Jointly published by Elsevier Science Ltd, Oxford
and Akadémiai Kiadó, Budapest

Scientometrics,
Vol. 46, No. 3 (1999) 371–382

THE LIFE SPAN OF A SPECIFIC TOPIC ON THE WEB

THE CASE OF "INFORMETRICS": A QUANTITATIVE ANALYSIS

JUDIT BAR-ILAN, BLUMA C. PERITZ

School of Library, Archive and Information Studies, The Hebrew University of Jerusalem,
P.O. Box 1255, Jerusalem, 91904 (Israel)

(Received August 31, 1999)

In this case study a first attempt was made to explore data on the Web for a certain period of time by using bibliometric methods for analysis. The period under investigation was between January 3, 1998 and June 7, 1998. An additional search was carried out on June 20, 1999. The terms used were "informetrics or informetric". The results show that substantial changes occurred to the "literature on the Web" on informetrics during this period. Three specific trends were observed: some documents disappeared, new ones were added and some underwent changes.

Introduction

The Internet, the newly and quickly emerging information medium aims to serve some of the information needs of scientific communities. The databases, electronic journals and papers, announcements, discussions, pointers to further information are just some of the services the Internet provides. Bibliometric studies can investigate the effectiveness of these and other services and aspects of the Internet.

Bibliometric methods to study documents on the Internet were used before. *Woodruff* et al. (1996) examined 2.6 million Web documents for size, number and type of tags, attributes, file extensions, protocols and ports and the number of in-links. They also analyzed automatically 92,000 documents for html syntax errors. The documents were collected by the Inktomi crawler, developed at Berkeley, now the searching and crawling force behind a number of search engines, including Hotbot and Microsoft Search. *Larson* (1996) used AltaVista to carry out a cocitation analysis of a set of 34 Earth Science related URLs. He demonstrated that cocitation analysis can be applied to the Web. *Bar-Ilan* (1997) used bibliometric techniques to study the growth and characteristics of Usenet news messages during a hot period: the eruption of mad-cow

0138–9130/99/US $ 15.00

disease. She was also able to identify core newsgroups dealing with the subject. *Almind* and *Ingwersen* (1997) collected the Danish Web pages in December 1995. They found 47,000 URLs and analyzed these for Denmark's visibility on the Web. Samples were analyzed for the distribution of Danish Web pages on large centers of learning; document type and frequency distributions. The sample sizes ranged between 100 and 400. They too demonstrated the applicability of informetric techniques to the Web, and coined the term "Webometrics". *Rousseau* (1997) analyzed the domain distribution and the distribution of citations to the 340 URLs that were located by AltaVista as results of the query "bibliometrics OR informetrics OR scientometrics". He showed that both distributions fit the Lotka function. Other characteristics of these pages were also studied. *Ingwersen* (1998) investigated the feasibility and reliability of computing Web impact factors. *Bar-Ilan* (1998) studied the overlap, precision and estimated recall of some of the major search engines on a given query. In a recent work, carried out in parallel to this study, *Koehler* (1999) examined a set of "random" Web pages and sites for constancy and permanence.

In this paper we studied the evolution of Web documents on a specific topic over a period of time. The period under study was five months, between January 3rd, 1998 and June 7th, 1998. Some documents changed (in contents or in format), some disappeared, and new ones appeared. During this period, we searched for and collected documents containing the keywords "informetrics" or "informetric". The Web pages identified during this period were revisited a year later, on June 20, 1999. At this time the Web was searched again for additional documents pertinent to the topic.

We decided on "informetrics" since we wanted to study a topic highly related to our research interests. Initially we considered the query terms "bibliometrics OR scientometrics OR informetrics", but the number of documents found by the search engines was too large, well over a thousand for some of them. Most of the search engines do not display more than the first thousand or five hundred links for a query, thus we would not have been able to retrieve all of the documents for the extended query.

