Jointly published by Kluwer Academic Publishers, Dordrecht
and Akadémiai Kiadó, Budapest

Scientometrics,
Vol. 50, No. 1 (2001) 7–32

Data collection methods on the Web for informetric purposes – A review and analysis

JUDIT BAR-ILAN

School of Library, Archive and Information Studies, The Hebrew University of Jerusalem, Jerusalem (Israel)

We present different methods of data collection from the Web for informetric purposes. For each method, some studies utilizing it are reviewed, and advantages and shortcomings of each technique are discussed. The paper emphasizes that data collection must be carried out with great care. Since the Web changes constantly, the findings of any study are valid only in the time frame in which it was carried out, and are dependent on the quality of the data collection tools, which are usually not under the control of the researcher. At the current time, the quality and the reliability of most of the available search tools are not satisfactory, thus informetric analyses of the Web mainly serve as demonstrations of the applicability of informetric methods to this medium, and not as a means for obtaining definite conclusions. A possible solution is for the scientific world to develop its own search and data collection tools.

Introduction

Informetrics, as defined by *Brookes* (1990) covers and extends both scientometrics and bibliometrics. While bibliometric analysis is based mainly on printed data, informetrics studies the quantitative aspects of information processes in general (see also *Tague-Sutcliffe*, 1992).

The Internet and the Web are becoming major, emerging electronic information sources. Even though they were aimed in the beginning for academic use, in February 1999, only about 6% of the Web servers had scientific/educational content (as estimated by *Lawrence* and *Giles* (1999)). Today the percentage may even be lower, since the commercial sector grows faster than the scientific sector. Still, 6% of the then estimated 16 million Web servers is a lot, taking into account that these 16 million servers together held about 800 million indexable Web pages at the beginning of 1999. The Web is still growing at an enormous speed: the estimated number of indexable Web pages in August 2000 was 2200 million, with a monthly increase of 7 million Web pages (as reported by *Moore* and *Murray* (2000)). As *Lawrence and Giles* (1999) point

0138–9130/2001/US $ 15.00

out: “The Web contains a diverse range of scientific material, including scientist, university and project home pages, preprints, technical reports, conference and journal papers, teaching resources and databases… Much of this material is not available in traditional databases.” In addition, national and international agencies and institutions publish large amounts of reports, science and technology indicators and other relevant R&D data (as can be seen in *Aguillo* (1997) or *Bar-Ilan* (2000b)).

Thus, the influence of the Web on science and technology cannot and should not be ignored. Several studies have already analyzed different aspects of this new medium. The first step in such an analysis is data collection. Data collection from the Internet may seem to be trivial compared with data collection from printed sources. Even though there are several freely available options (e.g., search engines and directories), it turns out that this task is far more complex than it seems. In this paper we review several data collection methods and discuss their advantages and shortcomings. For each data collection method we refer to some examples of studies using this method. We cannot hope to be exhaustive in our review, since even though the Web is only ten years old, there is already a rather large corpus of informetric analyses of the Web. Extensive online bibliographies on the subject are maintained by *Aguillo* (n.d.) and by *Boudourides* (n.d.)

Data collection methods

Characteristics of the “whole database”

Analyzing the whole Web. Around 1996 it was still possible to study and to analyze large portions of the Web. The statement “Lycos claims to have indexed the 91 percent of the Web with a spider technology” appeared more than once in the literature (see for example *Wired Cybrarian* (1997)). *Tim Bray* (1996) was more careful, and stated that questions like “How big is the Web?” and “What is the ‘average page’ like?” are unanswered by his study, he derived only some approximations to the answers. He estimated the size of the Web at 11 million paged in November 1995, based on a large crawl of the Open Text search engine. Based on this set of pages the distributions of the size of the pages, image count, incoming and outgoing links are presented. The visibility of a site was measured by the number of URLs pointing to it, a clear analog of citation counts for identifying “hot papers” (*Garfield* and *Welljams-Dorof*, 1992). For “hot papers” only recent citations are counted, but in 1995 everything on the Web was “recent”. At that time, “the top sites are a list of well-known universities, organizations (CERN and the World-Wide Web Consortium), and a few companies. The only commercial sites which make the top-10 list … are Yahoo!, …, and Netscape". The Web has gone a long way in just five years; it has grown about 200 times and is currently dominated by commercial sites. Then, the top site, UIUC (University of Illinois at Urbana-Champagne) had about 80,000 URLs pointing to it, while Yahoo! had only around 35,000 URLs linking to it. Today (27th August, 2000), Alta Vista's new service, Raging Search (http://www.raging.com) reports 106,697 incoming links to the UIUC site, as a result of the query "link:www.uiuc.edu" and 1,307,656 links to Yahoo, as a result of the query "link:www.yahoo.com".

Woodruff et al. (1996) analyzed a 2.6 million sized data set resulting from experimental crawls of the search engine Inktomi in November 1995, then being developed at UC Berkeley. They presented results on document size, number and types of tags, attributes, file extensions, protocols, ports and number of incoming links. Rather interestingly, their results on the most visible sites were very different from those of *Bray*, even though both search engines crawled (visited URLs) the Web at approximately the same time. Out of the 22 sites in Bray's list and of the 19 sites of *Woodruff* et al., only four sites appeared in both of them.

Such studies, attempting to characterize the "whole Web" are not feasible any more. As we pointed out in the introduction the current estimated size of the indexable Web is around 2,200 million pages. What is meant by "indexable"? These are static pages, freely accessible by search engines and Web users (for a critical view on Web size estimates, see *Dahn* (2000)). There are lots of other pages on the Web, a large portion of them are dynamically created pages which are based on querying some database not residing on the Web and being accessed through search forms. This part of the Web is called the "invisible Web" and is estimated to be 500 times bigger than the indexable Web (*Bergman*, 2000). This estimate may sound incredible, but it contains the whole catalog of the Library of Congress (nearly 119 million items (*Library of Congress*, 2000)), the white and yellow pages of most countries and all the records of Medline (over 10 million records), (*PubMed*, 2000), just to mention a few.

The constantly changing Web. The dynamic nature of this new medium must also be taken into account when carrying out informetric analysis of the Web. Unlike printed documents, the main data source of traditional bibliometrics, Web documents are constantly changing in several ways. The content of the page changes, documents are often removed from the Web (they become outdated and are removed by the author/publisher of the page, or the author's site simply closes down, etc.); their URL changes (due to reorganization of the site, a new host is chosen for the page, etc.); or the documents are temporarily inaccessible (due to communication problems, or to problems with the server on which the page resides). As *Feldman* (1997) put it: "It was here a minute ago!" (Actually, the URL through which I used to access this article, is not available anymore, I had to search for a new reference, which hopefully still contains the above-mentioned article at the time of reading). Thus, when analyzing the Web, one can at best hope to analyze an instantaneous snapshot of it.

Several works studied the dynamics of the Web. *Koehler* (1999) defined two measures: constancy (the rate of change) and permanence (the probability that pages carry the same URL over time, or can be reached through a forwarding address) of Web pages. He monitored a set of 361 URLs using the *WebCrawler random URL generator* (http://www.webcrawler.com/cgi-bin/random), which randomly selected URLs from the Web Crawler database, whose estimated size as of February 2000 was 5-6 million pages (*Notess*, 2000). These URLs were studied in weekly intervals for a period of one year (1997). The results showed that almost without exception the Web documents have undergone changes during the study period, and by the end of the year about 75% of the sample could still be accessed at the same URL.

