e.g., [2, 23, 41]).

Bar-Yossef et al. [12] proposed to measure decay, and computed the measure both for currently existing pages and for previous versions of the pages accessed through the Internet Archive. Fetterly et al. [22] based their similarity measure on “pre-images”, a modification of the “shingles” introduced by Broder et al. [15]. Kim and Lee [27] computed among other measures the modification rate of pages. Ntoulas et al. [34] assessed the degree of change between different versions of the same page based on TD·IDF and word distance. Kwon et al. [29] suggest to measure change based on edit distance. Toyoda and Kitsuregawa [41] computed the “novelty measure” assessing whether the newly identified pages at a given data collection point are really “new” or simply they were not discovered in the previous crawls.

2.3 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. We mention search engine dynamics as well, because search engines are our primary data collection sources when gathering information on a specific topic. The dynamics of search engines were observed and recorded in several studies (e.g., [5, 6, 32, 36, 37 and 40]). Bar-Ilan [7 and 8] introduced a set of measures to assess these dynamic changes.

3. THE GENERAL METHOD

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

1. Defining the topic. If search engines are utilized for data collection, it is essential to delineate the topic properly with the choice of proper keywords. Suppose that the chosen topic is the semantic web. Currently (as of January 30, 2008) Google reports about 5,100,00 results for semantic Web. However there are additional Web pages on the topic, where the term “semantic Web does not appear explicitly. For example the query ‘OWL Web Ontology Language -"semantic web"’ returns an additional 31,000 documents. We did not search for OWL alone, because both OWL and RDF have multiple meanings, which poses additional problems.

If data is collected through focused crawling [17] then examples of pages relevant to the topic are needed (depending on the method both positive and negative examples might be needed). The quality of the focused crawl is dependent on the representativeness of the examples provided.

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, since this is the technique applied in the case study.

Even if we found a set of keywords that cover the chosen topic, we are faced with additional problems. We describe some of the problems that are experienced when using Google, which is currently the most popular search engine and is one of the search engines with the widest coverage of the Web. Google currently does not allow more than 32 keywords per query. In addition Google does not allow to submit complex Boolean queries (e.g., conjunctions of disjunctions) – it does not recognize the use of parentheses in queries. It is also not very good in “search engine math”, and seemingly it provides only partial support for disjunctions. For example for the query “conjunction” it reported 49,300,000 results, but for “conjunction OR disjunction” only 16,300,000 results are reported; for disjunction 571,000 results, but for “disjunction –conjunction” 675,000 results (more on such inaccuracies can be found in [9]).

Thus we cannot learn about the growth of a topic based on the numbers reported by the search engine. We want to retrieve the actual web pages. Here we encounter 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” (breaking up the original query into subqueries), as demonstrated in section 4. This method is also advocated by Thelwall [39]. For larger queries, i.e. for queries retrieving hundreds of thousand results this method is not effective, but still if the search engines wish to cooperate, the result sets can be transferred to the interested parties for further analysis.

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 in order to gather information about the extent of the topic on the Web. One such example is poems or short stories – pages containing poems or short-stories usually do not contain these terms – in such cases different techniques have to be employed to study the evolution of these topics.

None of the search engines provides comprehensive coverage of the Web. It has been shown before [14, 24, 31] that the overlap between the search engines is small. Even though the experimental data on the overlap is old, and the exact overlap between the major search engines is unknown, in order to receive more comprehensive results, it is advisable to collect data from several search engines.

3. Follow-up data collection points. In order to study the evolution of the topic, data has to be gathered periodically. At the additional data collection points two methods are applied. The first method is identical to the initial data collection procedure. The second method is to revisit URLs identified at previous data collection points but not retrieved by the first method. The second method allows us to learn about changes that 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 simply the first data collection method was not perfect and “missed” these pages.

4. Data analysis. Our aim is to analyze the longitudinal patterns of the data set 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. The terminology *technical relevance* is applicable when the topic is defined by a query. A *technically relevant* URL is a URL that satisfies the query that defines the topic. We decided to use the terminology *technical relevance* instead of the more widely used term *relevant* in order to avoid the complex issues of defining relevance (see for example [38] or [32]).