To illustrate the changes over time in the number of results reported by the search engine Alta Vista, consider the following: on May 14th, 1997, it retrieved 340 documents as a result of the query "bibliometrics OR scientometrics OR informetrics", as reported by *Rousseau*, while on December 9th, 1997, Alta Vista reported 1389 hits for the same query, and by June 22nd, 1998 already 1804 were found by Alta Vista alone. On contrary to our expectations, on June 20th, 1999 *only* 1261 hits were reported, while on July 22nd, 1999, the number of results reported by the search engine increased again to 1404.

We had to narrow our query, and decided on "informetrics" only, following *Brookes*' definition (1988), who stated that the term informetrics is a generic term that embraces both biblio- and scientometrics. We extended our initial query from "informetrics" to "informetric OR informetrics" since documents containing the phrases "informetric functions" or "informetric methods" are as relevant to the subject as documents in which only the word "informetrics" appears.

Methodology

The search terms used were "informetrics OR informetric". The searches were carried out six times during a five month period between January 3rd 1998 and June 7th 1998. A search was carried out on the first Sunday of each month (January, February, March, April, May and June). On each of these dates (called search rounds) we submitted the query to six of the major search engines on the Web (in alphabethical order):

- AltaVista (http://altavista.digital.com/cgi-bin/query?pg=aq)
- Excite (http://www.excite.com)
- Hotbot (http://www.hotbot.com)
- Infoseek (http://www.infoseek.com)
- Lycos (http://www.lycos.com/index.html)
- Northern Light (http://www.northernlight.com)

These were the largest search engines at the time of the study (*Sullivan*). The searches were carried out on several search engines in parallel in order to make the results as exhaustive as possible. It is well known that the search engines cover only some fraction of the Web, and the overlap between them is small (see for example: *Lawrence* and *Giles*, 1998; *Bharat* and *Broder*, 1998; *Bar-Ilan*, 1998).

First, the query results were saved. A search engine displays ten to one hundred links to documents on one result page. Since hundreds of relevant documents were found, we had to save tens of pages of search results for each search engine. Next, the URLs (the addresses of the Web documents) and the titles were filtered out from the pages returned by the search engines using a Visual Basic program. The display of the results varies from search engine to search engine, and sometimes even changed between the dates the searches were carried out. The output of this filtering process resulted in a table with columns for the URL and the title. These tables were loaded into Microsoft Excel. A Visual Basic module in Excel eliminated the duplicates: URLs returned by more than one search engine. From this list, using a Visual Basic program, we created html pages with links to these new URLs with 50 links on each page. In the last phase of each search round each of these links were retrieved and the documents were saved on our local hard disk. This phase was carried out by "brute force", i.e., by connecting to the Web on about ten computers concurrently from our computer lab, and saving documents in parallel on each of them. The whole process took about seven hours. A few URLs were unreachable during the collection phase, due to communication problems. In the two days following each collection phase, additional attempts were made to download the documents from these URLs. Note that in the context of this paper, identical documents are considered different if their URLs are different and the comparison between the URLs is case sensitive.

The analyses of the results were carried out by constructing frequency and cross tables and by utilizing the filtering tool of Microsoft Excel. The retrieved documents were manually examined for relevance. The comparison between the different versions, saved in different search rounds, was carried out by a program written in Visual Basic.

A year later, on June 20th, 1999, all the URLs identified during the initial search rounds were revisited, and an additional search round took place, using the methodology of the initial rounds. We called this last round the *comparison round.*

Results and discussion

Total retrievals

During the six search rounds, a total of 1268 different URLs were retrieved. Some of the URLs were not retrievable due to communication problems, while some other URLs from the search engines' results list did not exist. Still some more URLs did not contain the search terms, and in others the words informetrics or informetric did appear, but not in "our" context. The definition of a document belonging to "our" context is quite wide. Any document for which it is clear that the words "informetrics" or "informetric" appear in the context of scientometrics and bibliometrics, or even in the wider context of information science belongs to this category. Examples of such documents are pages mentioning the *International Society for Scientometrics and Informetrics,* its conferences, references to papers appearing in the proceedings of these conferences, links to and pages of the electronic journal of scientometrics, informetrics and bibliometrics, *Cybermetrics* (http://www.cindoc.csic.es/cybermetrics/), references to papers and abstracts on informetrics, pointers to and pages of the *World-Wide Web Virtual Library*'s page on informetrics, bibliometrics and scientometrics (http://crrm.univ-mrs.fr/vl/metricts.html), informetrics as a course in information science, pointers and pages of the Danish *Centre for Informetric*

Studies. We called the documents belonging to this category relevant. Table 1 displays the results.

