Informetric Theories and Methods for Exploring the Internet: An Analytical Survey of Recent Research Literature

JUDIT BAR-ILAN AND BLUMA C. PERITZ

ABSTRACT

THE INTERNET, AND MORE SPECIFICALLY the World Wide Web, is quickly becoming one of our main information sources. Systematic evaluation and analysis can help us understand how this medium works, grows, and changes, and how it influences our lives and research. New approaches in informetrics can provide an appropriate means towards achieving the above goals, and towards establishing a sound theory. This paper presents a selective review of research based on the Internet, using bibliometric and informetric methods and tools. Some of these studies clearly show the applicability of bibliometric laws to the Internet, while others establish new definitions and methods based on the respective definitions for printed sources. Both informetrics and Internet research can gain from these additional methods.

INTRODUCTION

Tague-Sutcliffe (1992) defined *Informetrics* as "the study of the quantitative aspects of information in any form . . . and in any social group," and Brookes (1990) characterized it as "a generic term that embraces both biblio- and scientometrics." Along the lines of Tague-Sutcliffe, informetrics investigates: Characteristics and measurements of persons, groups, institutions, countries; publications and information sources; disciplines and fields; and information retrieval processes.

When the above definitions were offered, the World Wide Web was still non-existent, but today it is quickly becoming a major information source. Informetric methods can be and are applied to the Web, and new methods are being developed for this medium. This paper presents a selective

Judit Bar-Ilan and Bluma C. Peritz, School of Library, Archive, and Information Studies, The Hebrew University of Jerusalem, P.O. Box 1255, Jerusalem, 91904, Israel
LIBRARY TRENDS, Vol. 50, No. 3, Winter 2002, pp. 371–392

review of research based on the Internet, using bibliometric and informetric methods and tools. The review is organized according to the following methods:

- Data collection methods
- Informetric analysis
- Citation analysis
- Cocitation and coword analysis
- Content analysis
- Evaluation using existing/new measures
- Identifying and calculating indicators
- Models
- Fitting existing models and bibliometric laws

Data Collection Methods

Data collection from the Web is far from trivial, due to its size and its extremely dynamic nature. There are no methods to enumerate the "whole Web" (the total population under study) or to enable us to get a truly random sample of Web pages. When studying Web documents, sites, or the structure of parts of the Web, data collection is often carried out using the currently existing information retrieval tools, mainly the search engines, which are far from perfect. Bar-Ilan (2000a) discusses problems related to this type of data collection. Use studies rely mainly on surveys, interviews, and log analysis.

Surveys

Surveys on the Internet are employed mainly to receive information on the use of technology. Savolainen (1998) analyzes use studies of electronic networks. A considerable number of the reviewed studies collect data through quantitative surveys. Questionnaires can be sent out by regular mail or e-mail, can be filled out on Web pages, or can use a combination of these methods.

Lazinger, Bar-Ilan, & Peritz (1997) carried out an extensive survey on Internet use of the faculty members of Hebrew University of Jerusalem. The questionnaire was sent out by regular mail, in order to reach also faculty members who did not use e-mail. A follow up was sent to non-respondents. The overall response rate was 59.4 percent. More than 80 percent of the respondents used some Internet services, with e-mail being the most popular one (the questionnaires were sent in spring 1995, when the graphic browsers to the Web were just being introduced [*Life on the Internet,* n. d.]). Significant differences were found in the use patterns between the Humanities and Social Sciences faculty and the Science and Agriculture faculty.

Kovacs, Robinson, & Dixon (1995) investigated the use of discussion lists by library and information science professionals. The questionnaire was sent out to the participants of fifty-seven library and information science related discussion groups—approximately 10,000 participants. Filled out questionnaires were returned by e-mail. Only 576 responses were received. The majority of these respondents stated that discussion groups enhanced other sources of professional information. However, the majority also stated that discussion groups did not replace other sources of information.

The purpose of the survey conducted by Zhang (2000) was to enhance understanding of the scholarly use of Internet based e-sources among LIS researchers and to evaluate the potential of Web-based surveys. The population of the survey was 201 researchers with in-press publications in eight LIS journals. An e-mail was sent to these researchers requesting they participate in the survey. The respondents could either fill out a Web-based questionnaire or request a printed copy and return it by mail or fax. Only 10 percent of the researchers requested printed copies, and 20 percent of the researchers returned the questionnaires by regular mail or fax (some of them printed out the Web-based questionnaires by themselves). Three follow-ups were sent out, and the total response rate was 89.1 percent.

Spink, Bateman, & Jansen (1999) demonstrated a different use of Web-based surveys. The survey was made available from the Excite home page (a Web portal—http://www.excite.com[1]) for a five-day period in 1998. About 7.7 percent of the users who visited the survey page (approximately 3,700 visitors) filled out the survey and submitted it (p. 119).

Conducting surveys on the Web or through e-mail is becoming popular. Piper (1998) raises an important question: "Can experiments conducted on the Web avoid the many threats to internal validity, construct validity and external validity?" Her main concerns were "nonrepresentative, volunteer subjects and deception by subjects" (p. 10). Zhang (2000) also addresses the problems of biased sample, biased return, and low response rates. The above examples indicate that the key to achieving a reasonable response rate is to send the questionnaire to a population that enables researchers to also send personalized follow-ups to non-respondents (as in Lazinger, Bar-Ilan, & Peritz, 1997 and Zhang, 2000).

Monitoring/Logging

Another method of data collection—again, mainly concerned with the use of different Internet services—is by monitoring and analyzing log files of scientists' Internet use. Kaminer & Braunstein (1998) analyzed the log files of the sixty-three faculty members of Berkeley's College of Natural Resources in order to assess the impact of Internet use on scholarly production. They measured the number of distinct processes, the length of the sessions, and the types of services used. A questionnaire was also sent out to these faculty members. The main finding was that "adding measures of Internet use improves the explanatory power of the traditional model of scholarly productivity" (p. 729).