- 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 only at the specific *data check points*; these are the initial data collection point and all the follow-up data collection 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 might have even been removed.
- 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 $trel_{t1}(u)$=1 and $trel_{t2}(u)$=1, but there was a time t, $t1<t<t2$ such that $trel_t(u)$=0 (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 checking 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.

4. THE CASE STUDY

4.1 Data collection

The experiment started in January 1998. In the first stage (until June 1998) data was 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 *informetics* only, but because of Northern Light’s automatic stemming the query was extended. Data was collected once a month and changes between the data collected in consecutive data collection points were observed (see details in [10]). In June 1998, 866 URLs were identified through the collective effort of the above-mentioned search engines. The query was chosen because we were looking for information on the scientific field *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 above, we are well-aware that that the query is not sufficient for collecting all the pages belonging to the specific scientific field.

Search results fluctuated considerably between the data collection points (see [10]), 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 was collected in 2000 and in 2001. However, in retrospect this has not been a shortcoming of the research, since the growth and change patterns can be easily interpolated for the missing data collection points (see Fig 1).

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

1. Submitting the query *informetrics OR informetric* to the largest search engines at the time (multiple search engines were used in order to increase the number of documents that satisfy the query)
   a. In 1999 the same search engines were used as in 1998, namely AltaVista, Excite, Hotbot, InfoSeek, Lycos and Northern Light
   b. 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.). In [11] we report the findings for these data collection points.
   c. 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; and Hotbot served a different set of results as well.
   d. In 2005 and 2006, we queried Exalead, Google, MSN, Teoma (Ask) and Yahoo.

   Although the initial data set was rather small (less than 900 URLs), enormous growth was witnessed during the years, and in 2006 the search engines retrieved 24,272 different URLs (4,642 additional URLs were located through the “revisit” process in 2006).

   Search engines limit the number of displayed result for a query (the limitations as of June 2006 were: 1000 for Google, Yahoo, 2000 for Exalead, 250 for MSN and 200 for Teoma). In order to try to overcome these limitations we used several techniques:

   a. Including/excluding additional search terms (e.g. *informetrics -scientometrics* and *informetrics scientometrics*
   b. Limiting the query by site or filetype e.g. informetrics site:.es (pages from Spain only) – including/excluding sites or filetypes.
   c. Limiting the query by date (the betweendate feature of Ask)

   In one case we included/excluded 22 additional terms in order to break down the query results into small enough chunks.

   The whole set of searches on all the search engines were run within 1-2 hours 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 0-2 days of the searches, in order minimize the effect of the time elapsed between the search time and download time 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*.

2. All pages that contained either the term *informetrics* or the term *informetric* (i.e., satisfied the query) at least at the first time that they were identified by the search process were revisited at each of the later data collection points.

As discussed in section 3, 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. Note that newly retrieved pages are not necessarily newly created pages. It is possible that the page existed before and was indexed by some of the search engines, but it did not contain the search term; or because of the incomplete coverage of the Web by the search engines, it is quite plausible that the page existed for a long time and was relevant to the search but was only discovered at one of the later data collection points.

5. RESULTS

5.1 Growth

During the whole period 31,999 different URLs were identified that satisfied the query at least at 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 collection points. The growth over the years is considerable, 80.2% of the total unique URLs identified during the whole period located 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.

5.2 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 about the longitudinal patterns of the URLs identified during 1998-2005. We see that even after eight years 165 out of the initial 866 URLs still exist and still satisfy the query. Intermittence is quite negligible, i.e., 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 2-3 days. The data in Table 2 indicates that the most significant process beside 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.

5.3 Search engine coverage – Google

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. However, the results show that a considerable number of previously identified URLs are "forgotten" over time by the search engines. Because of the frequent changes in the search engine scenery, there is not a single search engine that participated in the data retrieval at all seven data check points; we chose to demonstrate the issue on the html documents retrieved by Google between 2002 and 2006.