Table 1
Total number of URLs, number of unretrieved URLs, number of non-existing URLs, number of URLs in which "informetrics" or "informetric" do not appear, number of URLs in which query terms do not appear in the context of scientometrics or bibliometrics, and number of relevant URLs per search round

Round – Date	No. of URLs	No. of URLs not retrieved	No. of non-existing URLs	No. of URLs in which the query terms do not appear	No. of URLs in which the query terms do not appear in "our" context	No. and % out of total no. of URLs in which the query term appears in "our" context
Round 1 – 4/Jan/1998	799	6	34	32	16	711 (89.0%)
Round 2 – 1/Feb/1998	810	3	44	30	17	716 (88.4%)
Round 3 – 1/Mar/1998	834	8	44	29	26	727 (87.2%)
Round 4 – 5/Apr/1998	869	7	59	28	26	749 (86.2%)
Round 5 – 3/May/1998	893	7	41	15	54	776 (86.9%)
Round 6 – 7/Jun/1998	942	12	53	11	57	809 (86.6%)
Total (URLs that in none of the rounds in which a search engine pointed to them, belonged to a higher category)*	1268	8	50	47	67	1096 (86.4%)

*The last row in the table displays the total number of different URLs in each category, and is *not* the sum of the column above it.

The definition of each category for a given search round is rather straightforward, but some difficulty arises when categorizing the set of total URLs the search engines pointed to during the six search rounds. Consider, for example a URL that the search engines pointed to it five times during the initial rounds, it was found to be in "our" context twice, in some other context twice, and we were unable to retrieve it once. To which category should such a document belong? We define that a URL is in "our" context if it was categorized as such in at least one of the search rounds. The reason for this decision is to be able to study the evolution of the largest possible set of relevant documents.

There is a natural order among the categories: not retrieved, not found, no informetric or informetrics, not in "our" context, in "our" context. When categorizing the set of total URLs, a URL belongs to a given category, if in none of the search round it was categorized as belonging to a higher category. This definition consistently extends the above definition of belonging to the set of relevant URLs.

The table shows slow, but monotonic growth in both the number of documents the search engines point to, and in the number of documents in the context of bibliometrics and scientometrics. The documents in which "informetric" or "informetrics" appear in a different context are mostly about commercial firms or businesses by these names, one of them is an ISP (Internet Service Provider).

On June 20th, 1999, 1147 different URLs were located by the search engines, out of which 1041 (90.8%) were categorized as relevant. Figure 1 illustrates the monotonic growth in the number of discovered URLs per search round.