Lawrence & Giles (1999) monitored the queries to different search engines presented by scientists at the NEC Research Institute. This set of queries constituted a sample of real-life queries. The analysis assumes that this is a truly random sample of user queries, although it is rather doubtful that the queries of NEC scientists at work are representative of the queries of the "typical" user. Based on this set, they calculated the coverage of the different search engines of the Web. At the time of the study (February 1999), the then largest search engine, Northern Light, covered only about 16 percent of the Web pages reachable and indexable by search engines.

Two studies on end-user searching on the Internet were based on huge logs from the search engine Excite (Jansen, Spink, & Saracevic, 2000; Ross & Wolfram, 2000). The first study analyzed 51,473 queries of more than 18,000 users and provided data on changes during the query sessions—on the number of search terms, on the usage of Boolean operators and query modifiers, and on the most highly used search terms. They also identified trends among user mistakes. The second study analyzed term cooccurrence in more than a million queries that "represent a subset of queries submitted to the Excite search engine on a single day" (p. 950).

Crawling

Today the Web is far too large and complex to even attempt to cover it all. Recently, Moore & Murray (2000) estimated that there were at least 2,100 million indexable pages on the Web in July 2000, with an estimated daily growth rate of seven million Web pages. A few years ago the Web was much smaller, and very large crawls of the Web probably depicted a reasonable picture on the structure of the "average Web page." In November 1995, Woodruff et al. (1996) analyzed 2.6 million documents collected by the search engine Inktomi, developed at the time at Berkeley. The characteristics examined included: Document size, number and types of tags, number of links, and ratio of document size to number of tags. They also listed the "most linked-to URLs."

At about the same time Bray (1996) analyzed the results of a 1.5 million sample collected by the search engine Open Text (does not exist anymore), and described the "average Web page" in terms of size, number of embedded images, and incoming and outgoing URLs. He also tabulated the biggest and most visible sites (defined according to the number of links pointing to them).

Retrieval by Sampling

Bharat & Broder (1998) attempted to create "random URLs" in order to compare the coverage of different search engines. Their objectives were similar to those of Lawrence & Giles (1998 and 1999), however their methodology was different: They sampled a weighted dictionary of Web words based on pages indexed by the human-edited directory service Yahoo (http://www.yahoo.com). Two term AND and OR queries were presented to the search engines and random URLs were selected from the result sets.

These URLs were assumed to be "random URLs." In spite of the different techniques, the results of Bharat & Broder's experiments are comparable to those of Lawrence &·Giles (1998). Both experiments took place in November 1997.

Exhaustive Retrieval from Databases

Retrieving all documents from the Web on a given topic or from a given domain or country allows the researchers to create random samples from the set. Almind & Ingwersen (1997) utilized this method. The initial set of Danish pages on the Web was retrieved from the Nordic Web Index (not operational anymore). To supplement this set, searches were also carried out on other search engines. These additional searches added only 200 new pages to the list of 47,000 Danish URLs retrieved from the Nordic Web Index. The very large overlap between the different sources points to the exhaustiveness of the set of pages indexed by the Nordic Web Index as of December 1995. The Danish Web pages were compared to those of other Scandinavian countries.

Bharat et al. (1998) built a huge snapshot of the link structure of the Web, based on a crawl of 100 million pages of AltaVista (http://www.altavista.com). The so-called "Connectivity Server" does not have data on the content of the different pages, but gives information on the incoming and outgoing links of sets of nodes. The Connectivity Server enables the researchers to carry out experiments in a relatively stable environment.

Search Engines and Other Retrieval Tools

The large general search engines are natural choices for collecting specific data from the Web. There are several ways to utilize the search engines: A single service can be used, or the results of several search tools can be compared or combined. The following three studies are examples of each of these uses.

Rousseau (1997) retrieved all the occurrences of the search terms "informetrics OR bibliometrics OR scientometrics" using AltaVista. AltaVista is one of the most popular search tools among Internet researchers, because it has a wide variety of useful options. However, it has been noted in several studies that its reliability is questionable (e.g., Ingwersen, 1998; Rousseau, 1999; Thelwall, 2000; Bar-Ilan, 2001).

Cronin et al. (1998) searched the Web using five search tools: Excite, Infoseek (currently the service can be found at http://www.go.com), Lycos (today this is an altogether different service, powered by Fast, but can still be found at http://www.lycos.com), WebCrawler (http://www.webcrawler.com), and Yahoo for pages mentioning five prominent professors in library and information science. The results retrieved from these engines were compared, and the combined results were also analyzed.

Bar-Ilan (1998) searched seven of the then largest search tools for pages mentioning the mathematician "Erdos." The results of the seven tools (AltaVista, Excite, Infoseek, Lycos, Magellan, Opentext, and Yahoo) were combined in order to get a picture of the way Erdos was depicted on the Web around the end of 1996.

Aguillo (2000) advocates the use of client-side based tools for the discovery of quality information on the World Wide Web. He recommends the use of quantitative indicators based on the visibility of sites.

Additional Data Collection Methods

Watson (1998) interviewed high school students in order to get a "close look at students' perceptions of using technology" (p. 1024), mainly the Internet. This method of open-ended interviews can only be used for very small populations—nine students in this case.

Rosenbaum (1998) analyzed the content of the Web sites of twenty-four Web-based community networks in Indiana. He did not have to search for these sites, since he already had knowledge of their existence. The same data collection method of retrieving data from known sites was applied by Koehler et al. (2000) when different "demographic aspects" of three e-journals (*Cybermetrics, Information Research,* and *Libres*), a print journal (*Journal of Internet Cataloging*) and a hybrid journal (*JASIS*) were analyzed. Results included data on the productivity of these journals, characteristics of papers, authors, and funding.

Haas & Grams (2000) used AltaVista's Surprise link (not existent anymore) to collect a set of pages and to characterize them and the types of links emanating from them. The Surprise link was supposed to link to "random" pages from the AltaVista database.

Bucy, Lang, Potter, & Grabe (1999) obtained data on page views from the 100hot's Insite Pro service (http://www.100hot.com). The InSite service does not seem to be operational anymore, but 100hot.com publishes the list of 100 most visited sites based on the usage patterns of over 100,000 Web users from all over the world (*100hot methodology,* n.d.).