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

	Total *trel* URLs (% out 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**

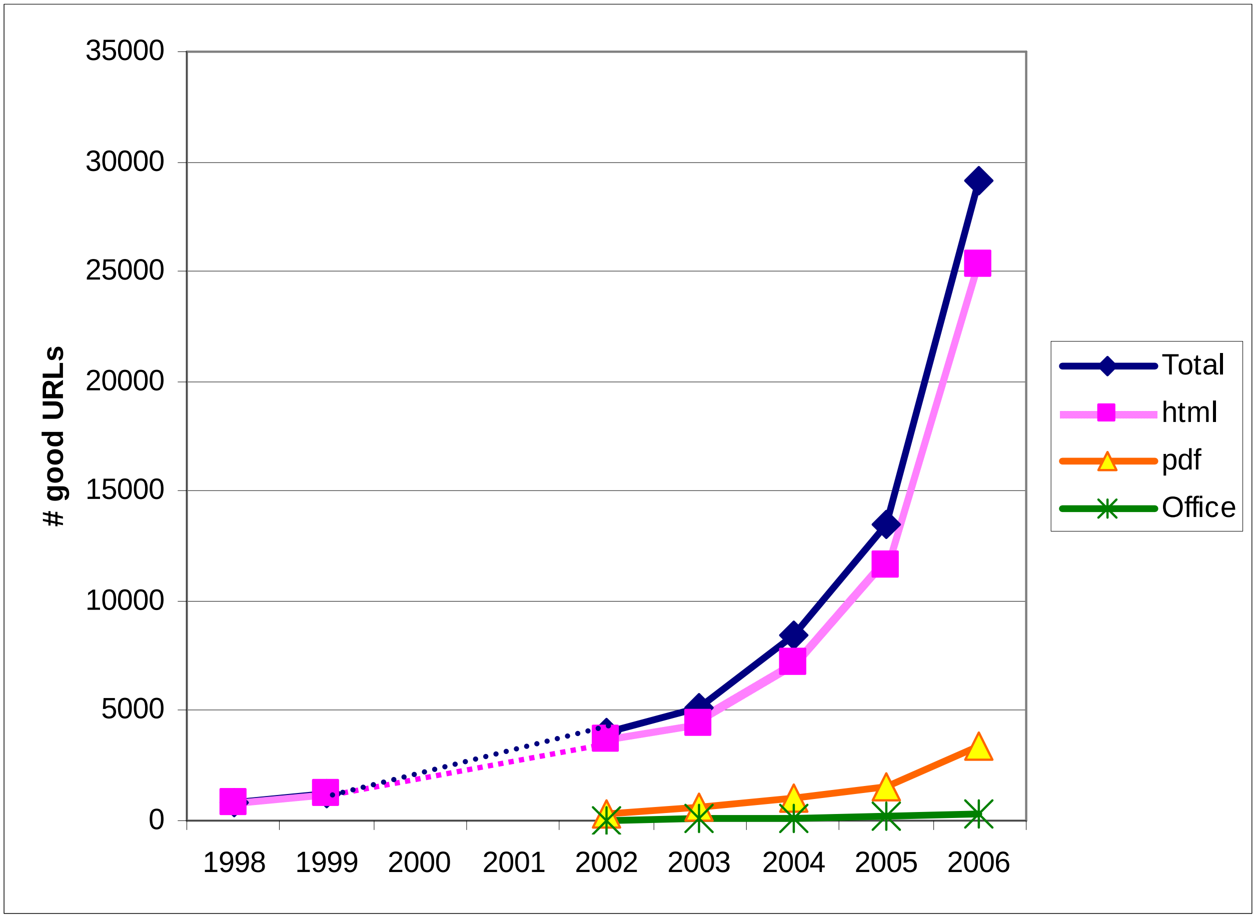

Figure 1: Growth curves for the different document types (growth curves interpolated for 2000 and 2001)