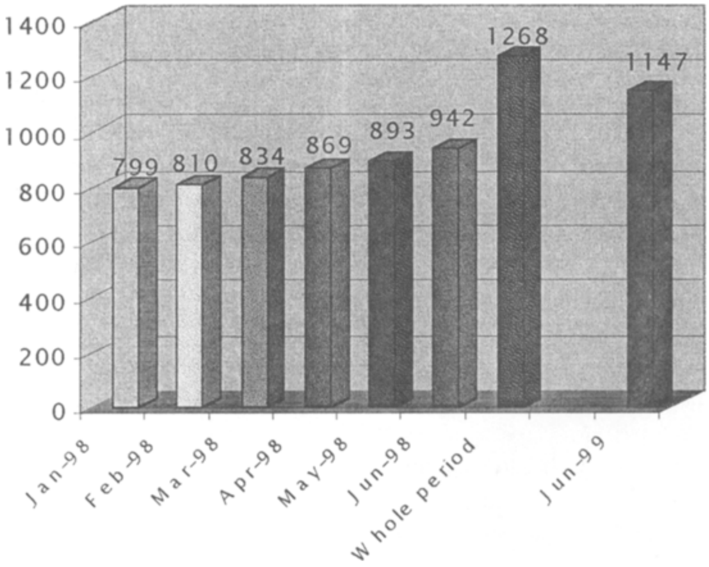

Fig. 1: Growth in the number of located URLs

As can be seen from Table 1 and Fig. 1, altogether 1268 URLs were identified during the initial search period. However, in any single round at most 942 URLs were located. This phenomenon can be explained by the fact that not only URLs are added to the Web, but also previously existing documents may either be totally removed from the Web or their content may not be relevant anymore to the query. One of the reasons for the removal of the documents is that they become outdated and may loose their relevance over time. Some of the newly discovered documents may have existed before, but they were simply moved to a new address (another URL). It may also be the case that documents existed previously on the Web but were discovered by the search engines only in one of the later rounds. To illustrate this, consider "Call for Papers for the Seventh ISSI Conference" which first appeared on the Web around mid-November 1997 (in which the word "informetrics" appears). It was retrieved for the first time only in the 5th round (in May 1998).

This behavior of appearance and disappearance was studied between consecutive search rounds, and the results are presented in Table 2. The percentages for documents that disappeared are calculated out of the total for the previous round, while the percentages for the newly added documents are calculated out of the total for the current round. Note that for the relevant URLs, the number of URLs in a given round minus the number of relevant URLs that disappeared plus the number of new relevant URLs do not add up to the number of relevant URLs in the next round, because a URL may be relevant in one of the rounds and non-relevant in the other.

Table 2
URLs appearing and disappearing in consecutive search rounds, in absolute numbers and percentages

	Newly added URLs in current round compared to previous round		URLs that disappeared in current round compared to previous round		Newly added relevant URLs in current round compared to previous round		Relevant URLs that disappeared in current round compared to previous round	
	no.	%	no.	%	no.	%	no.	%
Round 2	65	8.0	54	6.8	55	7.7	43	6.0
Round 3	131	15.7	107	13.2	102	14.0	84	11.7
Round 4	129	14.8	94	11.3	115	15.4	67	9.2
Round 5	156	17.5	132	15.2	113	14.6	87	11.6
Round 6	116	12.3	67	7.5	104	13.1	48	6.2

An additional factor must be taken into account when seeking explanations for the disappearance of documents from the Web: search engines are far from being perfect. The following examples illustrate this: four documents were retrieved in the first and sixth rounds only, three of them were exactly the same on both occasions and the fourth underwent some minor changes. Another five documents appeared in the first, fifth and sixth rounds only of which four have not changed at all, and the fifth changed only slightly in the last round. Why were these documents not retrieved by the search engines, have they disappeared and then reappeared? Or have the search engines dropped them? There are also quite a few documents that in the first rounds were retrieved by Alta Vista (for example), but in the later rounds were only found by Northern Light.

When comparing the list of URLs in the initial search period (1268 URLs) to the URLs located in the comparison round (1147 URLs) we observe the following: 607 "new" URLs were discovered, and 728 URLs "disappeared". When considering the relevant URLs only (1096 during the initial period, and 1041 in the comparison round), 601 "new" relevant URLs were discovered, while 598 URLs "disappeared".

In parallel to the comparison round we also revisited all 1268 URLs identified in the initial rounds. This time, 406 of them (31.7%) were not found or inaccessible. In a search round only 6.2% of the URLs were "not found" or inaccessible on the average. These results emphasize the dynamic nature of the Web.

Another interesting finding shows that the results displayed by the search engines are problematic. Out of the 1096 relevant URLs identified during the initial rounds, 745 were still relevant on June 20th, 1999, but *only* 440 of them were located by the search engines in the comparison round.

The "life span" of the documents

Next, we examined in how many search rounds a given URL appeared? The results are displayed in Table 3 and in Fig. 2.