Informetric Analysis

Irrespective of the data collection method, the collected data have to undergo some analysis in order to arrive at meaningful conclusions. Sometimes simple processing and standard statistical and mathematical analysis are sufficient, but at other times specific informetric methods, models, or laws are utilized. In the following sections we review the use of these methods, models, and laws for analyzing data from and about the Internet.

Citation Analysis

Harter (1996 and 1998) carried out one of the earliest attempts to assess the scholarly impact of electronic journals. He measured the number of citations of thirty-nine e-journals received by February 1996. The citations were extracted from ISI's Citation Indexes. Fifteen journals were not cited at all, and only seven were cited eleven times or more. Except for one or two exceptions the impact of these journals (in early 1996) was minimal.

Zhang (1998) investigated the citations to e-sources in library and information science journals during the period 1994 and 1996. E-sources were defined as: E-mail messages, messages posted to newsgroups and discussion lists, publications of any kind (not necessarily refereed), commercial sources, and other e-sources available from the Internet. Harter counted citations the specific e-journals received from journals indexed by ISI. Zhang, on the other hand, examined all types of references to e-sources appearing in the ten most highly cited library and information science journals and in four library and information science oriented e-journals. Except for the e-references appearing in the four e-journals, the impact of the e-sources was negligible.

At the very beginning, researchers noticed that incoming links to a Web page measure its visibility (see, e.g., Bray 1996 or Woodruff et al., 1996). Links can be seen as analogues of citations in the academic world. General search engines, like AltaVista and Hotbot (http://www.hotbot.lycos.com), retrieve lists of URLs in their database linking to a given URL or site. Recently Google (www.google.com) also added this option. Because of the limited coverage of the Web by these search tools, the link information is also limited. For example, consider the homepage of *Library Trends* (http://www.lis.uiuc.edu/puboff/catalog/trends/). AltaVista found 14 pages linking to it, Hotbot found 6 links, while Google found 129 pages linking to this URL. A similar search for the homepage of *JASIS* (http://www.asis.org/Publications/JASIS/jasis/html) resulted in 226 links reported by AltaVista, 160 links reported by Hotbot, and 245 links reported by Google. Even this small example illustrates that we cannot rely on search engines to produce reliable visibility data. All searches were carried out on November 18, 2000. The accuracy of the results was not examined.

Chakrabarti, Gibson, & McCurley (1999) advocate the provision of backlinks (pages that link to a given page) by the sites themselves and not through the search engines. Even though the implementation is not difficult, they are aware of privacy concerns and of other barriers of acceptance. For instance, commercial sites most likely will not be interested in linking to bad reviews about their products or to pages that also mention their competitors. In fact, it is hard to imagine that any site would be willing to include in the lists of pages that link to it those pages that have a negative attitude towards the site.

Cui (1999) used citation analysis to rank health Web sites. Again, the hypertext links were viewed as citations. The study analyzed the links appearing on the homepages of the libraries of the top U.S. medical schools, as compiled and published by *U.S. News and World Report.*

Lawrence, Giles, & Bollacker (1999) took a completely different approach to citation analysis on the Web. Instead of studying hypertext links as analogues of citations in the academic world, they looked for citations in the classical sense, and their "Autonomous Citation Indexing" (ACI) system can automatically create a citation index from literature in electronic format. The rationale behind this project is that an increasing number of authors, journals, institutions, and archives make research articles available on the Web, mainly in PDF or Postscript formats. ACI is implemented for computer science literature at the "ResearchIndex" site (http://www.researchindex.com/). The system allows one to search articles and citations. When searching for citations, it provides citation context (in the citing article), citation statistics, and links to the citing articles. For full-text articles the system also displays the exact bibliographic reference, the list of citations and the list of references, similar documents (textual similarity), and related documents (based on cocitations). The user interface needs some improvement.

Garfield (1999) related to this project in an address delivered at a symposium in honor of Manfred Kochen: " . . . without aposteriori human intelligence, the Internet will remain at best a mixed blessing. Artificial intelligence will help but not suffice. . . . The Internet has made it practical for future citation index databases to generate annotated bibliographies and reviews containing contextual quotations based on autonomous citation indexing. To see how this works in the field of computer science just go to www.researchindex.com."

A recent paper (Goodrum, McCain, Lawrence, & Giles, 2001) compares the ISI SCISEARCH Citation Index to ACI in the area of computer science. A major difference between the two systems is that ACI indexes PDF and Postscript formatted publication on the Web, while SCISEARCH indexes only a selected list of journals in the area.

Cocitation and Coword Analysis

The objective of the study carried out by Larson (1996) was to explore the applicability of classical cocitations on the Web, when citations are substituted with hyperlinks. He carried out a cocitation analysis of a set of Earth Science related Web sites. The starting point were two authoritative sites on the topic. The list of pages pointing to these two sites was retrieved using AltaVista, and the links appearing in the relevant pages were extracted. This set underwent a second round of relevance judgment by Larson, and a set of thirty-four "core" pages was created. Again, AltaVista's link option was utilized to retrieve the number of URLs linking to each of the 544 cocitation pairs. The data were converted to a correlation matrix and multidimensional scaling (MDS) was used to create the cocitation map. Larson concluded, "the mappings . . . seem to produce quite clear, reasonable and interpretable results."

Dean & Henzinger (1999) applied cocitation techniques in order to find "related pages" on the Web. A related page is one that addresses the same topic as the original page. One of their algorithms, the *cocitation* algorithm, looks for pages that link to the given page, and assumes that the nearby links point to pages with similar topics. These pages were collected, their cocitation degree computed, and those with the highest degrees were returned as the most related pages.

Kumar, Raghavan, Rajagopalan, & Tomkins (1999) used cocitation techniques in order to identify specific communities on the Web—groups of content creators sharing a common interest. The study exploits "cocitation in the Web graph to extract all communities that have taken shape on the Web, even before the participants have realized that they have formed a community" (p. 1483).

Ross & Wolfram (2000) used coword analysis to analyze term pair topics submitted to the search engine Excite. Their data were based on more than a million queries submitted to Excite on a single day. The most frequent term pairs were coded into thirty categories based on the semantic and pragmatic intent of the term pair; a term pair could belong to more than one category. Cluster analysis and MDS were used for the data analysis. A high proportion of the term pairs were for adult-oriented material.