A different approach was taken by *Bar-Ilan and Peritz* (1999) to study the life span of documents on a given topic. The chosen topic was informetrics, and documents were collected once a month over a five months period using the six largest search engines at the time of the study (between January and June 1998). A follow-up round of data collection took place in June 1999. This setting allowed us to study three different trends: the changes occurring to the content of Web pages over time, the appearance of new pages on the topic and the disappearance of previously existing pages. More than 50% of the pages remained unchanged during the initial period. Of the 1096 "relevant" URLs (informetrics used in an information scientific sense) collected during the initial period, 745 (68%) existed at the time of the follow-up round, and 400 (36%) of them remained unchanged. During the follow-up round, the search engines discovered 607 additional URLs. The findings of this study suggest much higher stability than the results of *Koehler*, however one must take into account the different data sets: we chose a set of URLs on a well-defined scientific topic, whereas the set chosen in *Koehler*'s study was intended to be a random set of URLs (only after the set was picked, a breakdown according to the domain types was tabulated).

Brewington and *Cybenko* (2000) studied a set of more than two million Web pages specified by over 250,000 users. The data collection started in March 1999. Their aim was to get estimates on how often search engines have to reindex pages in order to remain current, that is to capture changes occurring to Web pages "soon" after they occur.

The above studies reflect on the dynamic nature of the Web, which has to be taken into account when collecting data for informetric analysis. If the collected data represents a given point in time, an effort must be made to collect the whole data set at more or less the same time. The exact date of the data collection must also be given, regretfully; this important detail does not always appear in the published studies.

Use of a single general search engine

Studies based on a single search engine. A popular method of data collection for informetric purposes is the utilization of a single, large search engine. AltaVista (http://www.altavista.com) is used most often, mainly because it has a large set of features, which are very useful for informetric analysis - it allows to limit searches to domains and hosts, to time periods, and it also allows to find out which and how many pages link to a given set of pages (the clear analog of citations in bibliometrics).

One of the first studies to utilize AltaVista's capabilities was *Larson*'s cocitation analysis of a set of Earth Science related Web sites (*Larson*, 1996). The starting points of the data collection were two authoritative Web sites in the field. The next step was to find the list of pages linking to these initial points. Links appearing at least three times on these pages were visited, and a set of highly relevant pages (judged by the author) was created. For every pair of items in this set, the number of pages that link to both pages was acquired using AltaVista's "link:" option (the query "link:www.microsoft.com" returns the list of pages linking to this site, "link:" can be used both for particular pages and for URLs containing the specified string, see *AltaVista* (2000)). Looking for links to particular pairs of pages (using the query "link:pageA AND link:pageB") is a clear analog of cocitation of journal articles. *Larson* used conventional methods (MDS) to create a cocitation map of Earth Science related Web sites, which "seem[ed] to produce quite clear, reasonable and interpretable results" (*Larson*, 1996).

Rousseau (1998) also utilized AltaVista's "link:" option to investigate the citation pattern (called "sitations" by him) of pages retrieved as a result of the query "bibliometrics OR informetrics OR scientometrics". He was able to show that both the distribution of pages belonging to different sites and the number of sitations can be very well described by Lotka's law.

Ingwersen (1998) introduced the concept of Web impact factors (Web-IFs) (number of Web pages pointing to a country or a site divided by the number of Web pages of a country or a site), and calculated it for several countries, for some top level Internet domains (.com, .gov, .org and .edu) and for some academic institutions. AltaVista was chosen as the search tool and he used Boolean combinations of the "link:", the "domain:" (finds pages in the specified domain) and the "host:" (finds pages on a specific computer) options.

Recently, *Smith* (1999) and *Thelwall* (2000) calculated the Web-IFs of additional institutions and countries, both of them used AltaVista as their data collection tool. *Thelwall* takes a critical view and explains the problems resulting from AltaVista's uneven coverage of Web pages. In his opinion the relatively high impact factor of Finland in *Ingwersen*'s study is an artifact of this uneven coverage. *Smith* believes that external citations are most indicative of the overall significance of a Web page. In a site often there are links from every page back to the home page (for administrative/navigational purposes), which result in a huge number of non-significant self-citations. When overall impact is computed, sites not using this method are penalized.

Leydesdorff and *Curran* (2000) also utilized the specific capabilities of AltaVista in their study of university-industry-government relations ("triple helix" relations) on the Internet for the case of Brazil and the Netherlands. They looked for occurrences and cooccurrences of the "helix" words: university, government and industry. Their findings show similar patterns of growth in the number of Web pages for both countries. As of November-December 1999 the number of Brazilian pages was growing at a higher rate. The searches were carried out both in English and in the natural languages. The query "university" retrieved the highest number of hits in all languages and countries.

Other studies examined a specific search tool. A series of studies analyzed large sets of queries submitted to the search engine Excite (http://www.excite.com). Excite provided large sets of real-life queries for the research. *Jansen* et al. (2000) analyzed the number of pages viewed in a query session, the use of relevance feedback, the number of search terms in a query, and derived a list of most popular search terms, based on more than 50,000 queries. They noted the surprisingly high number of incorrect uses and mistakes. *Ross* and *Wolfram* (2000) created a map of search topics based on cooccurrence of terms in queries using both cluster analysis and multi dimensional scaling, based on more than 350,000 queries submitted to Excite on a single day.

Limitations resulting from search engine coverage, stability and semantics. Using a single search engine for data collection has its problems. *Ross* and *Wolfram* point out that their study is representative of the queries posed by Excite users on a specific day, and no general conclusions can be drawn before studying additional data sets from other search engines. A well-known limitation is the number of pages indexed by the search engines (for current numbers reported by the engines, consult *Sullivan* (2000)). Several studies estimated the relative sizes of the major search engines (*Bharat* and *Broder*, 1998; *Lawrence* and *Giles*, 1998 and 1999). In the 1999 study, the estimated total number of indexable Web pages was 800 million as of February 1999, at which time the then largest search engine, Northern Light (http://www.northernlight.com) covered only 16% of these pages. Today's (August 2000) estimates of the size of the indexable Web are as high as 2,200 million pages. Lately some of the search engines put large efforts in increasing their indices. Google (http://www.google.com) claims to have covered more than 1,000 million URLs, while the search results are based on 560 million URLs. Inktomi (http://www.inktomi.com) has also started using an index based on 500 million pages. The search engines invest efforts in trying to keep up with the exponential growth of the Web. Currently they seem to be successful, but it remains to be seen whether their technology is able to scale up with future growth of the Web.

Using different search engines for data collection can lead to different results. Besides AltaVista, the search engine Hotbot (http://www.lycos.hotbot.com) also has an option to limit the search to domains and to find out which and how many pages link to a given URL or site (citations in our terminology, another popular Internet terminology is "backlinks"). *Snyder* and *Rosenbaum* (1999) compared the performance of Hotbot and AltaVista. They noticed huge discrepancies between the results; the results from Hotbot were about an order of magnitude larger than those from AltaVista (in July 1998). They have also shown some inconsistencies in the results of AltaVista. In their study they reported 24,492,674 pages from the domain .com to the domain .com for Hotbot and 1,118,829 pages for AltaVista. We carried out the same search in September 2000, and got 141,586,217 hits on AltaVista. We were unable to repeat the search on Hotbot, since the linkdomain modifier has been removed (as was already noted by *Snyder* and *Rosenbaum*). We only got results for the domain:com query: 44,342,500 hits on Hotbot versus 204,179,019 hits on AltaVista. Again an order of magnitude difference, but this time in favor of AltaVista. This is not surprising, the Web changes constantly, and so do the search engines, they rebuild their indices, and change their algorithms, syntax and semantics.