Table 3
The number of search rounds in which a URL appears

Number of rounds	Number of URLs	% of total URLs	Number of URLs in "our" context	% of total URLs in "our" context
Once	193	15.2	152	13.9
Twice	184	14.6	136	12.4
Three times	131	10.4	109	10.0
Four times	147	11.6	114	10.4
Five times	73	5.6	67	6.1
Six times	540	42.6	518	47.2

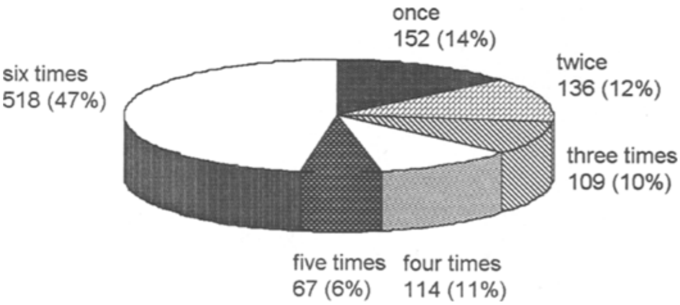

Fig. 2. The number of search rounds in which a relevant URL is retrieved

Note that the relevant documents seem to be a little more stable than the set of all documents. Documents from the set of all documents appeared in 3.8 search rounds on the average, while documents from the set of relevant documents appeared in 4.2 search rounds on the average.

The "stability" of the documents

Most of the URLs were retrieved in more than one search round. We examined whether the contents or the format of such documents changed from one search round to the other. The summarized results appear in Table 4.

Table 4
Stable versus changing documents in the set of all URLs and in the set of relevant URLs in absolute numbers and percentages

	Total URLs		Relevant URLs	
	No. of URLs	% of URLs	No. of URLs	% of URLs
Appeared in a single search round	193	15.2	152	13.9
Stable	638	50.3	567	51.7
Changed	437	34.5	377	34.4

The changes during the search rounds in 110 out of the 437 URLs were not further analyzed, because these URLs were not found, were non-retrievable or were retrieved erroneously in one or more of the rounds in which they were located. We examined the remaining 327 URLs (25.8% of total) in order to be able to characterize the changes that occurred in them. In the group of the relevant URLs we were able to further analyze 289 URLs (26.4% of the total number of relevant URLs).

We examined the changes that occurred to the documents between two different search rounds and characterized them as follows:

- lesser changes
  changes in 1-2 hypertext links, advertisements, copyright notices, small changes in format, change in date, changes in up to three words, changes to the html file that cannot be detected by the browser.
- considerable changes
  update (addition/removal of complete sentences, may include changes in formatting), major changes in look/format or appearance of a completely new document, changes in more than two hypertext links and multiple changes of any kind.

In order to summarize our findings, we defined two facets: type and stability. Type has two possible values:

- minor – each time the document changed, it underwent only lesser changes
- major – at least once the document underwent considerable changes

Stability has two possible values:

- stagnant – a document underwent changes less than 50% of the times it was compared
- dynamic – a document underwent changes 50% or more of the times it was compared

Note that if a URL was located in x rounds by the search engines, its contents was compared (x-1) times. These two facets give rise to four different combinations. The results appear in Table 5 for the total number of URLs that changed during the search period (327 URLs) and in Table 6 for the relevant URLs that changed during the search period (289 URLs).

Table 5
Characterization of the changed URLs in absolute numbers and percentages

	Stagnant		Dynamic		Total (type)	
Minor	32	(9.8%)	28	(8.6%)	60	(18.4%)
Major	109	(33.3%)	158	(48.3%)	267	(81.6%)
Total (stability)	141	(43.1%)	186	(56.9%)	327	(100%)

Table 6
Characterization of the changed relevant URLs in absolute numbers and percentages

	Stagnant		Dynamic		Total (type)	
Minor	27	(9.3%)	24	(8.3%)	51	(17.6%)
Major	104	(36.0%)	134	(46.4%)	238	(82.4%)
Total (stability)	131	(45.3%)	158	(54.7%)	289	(100%)

Most of the analyzed URLs, both in the set of the relevant URLs and the total URLs, have changed more than 50% of the time they appeared, and the changes can be characterized as major changes. Note that if a document underwent minor changes only it tends to be stagnant, while if major changes occurred to a document than the document is more likely to be dynamic.