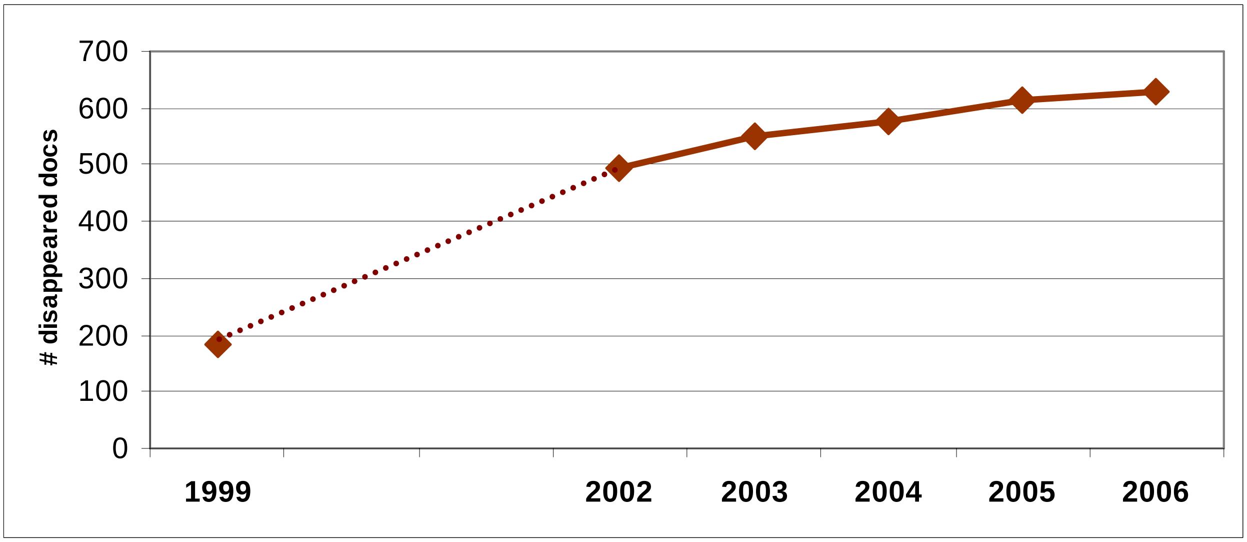

Figure 2: Disappearance rate of documents from the first dataset (1998). Curve interpolated for 2000 and 2001

Table 2: Longitudinal patterns of html documents

first identified in		1999	2002	2003	2004	2005	2006
1998	***trel***	648	291	242	216	176	156
(total 866)	**intermittent**			1	3	7	9
	inaccessible/disappeared	183	495	551	575	615	629
	term not in document	35	80	71	72	68	72
1999	***trel***		219	166	147	129	112
(total 601)	**intermittent**			1	0	2	1
	inaccessible/disappeared		321	390	408	432	447
	term not in document		61	44	46	38	41
2002	***trel***			2440	2025	1555	1351
(total 3196)	**intermittent**				46	59	134
	inaccessible/disappeared			574	850	1206	1326
	term not in document			182	275	376	385
2003	***trel***				1209	881	708
(total 1549)	**intermittent**					30	54
	inaccessible/disappeared				228	471	588
	term not in document				112	167	199
2004	***trel***					2318	1838
(total 3580)	**intermittent**						353
	inaccessible/disappeared					767	1006
	term not in document					495	383
2005	***trel***						4873
(total 6438)	**intermittent**						
	inaccessible/disappeared						915
	term not in document						650

As noted before, special care has been taken to try to retrieve all the documents the search engine reports to have in its database for the given query. There are two obstacles one faces: 1) 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). 2) There is a limit on the number of displayed search results (1000), 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 1000, 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’” [9], and it is quite possible that the number of URLs obtained by the subquery method is less than the total.

Data was collected from Google from 2002 onwards (five data collection points) – in 1999 it was not among the largest search engines, and no data was collected in 2000 and 2001. Even so, 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 in section 3, for a search engine we can compute additional measures ([7 and 8]) 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., $trel_{t2}(u)=0$.
- Our dataset is based on retrieval from additional search engines as well, thus we can also calculate the number of URLs that were retrieved by the given search engine for the first time at time t, but were located by other means at time $t1<t$. In this case as well, we are only counting technically relevant URLs both at $t1$ and t. Note that here we only count

URLs that were eventually discovered by the search engine. We call these URLs *missed* URLs.