An additional problem is the stability of the search tools. *Rousseau* (1999) carried out three searches daily both on Northern Light and on AltaVista for a 21 weeks period between July and December 1999. For all three queries, Northern Light has exhibited linear growth, while the AltaVista results have shown rather large daily fluctuations. He recommends "not using AltaVista for informetric research, unless one needs a unique feature particular to this search engine". *Bar-Ilan* (2000c) compared the results of Hotbot with those of the advanced interface of Snap (http://www.snap.com/search/power/form/0,179,home-0,00.html?st.sn.srch.0.pwr) on twenty queries over a period of ten days during September-October 1999. These two search tools were compared, because they both draw their results from Inktomi's database, and thus similar behavior was expected from both of them. To our surprise, even though the results of Snap were very stable on all queries during the search period, huge fluctuations were recorded for Hotbot. On some days Hotbot retrieved a relatively small number of results on all queries, while on other days the number of results was much greater, the average daily fluctuation was 3.98, that is on a "good day", almost four times as many results were retrieved than on a "bad day".

Limiting the search results to specific time periods is very useful for informetric purposes, it allows us to study the growth of a topic over a period time, without monitoring this growth synchronously. Currently, two major search engines, AltaVista and Northern Light (http://www.northernlight.com) allow limiting the search to a time period defined by the user. The question is what is the meaning of the date of the document? For AltaVista the date is "the date each page found was last modified" (*AltaVista*, 2000). Northern Light does not give any explanation. *Sullivan* (n.d.) presents an estimate from 1998, stating that only 70 percent of the Web servers returned the correct date, while 20 percent reported the current date, regardless of when the page was created or changed, and the remaining 10 percent reported no date at all. In another page in the "Search Engine Watch" site *Sullivan* (1998) quotes Northern Light's Director of Engineering, Marc Krellenstein: "We were shocked by how many Web documents were dated in the future".

Obviously, *Leydesdorff* and *Curran* (2000) were aware of the limitations of both the Internet and the search tool (AltaVista) they were using: "In our opinion, the Internet should be considered an emerging phenomenon that is an algorithmic result of the geometrical representations. From this perspective, the Internet itself remains an hypothetical domain to be indicated by using one or more search engines. Reflexively, we can study the quality of the search engine, for example, by making comparisons among them. … However, the dynamic representation is from a hindsight perspective, while the Web crawler is continuously rewriting the representation. For this dynamic reason and for the already mentioned static problem of sampling from the population, we cannot claim validity for our inferences beyond the AltaVista domain." I would add to the limitation of the domain, also the date the searches were carried out (November-December, 1999), as is implied by their statement that the Web crawler is continuously rewriting the representation. If the estimates reported by *Sullivan* (n.d.) are correct, then 20 percent of the documents will automatically get the date the search engine visited it for the last time. It is not clear how the engines deal with documents without dates. From our observations it seems that AltaVista assigns them the date it visited the page for the last time, while Northern Light does not assign any date to them.

The Helix data revisited. To illustrate the problem of using different search engines, and the importance of the exact time the searches are carried out, we repeated some of the searches that were conducted by *Leydesdorff* and *Curran*. Our searches were carried out on the 2nd September, 2000 on AltaVista, Northern Light, Hotbot and Fast (http://www.alltheweb.com). Hotbot and Fast do not allow to limit dates, thus for these engines only the totals were compared. Estimates on the results appearing in the *Leydesdorff-Curran* paper (denoted Helix from this point on) were obtained from the graphs appearing in the paper.

In Table 1, we present data for the total number of pages in the three examined domains: .br (Brazil) and .nl (Netherlands) and the so-called gTLD (top level domains: .com, .edu, .org, .gov, .net and .mil). The search engine Northern Light is abbreviated as NL in the table.

Table 1
No. of URLs in the different domains, retrieved by different engines

Domain	Helix	AltaVista until 1998	NL until 1998	AltaVista, no date limitation	NL, no date limitation	Fast, no date limitation	Hotbot, no date limitation
Brazil	1,670,000	795,088	453,345	1,758,178	2,534,645	4,536,221	518,200
Netherlands	1,420,000	305,115	514,849	1,286,653	2,957,220	4,268,293	856,800
gTLD	No data	No data	No data	66,262,285	No data	245,143,143	70,550,200

Notice, that in this case, the Helix data is much nearer to the total number of pages (including pages dated 1999 and 2000) of AltaVista than to the number of pages dated before 1999 (which is the data presented in the Helix paper). AltaVista and Fast have indexed more documents from Brazil than from the Netherlands, while the opposite is true for Northern Light and Hotbot. If we consider the pages from the gTLD, we observe huge differences between AltaVista and Fast, even though their reported sizes (AltaVista indexed 350 million pages, while Fast indexed 340 as reported in June, 2000 by *Sullivan* (2000)) are comparable.

The stability of the results was checked a week later, all the search engines reported more or less the same numbers on all searches (+/- 10%), except for Hotbot. The results for Netherlands and Brazil were comparable, but for the gTLD, 220,136,300 hits were reported versus the 70,550,200 hits the week before. Inktomi, the search engine behind Hotbot, has started using a two layered search, consulting first its database of the most popular pages (110 million), and if not "enough" (not defined) results are found there, it turns to the larger database containing an additional 390 million pages (for an explanation, see *Notess* (2000b)). This might explain the fluctuations, however it is hard to see why over 70 million hits are not enough (in any case, Hotbot only displays the first 1000 hits).

Additional inconsistencies were noted for AltaVista, when limiting the gTLD pages to the English language, the approximate document count was 155,491,524 versus the unlimited case (66,262,285)!!!! This inconsistency was also reported by Notess (2000c). When searching for the data limited to calendar year (between 1993 and 2000), quite often the sum of the hits per year was greater than the total number of hits reported (for Brazil, the sum was 2,165,304 versus 1,758,178 when no date was given, nearly 25% difference (!)), while in other cases the sum was less than the total. For Northern Light the sum was always about 10% less than the total, which can be explained by the estimate that about 10% of the servers do not give dates.

In Figure1, we compare the original results of the searches for “university” and for “industry” with the results obtained from AltaVista and NorthernLight. A possible explanation for the decrease in the current year for AltaVista is that this year is not over yet. In the Helix data there were about four times as many hits for “university” dated 1998 compared to AltaVista, and almost three times as many as for Northern Light. This could be caused by frequent updates in pages in which the word “university” appears, thus they have fresher dates now. On the other hand, the total number of pages with the term “university” is almost the same for the Helix data and for AltaVista, while almost twice as many pages were found by Northern Light compared with the Helix data.

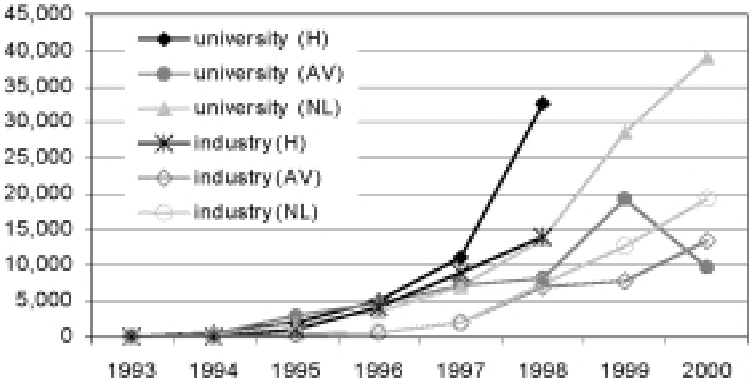

Fig. 1. “industry” and “university” in the Netherlands; any language

Figure 2 compares the Helix data and the AltaVista data for “industry AND government” (I-G), “university AND government” (U-G) and “university AND industry AND government” for gTLDs.