Leydesdorff & Curran (2000) studied the cooccurrence of the terms "university, "industry," and "government" in Web pages in three different domains. The domains were: Brazil, the Netherlands, and the so-called top level domains (.com, .edu, .gov, .org, .net, .mil). They studied the growth over time of these cooccurrences, using AltaVista's option to limit searches to given dates. The queries were presented both in English and in the local language. Similar trends were detected in all three domains.

Content Analysis

Content analyses of Web and Internet sources serve as exploratory tools for getting a better understanding of the Internet's content.

Bar-Ilan & Assouline (1997) analyzed the content of messages distributed by the PUBYAC (a discussion list for Children and Young Adult services) for a period of one month in spring 1997. Six content categories were defined (reference, library administration and policy, collection management, extension programs, announcements, and other). The most popular category was reference. The lifespan of topics, the number of active participants, and the productivity of the participants were also examined. From the answers received to a specific question sent to the participants of the discussion list, it seems that the librarians find the list very useful: "It helps them find answers to specific questions and assists in collection management and planning extension programs" (p. 170). Several other studies analyzed the content of discussion lists. Sometimes several groups were analyzed in parallel and their characteristics compared (e.g., Aires-de-Sousa, 1999; Schoch & White, 1997; Berman, 1996).

Not only discussion lists were analyzed, but also Web pages and Web sites. Cronin et al. (1998) searched the Web using five search tools for pages mentioning five prominent professors in library and information science. The retrieved Web pages were characterized according to the "forms of mention." Eleven categories of invocation were defined: Abstract, article, conference proceedings, current awareness, external home page, listserv, personal/parent organization home page, resource guide, book review, syllabus, and table of contents. The data were collected over a period of two months, though the dates are not given. The authors concluded: The Web "engenders new modes of scholarly interaction and signaling. Scholars do not merely post, or publish, their works on the Web: They seed ideas, discuss issues, and debate positions, in ways which, occasionally, deviate from, and challenge, established norms" (p. 1326).

A different kind of content analysis, examining not the form of invocation, but the different contexts in which the mathematician Paul Erdos was mentioned, appears in Bar-Ilan (1998). The paper analyzes the content of 2,685 Web documents collected between the end of 1996 and the beginning of 1997 (Paul Erdos passed away in September, 1996). Six main content categories were defined: Mathematical work, Erdos number, in honor/memory of Erdos, jokes/quotations, math education, and other. Almost 40 percent of the pages were classified as "mathematical work," but a rather surprising 13 percent of the pages belonged to the jokes/quotations category. (The most popular quotations/jokes were: "A mathematician is a machine for turning coffee into theorems" and "Why did the chicken cross the road? It was forced to do so by the chicken-hole principle"). The concept of Erdos number intrigues the authors of the Web pages; the concept was explained on ninety-one (3 percent) different pages (almost always exactly the same explanation), and 9 percent of the collected pages point to the home page of the "Erdos Number Project" (http://www.oakland.edu/~grossman/erdoshp.html). In 40 percent of the pages belonging to "mathematical work," Erdos's name was mentioned in bibliographical references.

Formal bibliographical references also appeared in Bar-Ilan (2000c), in a large portion of the pages (in 40.3 percent out of the 807 pages) containing the search terms "informetrics OR informetric." The searches were carried out in June 1998 using the six largest search engines at that time (AltaVista, Excite, Hotbot, Infoseek, Lycos, and Northern Light). The references extracted from these pages (called the "Web database") were compared with comparable data retrieved from commercial bibliographical databases. In all except one comparison, the Web database did at least as well as the commercial database, indicating that valuable, freely available data exist in the Web, but cannot be located easily.

Lawrence, Bollacker, & Giles (1999) were able to find large quantities of full-text papers in the area of computer science. They were looking for different formats, including PDF and PostScript. Bar-Ilan's findings were rather different; she located only a negligible number (4) of full text publications. This may be due to the fact that she collected information about a different subject area or to the fact that general search engines ignore formats like PDF and PostScript. The most productive and the most cited authors and sources, and the most cited papers (papers which are referred to in the largest number of collected Web pages) were also calculated.

Rosenbaum (1998) analyzed the content of the Web sites of twenty-four Web-based community networks in Indiana. The purpose of the study was to learn about the content and the structure of these sites.

Bar-Ilan (2000b) analyzed the content of Web pages containing the phrase "S&T indicators." Several facets were introduced, including the context in which the search phrase appeared, the type of document, the server, the domain, the geographical area, and the time period for which the indicators were computed. A rather interesting finding was the existence of a large number of Web pages with data from Malaysia. Since 1992, the Malaysian government has consistently published its Science and Technology reports on the Web.

Evaluation Using Existing/New Measures

Gordon & Pathak (1999) measured the retrieval effectiveness of Web search engines. Thirty-three members of the faculty at the University of Michigan Business School described to experienced searchers their information needs. The searchers presented appropriately phrased queries to eight search tools. The first twenty hits from each tool were retrieved and the 160 documents in some random order were presented to the faculty members, who judged the relevance of these documents. The absolute retrieval effectiveness was fairly low, and there were statistical differences in precision effectiveness.

A different approach was taken by Bar-Ilan (1999), who, instead of the subjective human relevance judgments, measured the technical precision of the retrieved documents. A document is technically relevant if it satisfies the query (i.e., all the search terms that are supposed to appear are actually present in the document, and all the terms that are not supposed to appear are missing). This is an objective measure, which can be computed simply, but it does not judge the quality of the document.

Oppenheim, Morris, McKnight, & Lowley (2000) gave an extensive review of the evaluation of Internet search engines. Precision was measured in most studies, but recall measuring is extremely difficult. Some suggested alternative methods were reviewed, and the authors recommended developing a standardized set of tools for search engine evaluation.

Page & Brin (1998) introduced a new method of measuring the quality of Web documents, called the PageRank. The method is based on the ideas of classical citation analysis, but instead of simply counting the number of links pointing to a document the quality of the page from which the link emanates is also taken into account. Similar ideas of weighing citations for classical citation analysis were introduced already in Pinski & Narin (1976). Egghe (2000) slightly disagrees with the analogy drawn between classical citations and hypertext links: Paper B citing paper A was necessarily written after paper A; however, this is not the case with Web pages, quite often there are reciprocal links between pages.