The extent of the dynamic changes that Google undergoes is considerable; these processes are especially visible for the URLs first retrieved by Google in 2002, as can be seen in Figure 4. Note that the number of "forgotten" pages sometimes decreases over the years; this is 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: (1) some of the originally *trel* pages cease to exist or cease to be technically relevant (2) there are dynamic changes in the lists of URLs covered by the search engine.

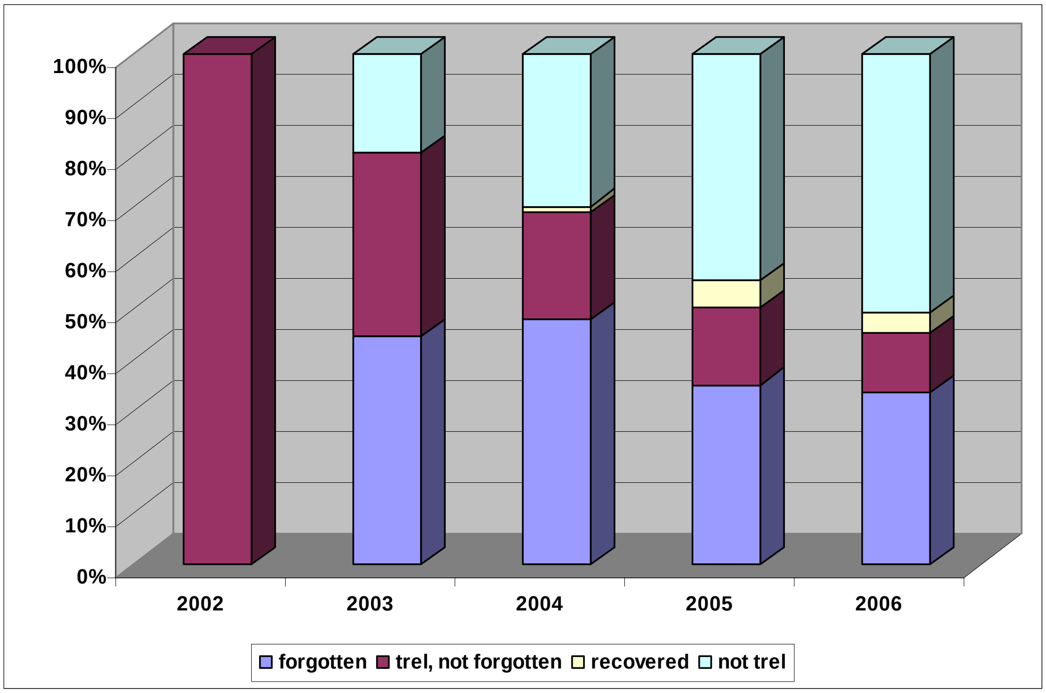

Figure 4: Dynamic changes in the retrieval of Google as reflected by the dataset first retrieved in 2002

6. CONCLUSIONS

This paper describes a general methodology for studying the evolution of a topic on the Web. To our knowledge our case study is the first extensive longitudinal study (eight years) of a topic on the Web that observes the "birth" of new pages, the "decay" and/or the "modification" of existing ones. The results show that models of the evolving Web (e.g., [1, 4]) have to take into account not only growth, but disappearance and modification as well. In addition for longitudinal studies one cannot rely on a single search engine, even if it is the largest, because of the dynamic changes in the content of the search engine's database. Although longitudinal studies are not easy to conduct, they are needed and recommended. It is our belief that the methods applied in this study can be applied in various settings in order to discover coverage, growth, decay and other longitudinal patterns and characteristics on the Web.

7. REFERENCES

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

[2] Baeza-Yates, R., and Poblete, B. Evolution of the Chilean Web structure composition. In Proceedings of the First Latin American Web Congress (LA-WEB 2003).

[3] Baeza-Yates, R., Castillo, C., and Saint-Jean, F. Web dynamics, structure and page quality. In (Eds. Mark Levene and Alex Poulovassilis) Web Dynamics, (2004), 93-112.

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

[5] Bar-Ilan, J. Search engine results over time: A case study on search engine stability. Cybermetrics, 2/3(1), (1999), paper 1 http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html

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

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

[8] Bar-Ilan, J. Search engine ability to cope with the changing Web. In (Eds. Mark Levene and Alex Poulovassilis) Web Dynamics (2004), 195-218.