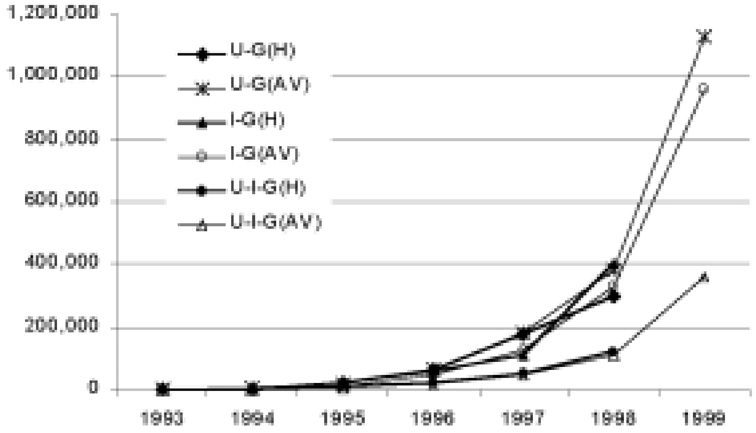

Fig. 2. gTLDs; some triple helix relations; any language

For the gTLDs the general trends are the same. *Leydesdorff* and *Curran* noted that “industry AND government” grew faster than “university AND government” in 1998. This trend was not reproduced in the current AltaVista data. Today’s results show that the “university AND government” relation is stronger than the “industry AND government” relation. This finding is also supported by Fast and Hotbot.

The results of our searches strengthen *Leydesdorff* and *Curran*’s statement, that the validity of their searches is limited to the AltaVista domain. Our data show that the time the searches were carried out is also crucial, since frequent changes occur even within the “AltaVista domain”.

Use of multiple general search engines

Studies using several search engines. We have already mentioned some studies (*Rousseau*, 1999; *Snyder* and *Rosenbaum,* 1999) utilizing two search engines, mainly for the sake of comparison. *Almind* and *Ingwersen* (1997) coined the term “webometrics” in their paper comparing the Danish portion of the Web to other Nordic countries. They used several search engines and the Nordic Web Index in order to increase the coverage.

Cronin et al. (1998) looked for Web pages mentioning five highly cited library and information science faculty using five search tools. They classify the retrieved pages according to eleven categories of invocation. The results are analyzed for each professor and each search tool separately, and are also combined for all the consulted search tools. They note and discuss the differences among the search engines. They state that they "were less concerned with exhaustivity…, than with deriving inductively a typology of invocations".

In the content analyses carried out by us we did try to be as exhaustive as (see *Bar-Ilan*, 1998b; 2000a and 2000b) possible, using freely available search tools. In these works our aim was to analyze the content of Web pages on a given topic, and to try to depict a picture of the kinds of information that can be found on the Web at the time of the study. In (*Bar-Ilan,* 1998b) we searched for "Erdos" (the world-famous mathematician, who passed away very near the time the searches were carried out), using seven of the largest search tools at that time. The searches were carried out several times during a two months period (November, 1996 - January, 1997), altogether 6681 different URLs were collected, out of which 2865 referred to the mathematician. These pages were classified. Even though, as expected, the largest category is related to his mathematical work, there was also great interest in the concept of Erdos numbers (for a definition, see *Grossman* and *Ion*, 2000)), and for a large set of "netizens", Erdos is simply the author of the saying: "A mathematician is a machine for turning coffee into theorems". In another study (*Bar-Ilan*, 2000a), we analyzed the content of Web pages containing the search term "informetrics". The searches were carried out in June 1998, using six of the largest search engines at that time. About 40% of the 807 URLs contained bibliographical references. The set of references was compared with the respective data in commercial bibliographical databases. Our set of references performed at least as well as the commercial databases (except for one case), indicating that valuable, free data is hidden in the Web waiting to be extracted from the millions of Web pages. Our most recent content analysis dealt with "S&T indicators" (*Bar-Ilan*, 2000b), where twelve search engines were queried in November, 1999. This study showed the high visibility of the European Union in this topic, and revealed that a large number of reports on the subject are published on the Web.

In order to gain an as accurate picture as possible in a content analysis of Web pages, it is important to try to cover large portions of the medium. Thus we used a number of search engines, however even with the combined results of the largest search engines, we are most probably very far from being exhaustive. It is definitely not enough to use the currently largest search engine, as it was shown in several works (see for example, *Bharat* and *Broder,* 1998; *Bar-Ilan,* 1998a; also implied in *Lawrence and Giles*, 1998 and 1999) that the overlap between the search engines is surprisingly small. Except for a very small core, each engine seems to be indexing a different part of the Web.

An example of search engine overlap. Here, we are going to illustrate the small overlap by searching for the term "webometrics" on today's largest search engines (search engines which indexed at least 100 million pages, as reported by *Sullivan,* 2000): Google, Webtop, AltaVista, Fast, Northern Light (http://www.nlresearch.com) and Inktomi (Iwon (http://www.iwon.com) was chosen as its representative, since other search tools that base their results on Inktomi: Hotbot, MSN search (www.search.msn.com) and Snap, retrieved only about half the number of URLs retrieved by Iwon). Webtop was discarded from our experiment, because it retrieved only 11 pages, compared to more than 35 pages retrieved by the other engines, even though Webtop (www.webtop.com) reportedly indexed 500 million pages. Similar findings are reported by *Notess* (2000b). Altogether, 308 different URLs were collected, out of which 285 (93%) contained the search term. Table 2 displays the number of URLs retrieved per search engine.

Table 2
No of URLs the retrieved for the query "webometrics" by search engine, total number, and "good URLs" – URLs containing the search term

Search engine	No. of URLs	% out of total number of URLs (308 URLs)	No. of "good URLs"	% out of total URLs number of good (285 URLs)
AltaVista	145	47%	139	49%
Excite	37	12%	32	11%
Fast	71	23%	64	22%
Google	135	44%	130	46%
Iwon	160	52%	158	55%
NorthernLight	68	22%	60	21%

Four engines, Hotbot, Iwon, MSN search and the advanced interface of Snap all draw their results from the Inktomi database. The results of Hotbot, MSN and Snap were very similar (all found around 70 URLs, but not exactly the same ones), Iwon retrieved a much larger set (160 URLs), there was a very large overlap, but still, four of the URLs appearing in the smaller sets were missing from the Iwon results. All of these four search tools are supposed to base their results on the new 500 millions database of Inktomi.

In the next stage, we compared the content of each of the 285 URLs with all the other URLs in order to identify content duplicates. Content duplicates appear sometimes on purpose, but most of the time they are a result of the naming system, which allows several different logical ways to name a URL (e.g.:
www.searchenginewatch.com
searchenginewatch.com
www.searchenginewatch.com/
www.searchenginewatch.com/index.html
searchenginewatch.com/index.html
all point to the same physical location). We combined all the content duplicates into a single entity and recorded the search engines, which retrieved at least one copy. The duplicate elimination resulted in 205 different documents, which contained the search terms. Figure 3 depicts the number of URLs retrieved by one, two, and up to six of the engines.

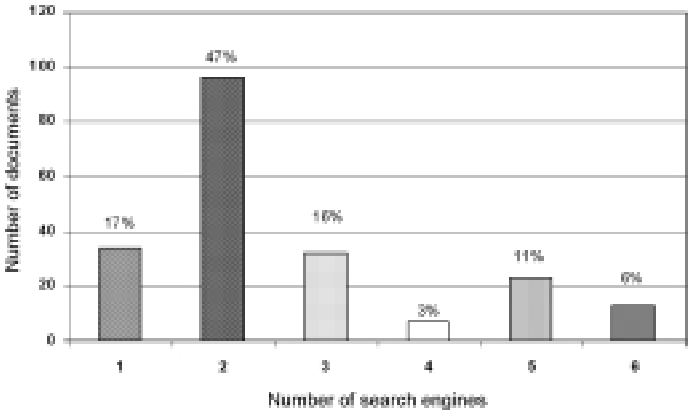

Fig. 3. Numbers and percentages of documents covered by: 1, 2, 3, 4, 5 and 6 search engines

Thirteen documents (6%) were retrieved by all six engines. Three pages are from the Cybermetrics site (http://www.cindoc.csic.es/cybermetrics/) including (*Aguillo*, n.d.), another three pages are about the Information Online and On Disc 99 Conference (http://www.csu.edu.au/special/online99/), where Alistair Smith gave a talk on “ANZAC webometrics”, two pages are from Alistair Smith’s site at the Victoria University of

Wellington in New-Zealand, the other pages are: the Webometrics bibliography (*Boudourides,* n.d.), Peter Ingwersen's homepage at the Royal School of Library and Information Science in Denmark (referring to the paper *Ingwersen*, (1998)), a page from the University of Western Ontario, mentioning a talk presented by J. Stephen Downie, the homepage of Don Turnball, a Ph.D student interested in webometrics at the University of Toronto, and a list from the Library of the University of Wales Aberystwyth (http://www.aber.ac.uk/~tplwww/e/pup_index.html), referring to the paper by *Almind* and *Ingwersen* (1997).