Henzinger, Heydon, Mitzenmacher, & Najork (1999) defined a new measure for search engines: "Search engine quality." The quality of a Web page is based on the links pointing to it. Some portion of the Web is crawled in order to estimate the "quality" of pages, and then the search engines are queried with a sample of the visited high quality pages to check if they index them.

One way to measure page popularity is through the number of links pointing to it (as in Page & Brin, 1998). Another possibility is to count the number of visitors to a site by an objective body (not self-adjustable counters on a Web page). Such a method is utilized by the Direct Hit service (http://www.directhit.com). The service monitors "which web sites Internet searchers select from the search results list, how much time the searchers spend at these sites and a number of other metrics, such as the position of a site relative to other sites. The sites that are selected by searchers are boosted in their ranking, while the sites that are consistently ignored by searchers are penalized in their rankings" (*Direct Hit Technology,* n.d.).

There are several works which examine formal features of Web pages and sites; for example, size and type of files and images, number of forms and other methods of interaction, applets, number and types of links. Bauer & Scharl (2000) used such data for "quantitative evaluation of Web site content and structure." Even though the data can be collected automatically, it is difficult to see how it evaluates the site, since evaluation is associated with quality. The authors suggested manual classification as one of the methods to analyze the raw data. Bucy, Lang, Potter, & Grabe (1999) used the data to deduce relationships between Web page complexity (banners, length, colors, graphical, dynamic, and interactive elements) and site traffic. They found significant relationships between site traffic and graphical elements for commercial pages, and between site traffic and asynchronous interactive elements for noncommercial documents. The page usage data were obtained from the 100hot's Insite Pro service, which tracks the usage patterns of over 100,000 Web users from all over the world.

Identifying and Calculating Indicators

Ingwersen (1998) was the first to define specific indicators for the Web. He defined the Web impact factor (WIF) as follows:

$$\frac{\text{\# of pages with a link to the site or country}}{\text{\# of pages in the site or country}}$$

He compared the WIF of different European countries, using AltaVista's link feature. WIFs of specific sites (like the site of *Science Magazine*) were also calculated. Just like the classical impact factor of journals, the WIF of a given country or site indicates its relative visibility on the Web.

Smith (1999) examined some methodological issues related to the WIF, and claimed that the external WIF (counting only links emanating from outside the site) is probably the best indicator. Internal WIFs do not really reflect on the visibility of a site, because a large portion of the links may simply be navigational links (back to the home page, etc.) or can be self-inflated, just like self-citations in classical citation analysis. His experiments show that WIFs for countries are not very reliable, but for large organizations the indicator seems useful.

Both Ingwersen and Smith used the link and domain options of AltaVista to calculate WIFs, since currently AltaVista is the only large search engine having both options. Thelwall (2000) warns that the uneven coverage of the Web by the search engines results in misleading calculations of the WIF.

Aguillo (1997) introduced a new procedure to obtain quantitative indicators of science and technology. The indicators are derived from the presence of research and development institutions on the Internet. Hypertext links between these institutions are treated the same way as citations in the ISI databases. The different types of multimedia objects are also subject to quantitative analysis. The planned database (*Internet World of Research and Development—IWR&D*) will include information on 20,000–25,000 sites. The suggested indicators include: Self-citations, density of links, visibility, WIF, and diversity. In Aguillo & Pareja (2000) some of these indicators were calculated for four Western European countries. The results showed that visualization measures based on WIFs are rather consistent and can be used to supplement scientometric data.

Models

Page & Brin (1998) were the first to rely on the structure of the Web in order to improve information retrieval. They modeled the Web as a directed graph with weights (called PageRank) on the nodes (the Web pages). These weights are a function of the number of incoming links and the weight of the pages they emanate from. This is the model behind the search engine Google.

At about the same time Kleinberg (1998) introduced the model of hubs and authorities. Authorities are pages with quality information, and hubs are pages with lists of links. Kleinberg developed an algorithm for identifying hubs and authorities. An initial set of Web pages on the topic is retrieved by a general search engine. This initial set is augmented with pages pointing to the set and to pages pointed to from the set (corresponding to the notions of cocitation and bibliographic coupling). An iterative weighing process results in a set of authorities and a set of hubs. The algorithm utilizes the link structure of the hypertext system; it does not rely on any linguistic characteristics. The process works even if the initial query has multiple meanings (e.g., jaguar). Kleinberg's ideas were implemented and extended in the IBM CLEVER Project (Chakrabarti et al., 1999).

Egghe (1997) studied the fractal features of hypertext systems and was able to find a link between the fractal theory of hypertext systems and informetrics. By his definition the fractal dimension is a function of the total number of Web pages and the average number of hyperlinks per page.

Fitting Existing Models/Bibliometric Laws

Growth and core. Bar-Ilan (1997) examined how newsgroups reacted to a crisis. The specific crisis was the outbreak of "mad cow disease" in Britain in the spring of 1996. Data were collected for a period of 100 days between April and July 1996, using AltaVista, which at the time indexed 14,000 newsgroups. The searches were carried out across the newsgroups using both popular (mad cow disease) and scientific terms (BSE, prion OR prions, bovine spongiform) related to the disease. The growth curve of the messages on the subject resembled the logistic growth function. It was possible to identify an initial period of extremely fast growth, then a second period of moderate growth. By the beginning of July 1996 interest in the disease went down considerably. A similar trend was detected in the number of relevant articles published in the *Times* and the *Sunday Times.* Rather interestingly the graph for BSE, unlike the other graphs, showed a linear growth throughout the whole period. This was due to the fact that B.S.E. is also an abbreviation in electrical engineering.

Data were retrieved from more than a thousand different newsgroups, Bradford's law was shown to be applicable, and it was possible to identify "core newgroups" that deal with the subject. Other characteristics of the messages were also examined (domains, most productive authors, most popular subjects, etc.).