Out of the total number of URLs (1268), 267 (21%) underwent major changes, 186 (14.7%) changed dynamically, and 158 (12.5%) underwent major, dynamic changes. When considering the relevant URLs only (1096 in total), the respective numbers are 238 (21.7%) – major changes, 158 (14.4%) – dynamic changes, and 134 (12.2%) both major and dynamic changes.

On June 20th, 1999 when the URLs identified in the initial rounds were revisited, out of the 1096 relevant URLs, 745 of them still existed on the Web, and 400 of them remained unchanged. These results suggest that Web documents containing the terms "informetrics" or "informetric" are relatively stable.

Conclusion

This paper describes a first attempt, known to us, to thoroughly study the dynamics of Web documents on a given topic over a period of time.

Each search round of our topic, informetrics, presented a slightly different picture of the subject on the Web. Three different trends that compliment each other were observed: documents disappear, others are added to the Web and are discovered by the search engines with time, and changes occur to existing individual documents. These changes are mainly comprised of updates.

The Web is a dynamic medium and one can never be sure whether information that existed on it one day would be there even in the near future, or would reappear in other forms.

We feel that informetrics is a rather stable topic on the Web, and suggest that for "hotter" topics (even if they are research/academic topics) more changes would occur. Since the Web is becoming more and more a main, maybe a first source of information, to confirm its reliability this type of research should be further investigated for other topics and fields.

*

We would like to thank Ms. Aviva *Joseph* for helping with the characterization of the changes that URLs had undergone and thanks to Tamar *Bar-Ilan* for helping with the data collection.

Bibliography

ALMIND, T. C., INGWERSEN, P. (1997), Informetric analyses on the World Wide Web: Methodological approaches to 'Webometrics'. *Journal of Documentation,* 53:404-426.

BAR-ILAN, J. (1997), The 'Mad Cow Disease', Usenet newsgroups and bibliometric laws. *Scientometrics,* 39:29-55.

BAR-ILAN, J. (1998), On the overlap, the precision and estimated recall of search engines. A case study of the query 'Erdos'. *Scientometrics,* 42:207-228.

BHARAT, K., BRODER, A. (1998), A technique for measuring the relative size and overlap of public Web search engines. *Proceedings of the 7th International World Wide Web Conference,* April 1998, *Computer Networks and ISDN Systems, 30,* 379-388.

BROOKES, B. C. (1988), Biblio-, Sciento-, Infor-metrics??? What are we talking about? In: L. EGGHE and R. ROUSSEAU Eds., *Informetrics 89/90*: 31-42. Amserdam: Elsevier.

INGWERSEN. P. (1998), The calculation of Web Impact Factors. *Journal of Documentation,* 54:236-243.

KOEHLER, W. (1999), An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science,* 50:162-180.

LARSON, R. (1996), Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. *ASIS96.* 1996. Online. Available:
http://sherlock.berkeley.edu/asis96/asis96.html. (Date of access: December 1997).

LAWRENCE, S., GILES, C. L. (1998), Accessibility and distribution of information on the Web. *Nature,* 400:107-109.

ROUSSEAU, R. (1997), Sitations: an exploratory study. *Cybermetrics* 1. Online. Available:
http://www.cindoc.es/cybermetrics/articles/v1i1p1.htm. (Date of access: November 1997).

SULLIVAN, D., Search engine watch. Online. Available: http://searchenginewatch.com/. (Date of access: June 1998).

WOODRUFF, A. et al. (1996), An investigation of documents from the World Wide Web. *Proceedings of the Fifth International World Wide Web Conference.* Paris, France, May 6-10, 1996: 963-980. Amsterdam: Elsevier.