For the specific query, it would have been sufficient to use just three search engines: AltaVista, Google and Iwon, since together they retrieved 198 (97%) of the documents. The most productive domains were France, Spain, the United Kingdom and Denmark. 69% of the "good" URLs originated from these four European countries. A closer inspection of the pages from France showed that all except two of the pages are pages from Mannina Bruno's "Automated Research System" (AURESYS) project (http://ms161u06.u-3mrs.fr/hom.html). This system creates digital libraries automatically, and one of the examples is related to scientometrics and bibliometrics, thus these pages are either archived copies of the pages found on the Web (from the beginning of 1999, presumably) or give details on the collected pages, including keywords. Details of the project can be found in (*Mannina* and *Quoniam*, 2000). All 53 Spanish documents are pages from the *Cybermetrics* (e-journal) website (http://www.cindoc.csic.es/cybermetrics). However, only on six of the pages the term "webometrics" actually appears, in the rest of the pages it only appears as a content meta tag - hidden from the users view. It is well known that only some search engines index meta tags, our results show that among the engines checked, only AltaVista and Inktomi take meta tags into account.

Search engine reliability. It is difficult enough to cope with the dynamic nature of the Web, but unfortunately the search tools add an additional dimension of instability, which cannot be accounted for by the changes taking place on the Web. We have already mentioned the daily fluctuations in the search results of AltaVista (*Rousseau*, 1999) and Hotbot (*Bar-Ilan*, 2000c). While searching for informetrics once a month during a period of five months between January and June, 1998 using the six largest search engines at the time of the study (AltaVista, Excite, Hotbot, Infoseek, Lycos and Northern Light), we (*Bar-Ilan*, 1999) observed that the search engines lose information over time: relevant URLs that were retrieved by them at one time, are dropped from their list of URLs at a later time, even though these URLs continue to exist and to be of relevance. Sometimes, at an even later time, they once again include some of these "dropped" URLs in the list of results returned for the query. This problem was most acute for Excite and AltaVista. These search engines did become aware of the problem, and announced that they are taking steps to avoid it (*Sherman*, 1999). It remains to be seen whether they live up to their promise.

We have seen so far that the reliability of the existing freely available search tools is rather questionable. *Lawrence* and *Giles* (1999) point out that it may "not be economical for the engines to improve coverage or timeliness. The engines may be limited by the scalability of their indexing and retrieval technology, or by network bandwidth… Search engines might generate more revenue by putting resources into other areas (for example, free e-mail)." The informetrican is interested in the whole set of results for his query (or at least in the size of this set), whereas the average user only needs a few "most relevant" URLs. Thus the high popularity of the human edited general directory services (e.g. Yahoo (http://www.yahoo.com) or the Open Directory (http://www.dmoz.org)). These services cover at the present time at most two million URLs (*Sullivan*, 2000) out of the estimated 2200 million pages. Thus these services cannot serve as a means for extensive collection of data, however they may point to some authoritative sites, which can serve as starting points.

The natural question that arises is can we do any better? If you have ample financial, hardware, software and communication resources, you can:

Get your own crawler

The structure of the Web. Theoretically, powerful crawlers can cover the whole Web. In practice, however there are a few problems. The crawler starts at some initial URL (or at a list of URLs) and visits every page that is linked to these initial pages, then visits the links appearing in the new pages in the previous step and so on. Thus, not all the pages are visited at exactly the same time, which means that we do not get an exact snapshot of the Web at some specified time (some of the documents may undergo changes during the data collection period). A crawler using the above method is not able to cover the whole Web, since it will only be able to reach those Web pages that have a directed path of links from the set of initial pages. Still, a lot can be learnt about the Web from such experiments.

Bharat et al. (1998) built the so-called "Connectivity Server", based on a large (100 million URLs) crawl of AltaVista. This server enabled the researchers to visualize the neighborhood of a given set of pages, and to carry out experiments on the use of new retrieval strategies (e.g. *Bharat* and *Henzinger,* 1998; *Dean* and *Henzinger*, 1999). An improved version of the Connectivity Server (CS2) provided the data for a recent work describing the graph structure of the Web (*Broder* et al., 2000). Their findings are based on two crawls of AltaVista each with over 200 million pages and 1500 million links. The results indicate that the graph created from these huge crawls (which approximates the Web in May 1999) has a large strongly connected component, denoted SCC (i.e. there exists a directed path of links between any two nodes in this set), a set of nodes called IN, from which there is a directed path SCC, a set of nodes called OUT, for which there exits a directed path from SCC, "tubers", which connect IN with OUT without going through SCC, some tendrils in and out of IN and OUT leading nowhere and some disconnected components. The results of this work verify the results of previous studies by *Barabasi* and *Albert* (1999) and *Albert, Jeong* and *Barabasi* (1999), based on a 325 thousand node subset of the Web, showing that the distributions of both the number of links pointing to a page and the number of links outgoing from a page follow power laws with respective exponents. *Albert* et al. claim that the average diameter (number of links one needs to follow form any node *a* to reach any given node *b*) is about 20, while the *Broder* et al. found that the average diameter is only 16, when the average is over pairs of nodes for which such a path exits, however their observations show that only for 25% of the pairs of nodes there exists a directed path, thus most pairs are disconnected. The number of pages per site also follows a power law distribution, according to *Huberman* and *Adamic* (1999), based on large crawls of Alexa (http://www.alexa.com) and Infoseek (this service can be found today at http://www.go.com) covering ¼ and ½ million sites respectively. They found that Zipf like distributions do not fit (based on the rank of the site instead of its actual size). In contrast, *Rousseau* (1997) was able to show that the distribution of the number of pages retrieved as a result of the query "bibliometrics OR informetrics OR scientometrics" per site does follow a Lotka distribution. Note, that these two experiments have different settings (counting the total number of pages per site vs. counting only pages that are retrieved for a given query), different dates (1997 vs. 1999) and different sizes (38 sites vs. ½ million sites). On the other hand, the other finding of *Rousseau* on the number of links pointing to a site (again a Lotka distribution), was supported by the results of *Broder* et al.: they found that a Zipf distribution gave a better fit than the power law distribution.

Link analysis. The crawler can also assist the researchers in the study of the link structure of the Web. The analogy between citations in scientific papers and hypertext links is obvious. *Bray* (1996) used the number of links to a site as a measure of its visibility. *Carriere* and *Kazman* (1997) built a tool for ranking search based on the sum of the incoming and outgoing links of the documents. Incoming links correspond to citations, while outgoing links correspond to references.

In the academic world the impact of an article or of a journal (possibly the analogue of a Web site) is based on the number of citations it receives. Each citing article has exactly the same weight, as long as it is indexed by the Citation Index (otherwise it has no weight at all). This approach can be justified by the responsibilities of the authors, the referees and the editors to include only relevant references and to make an effort not to leave out references to major, relevant previous works. Still, there are differences between citations: it is more prestigious to receive a citation from a paper appearing in *Science,* for example, than to receive a citation from some low impact journal.