Two informetric papers studied the growth of differerent topics. Rousseau (1999) carried out three daily single word searches (trumpet*, pope, and saxophone*) in AltaVista and in Northern Light for a period of twenty-one weeks between July and December 1999. The results for Northern Light show slow monotonic growth in the number of results, while large fluctuations were observed for AltaVista. Curves were fitted to the Northern Light data, from which predictions were made as to the growth of the Northern Light database. Leydesdorff & Curran (2000) studied the growth of the number of Web pages containing the term "university," "industry," and "government" (and combinations of these terms) for Brazil, the Netherlands, and the top level domain. AltaVista was queried with dates limited to calendar years between 1993 and 1998. When taking this approach, one must be aware that the date of a Web page is at best the last time the page was updated, and if data are unavailable or unreasonable (e.g., dated in the future), the date is the last time the crawler of the search engine visited that page.

Power laws and Zipf-type laws. Rousseau (1997) in an early paper illustrated that bibliometric laws are applicable to the Internet. In May 1997, AltaVista was searched for "bibliometrics OR informetrics OR scientometrics." The results set, consisting of 343 documents, was retrieved. The number of pages citing each of the pages in the results set was determined using AltaVista's link option. Rousseau was able to fit appropriate Lotka functions to the data both for the number of retrieved pages per site, and for the number of citations to a site.

It turns out that Rousseau's results can be generalized to several characteristics of the Web. Huberman, Pirolli, Pitkow, & Luskose (1998) showed that the surfing behavior of Web users follows Zipf-like distributions. The authors proposed a model of Web surfing that explains the empirical findings on distributions of page hits observed at Web sites. Albert, Jeong, & Barabasi (1999), based on a subset of the Web of about 325,000 pages, showed that both incoming and outgoing links obey appropriate power laws. Huberman & Adamic (1999) explained the distribution of the number of pages per site again by a power law. The largest test to date was run by Broder et al. (2000), based on a Web crawl of approximately 200 million pages. This experiment validates the power law distributions both for incoming and outgoing links. The authors noted that Zipf-like distributions (based on rank instead of magnitude) for incoming links give a better fit than the power law distribution. The Web is a complex system, characterized by growth and preferential attachment, as explained by Barabasi & Albert (1999).

Obsolescence. Obsolescene, or more precisely, the characterization of the changes occurring to Web documents, has been studied both in the informetric and the practical setting.

The content and the format of printed literature do not change after publication. This is obviously not the case for Web documents. On the one hand, new documents are continuously being published on the Web. On the other hand, existing documents are removed from the Web, have their location changed, or undergo changes in content or in format. Documents may be removed from the Web for several reasons; for example, the page may become outdated, the server on which it resides ceases to exist or malfunctions, the author of the page is not allowed to use the server anymore, or the author simply loses interest in the topic. The change in the location of a document is usually due to technical reasons. Sometimes there is a forwarding note or an automatic redirect to the new location. Internal changes to a document indicate that the page was updated. Some documents are never removed from the Web, even if they contain totally outdated information. Unfortunately, most of Web documents are not dated. Thus, it is sometimes almost impossible to decide whether the information is current (e.g., opening times of events, entrance fees, or sizes—demographic or Web data).

Bar-Ilan & Peritz (1999) studied changes that occur to Web documents over time regarding a given scientific topic. Documents containing the terms "informetrics OR informetric" were retrieved from the six largest engines at the same time once a month for a period of five months (spring–summer 1998). The set of documents on informetrics seems to be much more stable than documents on general, popular topics. Most of the documents that were retrieved more than once were stable. Among the pages that did change during the observation period, the majority underwent frequent, major changes. Thus, pages are either completely static or are changed often and considerably.

Koehler (1999) analyzed Web page and Web site constancy and permanence. The sample of the URLs was identified by using the random URL generator feature of WebCrawler. There was thus no apriori characterization of the observed Web pages. The pages were retrieved once a week between January 1997 and January 1998. The permanence of these pages was investigated. Three categories were defined: Always present, intermittent, and comatose. At the end of the period about 30 percent of the pages failed to respond. The changes that the pages underwent were also categorized. Nearly all Web pages changed during the year. The influence of the type of URL, and quantitative aspects of the pages (sizes, multimedia, e-mail links, etc.) on constancy and permanence were also studied.

It is important to mention here two other works that studied changes to Web pages for practical reasons. Douglis, Feldmann, Krishnamurthy, & Mogul (1997) observed the rate of change of Web pages in order to assess the benefits of caching (the less changes to the pages, the more useful it is). They found that content type and rate of access have a strong influence, while domain and size have little effect. The purpose of Brewington & Cybenko's study (2000) was to estimate the rate at which Web search engines must re-index the Web in order to remain current. Both studies based their findings on large data sets.

Conclusion

This review was based on the different methods of classical informetric analysis. A tabulated summary of the review is presented by way of the topics informetrics investigates. Some of the reviewed studies clearly showed the applicability of bibliometric laws to the Internet, while others developed new definitions and methods based on the respective definitions for printed sources. In some cases the Web research community introduced or re-introduced (as in the case with weighted links) models and methods that may also be applied to printed sources. Both informetric and Internet research can gain from these new developments.

Table 1. Characteristics and Measurements of Countries, Groups, Persons (Authors).

Productivity	Almind & Ingwersen (1997)	Country: Denmark
Growth	Almind & Ingwersen (1997)	Country: Denmark
Interaction	Bar-Ilan & Assouline (1997)	Group: Participants of the discussion list
Topics/Subjects	Bar-Ilan & Assouline (1997)	Group: Participants of the discussion list
Use	Lazinger, Bar-Ilan & Peritz (1997)	Group: Faculty members of the Hebrew University
Visibility	Cronin et al. (1998)	Persons: Prominent library and information scientists

Table 2. Characteristics and Measurements of Publications and Publication Sources.