[9] Bar-Ilan, J. Expectation versus reality – Search engine features needed for Web research at mid-2005. Cybermetrics, 9(2005), paper 2,

http://www.cindoc.csic.es/cybermetrics/articles/v9i1p2.html

[10] Bar-Ilan, J., and Peritz, B. C. The life-span of a specific topic on the Web – The case of “informetrics”: a quantitative analysis. Scientometrics, 46 (1999), 371-382.

[11] Bar-Ilan, J., and Peritz, B. C. 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 (2004), 980-990.

[12] Bar-Yossef, Z., and Gurevich, M. Random sampling from a search engine’s index. In Proceedings of WWW2006 (2006), 367-376.

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

[14] Bharat, K. and Broder, A. A technique for measuring the relative size and overlap of public Web search engines. In Proceedings of the 7th International World Wide Web Conference, April 1998, Computer Networks and ISDN Systems, 30 (1998): 379-388.

[15] Broder A, Glassman S, Manasse M, Zweig G. Syntactic clustering of the Web. In Proceedings of the 6th International World Wide Web Conference, April 1997; 391–404.

[16] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. Graph structure in the Web. Proceedings of the 9th International World Wide Web Conference, May 2000. http://www9.org/w9cdrom/160/160.html

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

[18] Cho, J., and Garcia-Molina, H. The Evolution of the Web and implications for an incremental crawler. In Proceedings of 26th International Conference on Very Large Databases (VLDB), September 2000, 200-210.

[19] Dorogovtsev, S. N., Mendes, J. F. F. (2000). Scaling behaviour of developing and decaying networks. Europhysics Letters, 52, 33.

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

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

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

[23] Gomes, D., and Silva, M. J. Modeling information persistence on the Web. In Proceedings of the 6th International Conference on Web Engineering (ICWE06), (2006), 193-200.

[24] Gulli, A., and Signorini, A. The indexable Web is more than 11.5 billion pages. In Proceedings of 14th International World Wide Conference (2005), 902-903.

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

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

[27] Kim, S. J., and Lee, S. H. An empirical study on the change of Web pages. In Proceedings of APWeb 2005, LNCS 3399, (2005), 632 – 642.

[28] Koehler, W. A longitudinal study of Web pages continued: A report after six years. Information Research, 9(2) (2004) paper 174. http://InformationR.net/ir/9-2/paper174.html

[29] Kwon, S. Y., Lee, S. H., and Kim, S. J. A precise metric for measuring how much Web pages change. In DASFAA 2006, LNCS 3882 (2006) 557 – 571.

[30] Lawrence, S., and Giles, C. L. Searching the World Wide Web. Science, 280 (1998), 98-100.

[31] Lawrence, S., and Giles, C. L. Accessibility of information on the web. Nature, 400 (1999), 107-109.

[32] Mettrop, W., & Nieuwenhuysen, P. Internet search engines - fluctuations in document accessibility. Journal of Documentation, 57(5) (2001), 623-651.

[33] Mizzaro, S. How many relevances in information retrieval? Interacting With Computers, 10(1998), 305-322. http://www.dimi.uniud.it/mizzaro/research/papers/IwC.pdf

[34] Ntoulas, A., Cho, J., and Olston, C. What's new on the Web? The evolution of the Web from a search engine perspective. In Proceedings of the World-Wide Web Conference (WWW), May 2004, 1-12.

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

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

[37] Rousseau, R. Daily time series of common single word searches in AltaVista and Northern Light. Cybermetrics, 2/3(1), paper 2. (1999) http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

[38] Saracevic, T. Relevance reconsidered. In Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2), Copenhagen, Denmark (1996), pp. 201-218.

[39] Thelwall, M. Extracting accurate and complete results from search engines: Case study Windows Live. JASIST, 59(2008), 38-50.

[40] Thelwall, M. The responsiveness of search engine indexes. Cybermetrics, 5(1), (2001), paper 1 http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html

[41] Toyoda, M., and Kitsuregawa, M. What’s really new on the Web? Identifying new pages from a series of unstable web snapshots. In Proceedings of WWW2006 (2006), 233-241.