On the Web there are no rules, no one is responsible for anything, as *Chakrabarti* et al. (1999) put it: "the Web has evolved into a global mess of previously unimagined proportions". Almost anyone can publish Web pages on any subject. Many authors strive for attention, and will do anything to get higher rankings with the search engines. Some of them devise all sorts of tricks to appear higher in the search engines' lists on certain queries (this is called "spamming"). The search engines are in constant war with these spammers.

In traditional information retrieval systems the ranking is based on TF×IDF (term frequency times the inverse document frequency, see for example *Salton* (1989)). At first, most of the search engines based their rankings on some function of the term frequency, the inverse document frequency and some additional textual features like the location of the term in the document. Hypertext documents, are not just texts, they also include links to other documents. *Carriere* and *Kazman* (1997) proposed to rank search results in decreasing order of the sum of the incoming and outgoing links from a document. A simple count of incoming links cannot work on the Internet, since spammers can easily create dummy documents that point to the one they want to be highly ranked, or groups of people can decide to exchange links whether they are relevant or not. To overcome this problem, *Brin* and *Page* (1998) defined the PageRank measure, which gives weights to pages, based on the links pointing to them: the links are weighted according to the weight of the page they emanate from. The ranks can be calculated using an iterative algorithm. *Brin* and *Page*, then Ph.D students at Stanford had the necessary resources, created their own search engine, called Google, crawled 24 million pages and built a link graph containing 518 million links based on these 24 million documents. The PageRank for each of these pages was computed from this graph. They were able to demonstrate the usefulness of their ranking algorithm. Today, *Page* and *Brin* are the CEO and Presidents (respectively) of one of the most successful Web search engines, Google. Note, that the PageRank is computed for all the URLs in the database based on the link structure and regardless of a specific query.

It might easily happen that a page is very prestigious on a certain topic, but also happens to mention an additional topic in a superficial manner. This page would be ranked very high for the superficially mentioned topic, without any real justification.

The HITS algorithm (*Kleinberg*, 1998) starts with a set of documents that are potentially relevant to the query - documents retrieved by a standard search engine. This initial set is extended to the set of all pages pointing to and from it (with some restrictions). Then an iterative algorithm identifies the set of best *authorities,* authoritative documents representing a high-quality response to a broad user query, and the set of best *hubs*, link pages pointing to authorities. In the academic setting, authorities correspond to high impact articles, while hubs are less emphasized, they can be viewed as review articles or bibliographies or results of searches in bibliographic databases. The basic idea behind *Kleinberg*'s method is that: "Hubs and authorities exhibit what could be called a *mutually reinforcing relationship:* a good *hub* is a page that points to many authorities; a good *authority* is a page that is pointed to by many hubs". For a clear overview of the ideas, see *Chakrabarti* et al. (1999). *Kleinberg* was able to test his method on the IBM Research Intranet and an experimental system was based around this technique. Improvements to the basic algorithm were and are implemented on the IBM CLEVER Search Engine (*The Clever Project,* n.d.). One of the improvements is to incorporate link and content analysis: to give weight to the links according to how well the text in the vicinity of the link in the referring page matches the original query (*Chakrabarti* et al., 1998). Similar ideas, trying to fix problems with the basic algorithm resulting from a number of idiosyncrasies of the Web appear in (*Bharat* and *Henzinger*, 1998). Both these teams have access to the link structure of very large snapshots of the Web, which enables them to experiment with the different algorithms.

Actually, the idea to give weights to links is not so new, it appeared before both in the context of social networks (*Katz*, 1953) and in the context of citation analysis (*Pinski* and *Narin*, 1976). Katz, in his paper suggested "a new method of computing status, taking into account not only the number of direct 'votes' received by each individual but, also, the status of each individual who chooses the first, the status of each who chooses these in turn, etc.". *Pinski* and *Narin* developed a methodology for determining an influence weight for each journal. This measure takes into account not only the number of citations a journal received, but also the weight of the journals it receives the citations from. *Pinski* and *Narin* point out: "It seems more reasonable to give higher weight to a citation from a prestigious journal than to a citation from a peripherial one". The method of *Katz* for computing status, and of *Pinski* and *Narin* to compute influence weight are very similar to the methods based on *link analysis* on the Web (corresponding to citation analysis in the academic setting) developed by *Page* and *Brin* and by *Kleinberg*.

Additional techniques

When the researcher knows where to look for the data on the Web, there is no need for search tools or crawlers. *Rosenbaum* (1998) used this technique when he analyzed the content of community network sites in Indiana. *Aguillo* and *Pareja* (2000) studied the structure of the Research Councils of four European countries, again by going to the appropriate sites directly. Complementary data was obtained from commercial search engines. *Cui* (1999) used the list of the twenty-five top US medical schools as a starting point for rating health Web sites.

A hybrid approach for collecting computer science papers is being used in the *ResearchIndex (www.researchindex.com)* project. This project aims to automatically create citation databases similar to the ISI Citation Indexes. The system contains full text articles that are linked to the cited works, it is also possible to look for citations of authors or papers. The data collection is carried out by using commercial search engines, by monitoring mailing lists and newsgroups, by visiting sites of computer science departments and by employing other heuristics (*Lawrence* et al., 1999). The rationale for their project is, that large amounts of scientific literature is available on the Web, but in a very disorganized fashion; finding articles is currently difficult, because Web search engines have difficulty in keeping up to date and they do not index contents in PostScript and PDF formats. Other projects (e.g.; the ArXiv.org (http://xxx.lanl.gov/help/general), a preprint service in physics and computer science; or DoIS (http://dois.mimas.ac.uk), a new initiative in information science) do not provide citation data, but create large free, online databases of scientific literature. These databases may serve as data collection tools, after the researcher inspects their coverage and quality. Recently, *Aguillo* (2000-b) sent a message to the SIGMETRICS discussion list (http://web.utk.edu/~gwhitney/sigmetrics.html) offering to archive scientometrics related papers at the Cybermetrics site.

One of the more promising, new developments is to crawl the Web in a *focused* manner, collecting only documents that are relevant to a pre-defined set of topics (*Chakrabarti* et al., 1999). Focused crawlers allow the creation of large *vertical search engines* that concentrate on specific topics. Because of the size of the Web, human created lists of relevant sites cannot be as comprehensive as the ones collecting information through focused crawling. There are also efforts to automatically classify these documents into subject categories, which may lead to create much more extensive directories than the currently existing "hand-made" ones (like Yahoo and Open Directory), while achieving similar quality (*McCallum* et al., 2000). When such tools become widely available, they can serve as data collection tools for informetric purposes.

Aguillo (2000-a) advocates the use of client-side based, low cost (shareware) programs for data collection, tracing, indexing, analysis and visualization.

Conclusion

Commercial search engines have their own priorities, which are not always the same as those of scientists. They aim to provide extremely fast services, and sometimes quality is sacrificed for speed. It is impossible (costs, time and resources) and not even desirable (because of the congestion it would cause in the communication channels) for every researcher to have his own crawler.

A possible solution to the data collection problem is for the scientific world to create its own search and data collection tools. Scientific content on the Web is estimated to be less than 6% (of the sites, there is no estimate on the number of pages), thus the total information in a specific scientific discipline can probably be handled. We propose to create vertical search engines and directories per disciplines with high quality control. New technologies (focused crawlers), human monitoring and submission of sites should be incorporated. The quality of the service should be controlled by experts in the specific discipline (along the ideas of the Open Directory, http://www.dmoz.org), who would serve as "editors" of the tool. Resources could be provided by scientific institutions, foundations and research councils. Easy interaction between the search tools in the different disciplines should be provided.