Productivity	Koehler et al. (2000)	Productivity of e-journals in information science
	Harter (1998), Zhang (1998)	Impact of e-journals
Growth	Bar-Ilan (1997)	News-group postings viewed as publications
Obsolescence	Bar-Ilan & Peritz (1999) Koehler (1999)	For a specific topic For a set of pages
Core	Bar-Ilan (1997)	News groups viewed as publication sources
Ranking/Visibility	*Direct Hit Technology* (n.d.)	Measures the number of visits to a page
	Page & Brin (1998)—PageRank	Measures the link popularity
Topics	Bar-Ilan (2000a) Bar-Ilan (2000b)	Characterization of the topics of pages retrieved for the query "informetrics" and for "S&T indicators"
Structure	Aguillo & Pareja (2000)	Structure of R&D sites of four western European countries
Linguistic Characteristics	Leydesdorff & Curran (2000)	Comparing occurrences of words in English vs. the local language
Citation Patterns	Lawrence, Bollacker, & Giles (1999) Chakrabarti, Gibson, & McCurley (1999)	Classical citations Hypertext links as citations
Use	*100hot Methodology*, n.d.	Displays list of 100 top viewed sites based on surfing patterns of over 100,000 users. Data is updated weekly

Web pages and e-mail messages are viewed as publications, and sites, discussion lists, and news-groups as publication sources.

Table 3. Characteristics and Measurements of Disciplines, Fields, Subfields, and Topics.

Growth	Bar-Ilan & Peritz (1999)	Informetrics
Development/ Structure	Larson (1996)	Cocitation analysis of earth science sites
Interdisciplinarity/ Interaction		Between fields of research
Indicators	Ingwersen (1998)	WIF
	Aguillo (1997)	A list of indicators
Prediction/Planning	Moore & Murray (2000)	Study on the rate of growth of the Web
	Rousseau (1999)	Predictions on the growth rate of Northern Light

Table 4. Characteristics and Measurements of Databases and of the Information Retrieval Process.

Evaluation	Oppenheim, Morris, McKnight, & Lowley (2000)	Review of search engine evaluation methods
Use	Spink, Bateman, & Jansen (1999)	Survey of Excite users
Coverage	Bharat & Broder (1998) Lawrence & Giles (1998, 1999)	Estimates on search engine coverage

Note

[1] The URL of a service or a site is given only the first time the service or site appears in the text.

References

100hot methodology (n.d.). Retrieved November 25, 2000 from http://www.100hot.com/help/methodology.html.

Aguillo, I. F. (1997). STM information on the Web and development of new Internet R & D databases and indicators. *Online Information '97 Proceedings* (pp. 239–243). Oxford: Learned Information Europe, Ltd.

Aguillo, I. F. (2000). 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., & Pareja, V. M. (2000). Indicators of the Internet presence of the Western European Research Councils. Poster Presentation at S&T 2000, Leiden, the Netherlands, May 2000. Retrieved November 23, 2000 from http://sahara.fsw.leidenuniv.nl/cwts/abs/AGUILLO.txt.

Aires-de-Sousa, J. (1999). An electronic discussion forum for organic chemistry—The ORGLIST case. *Internet Journal of Chemistry, 2*(21), 1–10.

Albert, R.; Jeong, H.; & Barabasi, A. L. (1999). Diameter of the World Wide Web. *Nature, 401*(6749), 130–131.

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to "webometrics." *Journal of Documentation, 53*(4), 404–426.

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.

Bar-Ilan, J. (1997). The 'Mad Cow Disease', usenet newsgroups and bibliometric laws. *Scientometrics, 39*(1), 29–55.

Bar-Ilan, J. (1998). 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. Retrieved November 15, 2000 from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html.

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

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) The Web as information source on informetrics?—A content analysis. *Journal of the American Society for Information Science, 51*(5), 432–443.

Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes—A review and analysis. *Scientometrics, 50*(1), 7–32.

Bar-Ilan J., & Assouline B. (1997). A content analysis of PUBYAC—A preliminary study. *Information Technology and Libraries, 16*(4), 165–174.

Bar-Ilan, J., & Peritz B. C. (1999). The life span of a specific topic on the Web; The case of 'informetrics': A quantitative analysis. *Scientometrics, 46*(3), 371–382.

Bauer C., & Scharl, A. (2000). Quantitative evaluation of Web site content and structure. *Internet Research: Electronic Networking Applications and Policy, 10*(1), 31–43.

Berman, Y. (1996). Discussion groups on the Internet as sources of information: The case of social work. *ASLIB Proceedings, 48*(2), 31–36.

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. *Computer Networks and ISDN Systems, 30,* 379–388 (Proceedings of the 7th International World Wide Web Conference, April 1998). Retrieved November 15, 2000 from http://decweb.ethz.ch/WWW7/1937/com1937.htm.

Bharat, K.; Broder, A.; Henzinger, M.; Kumar, P.; & Venkatasubramanian, S. (1998). The connectivity server: Fast access to linkage information on the Web. *Computer Networks and ISDN Systems, 30,* 469–477 (Proceedings of the 7th International World Wide Web Conference, April 1998). Retrieved November 15, 2000 from http://www7.scu.edu.au/programme/fullpapers/1938/com1938.htm.

Bray, T. (1996). Measuring the Web. *Computer Networks and ISDN Systems, 28*(7–11), 933–1005 (Proceedings of the 5th International World Wide Web Conference).

Brewington, B. E., & Cybenko, G. (2000). How dynamic is the Web? *Computer Networks and ISDN Systems, 33,* 257–276 (Proceedings of the 9th International World Wide Web Conference, May 2000). Retrieved November 15, 2000 from http://www9.org/w9cdrom/264/264/html.

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomlins, A.; & Wiener, J. (2000). Graph structure in the Web. *Computer Networks and ISDN Systems, 33,* 309–320 (Proceedings of the 9th International World Wide Web Conference, May 2000). Retrieved August 20, 2000 from www9.org/w9cdrom/160.160.html.

Brookes, B. C. (1990). Biblio-, sciento-, infor-metrics??? What are we talking about? In L. Egghe & R. Rousseau (Eds.), *Informetrics* (vol. 89/90, pp. 31–42). Amsterdam: Elsevier.

Bucy, E. P.; Lang, A.; Potter, R. F.; & Grabe, M. E. (1999). Formal features of cyberspace: Relationships between Web page complexity and site traffic. *Journal of the American Society for Information Science, 50*(13), 1246–1256.