Until these high quality tools become a reality, we can and should continue to experiment with informetric techniques and methods, and try to develop new ones, while always keeping in mind that the validity of the results is influenced by the currently available data collection methods.

*

Many thanks to Bluma *Peritz* for the very useful discussions and for her helpful remarks on the drafts of this paper.

References

AGUILLO, I. F. (1997). STM information on the Web and development of New Internet R&D databases and indicators. *Online Information 97 Proceedings*, 239–243.

AGUILLO, I. F. (2000-a). A new generation of tools for search, recovery and quality evaluation of World Wide Web medical resources. *Online Information Review,* 24 (2); 138-143.

AGUILLO, I. F. (2000-b). Mirroring individual scientometric contributions in the Cybermetrics site.
In *SIGMETRICS discussion list*. [Online].
Available: http://listserv.utk.edu/cgi-bin/wa?A1=ind0009&L=sigmetrics (September 2000).

AGUILLO, I. F. (no date). *Cybermetrics. Papers and Abstracts*. [Online].
Available: http://www.cindoc.csic.es/cybermetrics/links03.html (September 2000).

AGUILLO, I. F., PAREJA, V. M. (2000). Indicators of the Internet presence of the Western European Research Councils. *Poster Presentation in* S&T 2000, Leiden, May 2000 [Online].
Available: http://sahara.fsw.leidenuniv.nl/cwts/abs/AGUILLO.txt (September 2000).

ALMIND, T. C. & INGWERSEN, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to 'Webometrics'. *Journal of Documentation*, 53 (4), 404–426.

ALBERT, R., JEONG, H., BARABASI, A. L. (1999). Diameter of the World Wide Web. *Natur*e, 401: 130–131.

ALTAVISTA (2000*). Advanced Search Tutorial.* [Online].
Available: http://doc.altavista.com/adv_search/ast_i_index.html (September 2000).

BARABASI, A. L., ALBERT, R. (1999). Emergence of scaling in random networks. *Science*, 286 (5439): 509-512.

BAR-ILAN, J. (1998a). On the overlap, the precision and estimated recall of search engines – A case study of the query 'Erdos'. *Scientometrics* , 42 (2): 207–228.

BAR-ILAN, J. (1998b). The mathematician, Paul Erdos (1913-1996) in the eyes of the Internet. *Scientometrics*, 43 (2): 257–267.

BAR-ILAN, J. (1999). Search engine results over time – A case study on search engine stability. *Cybermetrics*, 2/3(1), paper 1. [Online].
Available: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html (September 2000).

BAR-ILAN, J. (2000a) The Web as information source on informetrics? – A content analysis. *JASIS*, 51 (5): 432-443.

BAR-ILAN, J. (2000b). Results of an extensive search for "S&T indicators" on the Web – A content analysis. *Scientometrics,* 49 (2): 257–277.

BAR-ILAN, J. (2000c). Evaluating the stability of the search tools Hotbot and Snap: A case study. *Online Information Review*, 24(6).

BAR-ILAN, J., PERITZ B. C. (1999). The life span of a specific topic on the Web; the case of 'Iformetrics' a quantitative analysis. *Scientometrics*, 46 (3): 371–382.

BERGMAN, M. K. (2000). *White Paper – The Deep Web: Surfacing Hidden Value*. [Online].
Available: http://128.121.227.57.download/deepwebwhitepaper.pdf (August 2000).

BHARAT, K., BRODER, A. (1998). 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: 379-388 [Also online].
Available: http://decweb.ethz.ch/WWW7/1937/com1937.htm (September 2000).

BHARAT, K., BRODER, A., HENZINGER, M., KUMAR, P., VENKATASUBRAMANIAN, S. (1998). The connectivity server: Fast access to linkage information on the Web. In: *Proceedings of the 7th International World Wide Web Conference*, April 1998, *Computer Networks and ISDN Systems*, 30: 469-477 [Also online].
Available: http://www7.scu.edu.au/programme/fullpapers/1938/com1938.htm (September 2000)

BHARAT, K., HENZINGER, M. (1998). Improved algorithms for topic distillation in a hypertext environment. In: *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval,* August, 1998, 104–111.

BOUDOURIDES, M. A. (no date). *Webometrics and Organizations.* [Online].
Available: http://hyperion.math.upatras.gr/weborg/ (September, 2000).

BRAY, T. (1996). Measuring the Web. In *Proceedings of the 5th International World Wide Web Conference*, May 1996, *Computer Networks and ISDN Systems*, 28, 993-1005. [Also online]
Available: http://www5conf.inria.fr/fich_html/papers/P9/Overview.html (September 2000).

BREWINGTON, B. E., CYBENKO, G. (2000). How dynamic is the Web? In *Proceedings of the 9th International World Wide Web Conference*, May 2000, *Computer Networks and ISDN Systems*, 33, 257-276. [Also online]
Available: http://www9.org/w9cdrom/264/264.html (August 2000).

BRIN, S., PAGE, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In: *Proceedings of the 7th International World Wide Web Conference*, April 1998, *Computer Networks and ISDN Systems*, 30: 107 - 117. [Also online]
Available: http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm (September 2000).

BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN. P., RAJAGOPALAN, S., STATA, R., TOMLINS, A., WIENER, J. (2000). Graph structure in the Web. In: *Proceedings of the 9th International World Wide Web Conference*, May 2000, *Computer Networks and ISDN Systems, 33:* 309-320. [Also online].
Available: http://www9.org/w9cdrom/160/160.html (September 2000).

BROOKES, B. C. (1990). Biblio-, sciento-, infor-metrics??? What are we talking about? In L. EGGHE and R. ROUSSEAU (Eds), *Informetrics 89/90,* 31-42. Amsterdam: Elsevier.

CARRIERE, J., KAZMAN, R. (1997). WebQuery: Searching and visualiznig the Web through connectivity. In: *Proceedings of the 6th International World Wide Web Conference*, May 1997, 701-711. [Also online].
Available: http://www.cgl.uwaterloo.ca/Projects/Vanish/webquery-1.html (September 2000).

CHAKRABARTI, S., DOM B., KUMAR, R. S., RAGHAVAN, P., RAJAGOPALAN, S., TOMKINS, A., KLEINBERG, J. M., GIBSON, D. (1999). Hypersearching the Web. *Scientific American,* 280(6): 54-60. [Also online].
Available: http://www.sciam.com/1999/0699issue/0699raghavan.html (September 2000).

CHAKRABARTI, S., DOM B., RAGHAVAN, P., RAJAGOPALAN, S., GIBSON, D., KLEINBERG, J. M. (1998). Automatic Resource Compliation by Analyzing Hyperlink Structure and Assoicated Text. In: *Proceedings of the 7th International World Wide Web Conference*, April 1998, *Computer Networks and ISDN Systems*, 30: 65-74 [Also online].
Available: http://decweb.ethz.ch/WWW7/1898/com1898.htm (September 2000).

CHAKRABARTI, S., VAN DEN BERG, M., DOM, B. (1999). Focused crawling: A new approach to topic-specific Web resource discovery. In: *Proceedings of the 8th International World Wide Web Conference*, May 1999, 545-562. [Also online].
Available: http://www8.org/w8-papers/5a-search-query/crawling/index.html (September 2000).

The Clever Project. (no date). [Online].
Available: http://www.almaden.ibm.com/cs/k53/clever.html (September 2000).

CRONIN, B., SNYDER, H. W., ROSENBAUM, H., MARTINSON, A., CALLAHAN, E. (1998). Invoked on the Web. *Journal of American Society for Information Science,* 49 (14): 1319–1328.

CUI, L. (1999). Rating health Web sites using the principles of citation analysis: A bibliometric approach. *Journal of Medical Internet Research*, 1(1): e4. [Online].
Available: http://www.jmir.org/1999/1/e4/index.htm (September 2000).

DAHN, M. (2000).Counting angels on a pinhead: Critically interpreting Web size estimates. *Online* 24 (1): 35–40. [Also online].
Available: http://www.onlineinc.com/onlinemag/OL2000/dahn1.html (August 2000).

DEAN, J., HENZINGER, M. (1999). Finding related pages in the World Wide Web. In: *Proceedings of the 8th International World Wide Web Conference*, May 1999, 389-401. [Also online].
Available: http://www8.org/w8-papers/4a-search-mining/finding/finding.html

FELDMAN, S. (1997). 'It Was Here a Minute Ago!': Archiving the Net. *Searcher*, 5 (9), 52. [Also Online].
Available: http://www.info-sec.com/internet/internet_120397a.html-ssi (August 2000).

GARFIELD, E., WELLJAMS-DOROF, A. (1992). Citation data: Their use as quantitative indicators for science and technology evaluation and policy making. *Science & Public Policy,* 19 (5): 321–327. [Also online].
Available: http://www.garfield.library.upenn.edu/papers/sciandpubpolv19(5)p132y1992.html
(August 2000).

GROSSMAN, J. W., ION, P. D. F. (2000). *The Erdos Number Project.* [Online].
Available: http://www.oakland.edu/~grossman/erdoshp.html. (September 2000).

HUBERMAN, B. A., ADAMIC, L. A. (1999). Growth dynamics of the World-Wide Web, *Nature*, 401, 131.

INGWERSEN. P. (1998). The calculation of Web impact factors. *Journal of Documentation*, 4(2): 236-243.

JANSEN, B. J., SPINK, A., SARACEVIC, T. (2000). Real life, real users and real needs: A study and analysis of user queries on the Web. *Information Processing and Management*, 36: 207–227.

KATZ, L. (1953). A new status index derived from sociometric analysis. *Psycometrika.* 18 (1): 39–43.

KLEINBERG, J. M. (1998). Authoritative sources in a hyperlinked environment. In: *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.* Also appeared in: *Journal of the ACM,* 46(5): 604–632, 1999. [Also online].
Available: http://www.cs.cornell.edu/home/kleinber/auth.ps (September 2000).

KOEHLER, W. (1999). An analysis of Web page and Web site constancy and permanence. *Journal of the American Society for Information Science,* 50 (2): 162–180.

LARSON, R. (1966). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. *ASIS96.* [Online].
Available: http://sherlock.berkeley.edu/asis96/asis96.html (September 2000).

LAWRENCE, S., GILES, C. L. (1998). Searching the World Wide Web. *Science*, 280, 98–100.

LAWRENCE, S., GILES, C. L. (1999). Accessibility and distribution of information on the Web. *Nature*, 400: 107–110.

LAWRENCE, S., BOLLACKER, K., GILES, C. L. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer,* 32 (6): 67–71.

LEYDESDORFF, L., CURRAN, M. (2000). Mapping university-industry-government relations on the Internet: The construction of indicators for a knowledge-based economy. *Cybermetrics,* 4 (1), paper 2. [Online]. Available: http://www.cindoc.csis.es/cybermetrics/articles/v4i1p2.html (August 2000).

THE LIBRARY OF CONGRESS (2000). *Facsinating Facts about the Library of Congress.* [Online]. Available: http://www.loc.gov/today/fascinate.html (August 2000).

MANNINA, B., QUONIAM, L. (2000) How to hold a virtual library active? *Cybermetics*, 4 (1), paper 1. [Online]. Available: http://www.cindoc.csic.es/cybermetrics/articles/v4i1p1.html (September 2000)

MCCALLUM, A. K., NIGAM, K., RENNIE, J., SEYMORE, K. (2000). Automating the construction of Internet portals with machine learning. *Information Retrieval,* 3, 127–163.

MOORE, A., MURRAY, B. H. (2000). *Sizing the Internet*. [Online]. Available: http://www.cyveillance.com/resources/7921S_Sizing_the_Internet.pdf (August 2000).

NOTESS, G. R. (2000). *Search Engine Statistics: Database Total Size Estimates*. [Online]. Available: http://www.searchengineshowdown.com/stats/0002sizeest.shtml (August 2000).

NOTESS, G. R. (2000b). *The Half Billion Crew: Google, Inktomi GEN3 & Webtop*. [Online]. Available: http://www.searchengineshowdown.com/stats/500million.html (September, 2000).

NOTESS, G. R. (2000c). *Inconsistencies Reports*. [Online]. Available: http://www.searchengineshowdown.com/inconsistent.shtml (September 2000).

PINSKI, G., NARIN, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing and Management,* 12: 297-312.

PUBMED OVERVIEW (2000). [Online]. Available: http://www.ncbi.nlm.nih.gov:80/entrez/query/stattic/overview.html (August 2000).

ROSENBAUM, H. (1998). Web-based community networks: A study of information organization and access. In: *ASIS'98 Contributed Papers,* 516–530.

ROSS, N. C. M., WOLFRAM, D. (2000). End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. *Journal of the American Society for Information Science,* 51 (10): 949–958.

ROUSSEAU, R. (1997). Sitations: An exploratory study. *Cybermetrics*, 1 (1), [Online]. Available: http://www.cindoc.es/cybermetrics/articles/v1i1p1.htm (August 2000).

ROUSSEAU, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics*, 2/3 (1), paper 2, [Online]. Available: http://www.cindoc.csis.es/cybermetricc/articles/v2i1p2.html (August 2000).

SALTON, G. (1989). *Automatic Text Processing.* Addison-Wesley, Reading, MA.

SHERMAN, C. (1999). The search engines speak. In *Web Search.* [Online]. Avaliable: http://websearch.about.com/internet/websearch/library/weekly/aa120399.htm (September 2000).

SMITH, A. G. (1999). A tale of two Web spaces: Comparing sites using Web impact factors. *Journal of Documentation*, 55 (5): 577–592.

SNYDER, H., ROSENBAUM, H. (1999). Can search engines be used as tools for Web-link analysis? A critical view. *Journal of Documentation*, 55 (4): 375–384

SULLIVAN, D. (1998). Northern light adds search functions, freshens index. In *SearchEngineWatch.* [Online]. Available: http://www.searchenginewatch.internet.com/sereport/98/08northernlight.html (September 2000).

SULLIVAN, D. (2000). Search engine sizes. In: *SearchEngineWatch.* [Online]. Available: http://www.searchenginewatch.com/reports.sizes.html (August 2000).

SULLIVAN, D. (no date). Search assistance features. In: *SearchEngineWatch*. [Online]. Available: http://searchenginewatch.internet.com/facts/assistance.html (September 2000).

TAGUE-SUTCLIFFE, J. (1992). An introduction to informetrics. *Information Processing and Management*, 28 (1): 1–3.

THELWALL, M. (2000). Web impact factors and search engine coverage. *Journal of Documentation*, 56 (2): 185–189.

WIRED CYBRARIAN. (1997). [Online]. Available: http://hotwired.lycos.com/cybrarian/ reference/search.html (August 2000).

WOODRUFF A., AOKI, P. M., BREWER E., GAUTHIER P., ROWE, L. A. (1996). An investigation of documents from the World Wide Web. In *Proceedings of the 5th International World Wide Web Conference, May 1996, Computer Networks and ISDN Systems,* 28: 963-980. [Also online]. Available: http://www5conf.inria.fr/fich_html/papers/P7/Overview.html (September 2000).

Received October 12, 2000.

Address for correspondence:
JUDIT BAR-ILAN
School of Library, Archive and Information Studies,
The Hebrew University of Jerusalem,
P. O. Box 1255, Jerusalem, 91904 (Israel)
E-mail: judit@cc.huji.ac.il