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. Retrieved November 15, 2000 from http://www.sciam.com/1999/0699issue/0699raghavan.html.

Chakrabarti, S.; Gibson, D.; & McCurley, K. S. (1999). Surfing the Web backwards. *Computer Networks and ISDN Systems, 31*(11–16), 1679–1693 (Proceedings of the 8th International World Wide Web Conference, May 1999). Retrieved November 15, 2000 from http://www8.org/w8-papers/5b-hypertext-media/surfing/surfing.html.

Cronin, B.; Snyder, H.; Rosenbaum, H.; Martinson, A.; & Callahan, E. (1998). Invoked on the Web. *Journal of the 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. Retrieved November 18, 2000 from http://www.jmir.org/1999/1/e4/index.htm.

Dean, J., & Henzinger, M. (1999). Finding related pages in the World Wide Web. *Computer Networks and ISDN Systems, 31,* 389–401 (Proceedings of the 8th International World Wide Web Conference, May 1999). Retrieved November 18, 2000 from http://www8.org/w8-papers/4a-search-mining/finding/finding.html.

Direct Hit Technology. (n.d.). Retrieved November 23, 2000 from http://www.directhit.com/about/products/.

Douglis, F.; Feldmann, A.; Krishnamurthy, B.; & Mogul, J. (1997). Rate of change and other metrics: A live study of the World Wide Web. In *Proceedings of the Symposium on Internet Technologies and Systems,* Monterey, California, December 8–11, 1997. Retrieved November 20, 2000 from http://www.usenix.org/publications/library/proceedings/usits97/full_papers/douglis_rate.

Egghe, L. (1997). Fractal and informetric aspects of hypertext systems. In B. C. Peritz & L. Egghe (Eds.), *Proceedings of the 6th Conference of the International Society for Scientometrics and Informetrics,* Jerusalem, June 16–19, 1997. Reprinted in *Scientometrics, 40*(3) (1997), 455–464.

Egghe, L. (2000). New informetric aspects of the Internet: Some reflections—Many problems. *Journal of Information Science, 26*(5), 329–335.

Garfield, E. (1999). From the World Brain to the informatorium—With a little help from Manfred Kochen. Presented at the Symposium in Honor of Manfred Kochen at the University of Michigan, September 21, 1999. Retrieved November 18, 2000 from http://www.garfield.library.upenn.edu/papers/kochen_worldbrain.html.

Goodrum, A. A.; McCain, K. W.; Lawrence, S.; & Giles, L. C. (2001). Scholarly publishing in the Internet age: A citation analysis of computer science literature. *Information Processing and Management, 37*(5), 661–675.

Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: The retrieval effectiveness of search engines. *Information Processing and Management, 35*(2), 141–180.

Haas, S. W., & Grams, E. S. (2000). Readers, authors and page structure: A discussion of four questions arising from a content analysis of Web pages. *Journal of the American Society for Information Science, 51*(2), 181–192.

Harter, S. P. (1996). The impact of electronic journals on scholarly communication: A citation analysis. *The Public-Access Computer Systems Review, 7*(5). Retrieved November 12, 2000 from http://info.lib.uh.edu/pr/v7/n5/hart7n5.html.

Harter, S. P. (1998). Scholarly communication and electronic journals: An impact study. *Journal of the American Society for Information Science, 49*(6), 507–516.

Henzinger, M. R.; Heydon, A.; Mitzenmacher, M.; & Najork, M. (1999). Measuring index quality using random walks on the Web. *Computer Networks—The International Journal of Computer and Telecommunications Networking, 31*(11–16), 1291–1303 (Proceedings of the 8th International World Wide Web Conference, May 1999). Retrieved February 11, 2002 from http://www8.org/w8-papers/2c-search-discover/measuring/measuring.html

Huberman, B. A., & Adamic L. A. (1999). Growth dynamics of the World Wide Web. *Nature, 401*(6749), 131.

Huberman, B. A.; Pirolli, P. L. T.; Pitkow, J. E.; & Lukose, R. M. (1998). Strong regularities in World Wide Web surfing. *Science, 280*(5360), 95–97.

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.

Kaminer N., & Braunstein Y. M. (1998). Bibliometric analysis of the impact of Internet use on scholarly productivity. *Journal of the American Society for Information Science, 49*(8), 720–730.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms,* 1998. Reprinted in *Journal of the ACM, 46*(5) (1999), 604–632. Retrieved November 12, 2000 from http://www.cs.cornell.edu/home/kleinber/auth.ps.

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.

Koehler, W., et al. (2000). A bibliometric analysis of select information science print and electronic journals in the 1990s. *Information Research, 6*(1). Retrieved November 25, 2000 from http://www.shef.ac.uk/~is/publications/infres/paper88.html.

Kovacs, D. K.; Robinson, K. L.; & Dixon, J. (1995). Scholarly e-conferences on the academic networks. *Journal of the American Society for Information Science, 46*(4), 244–253.

Kumar, R.; Raghauam, P.; Rajagopalan, S.; & Tomkins, A. (1999). Trawling the Web for emerging cyber-communities. *Computer Networks, 31*(11), 1481–1493.

Larson, R. (1996). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. In *Proceedings of the ASIS Annual Meeting, 33,* 71–78. Medford: Information Today, Inc. Retrieved November 12, 2000 from http://sherlock.berkeley.edu/asis96/asis96.html.

Lawrence, S.; Bollacker, K.; & Giles, C. L. (1999). Digital libraries and autonomous citation indexing. *IEEE Intelligent Systems, 32*(6), 67–71.

Lawrence, S., & Giles, C. L. (1998). Searching the World Wide Web. *Science, 280*(5360), 98–100.

Lawrence, S., & Giles, C. L. (1999). Accessibility and distribution of information on the Web. *Nature, 400*(6740), 107–109.

Lazinger, S. S.; Bar-Ilan, J.; & Peritz, B. C. (1997). Internet use by faculty members in various disciplines: A comparative case study. *Journal of the American Society for Information Science, 48*(6), 508–518.

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. Retrieved November 12, 2000 from http://www.cindoc.csis.es/cybermetrics/articles/v4i1p2/html.

Life on the Internet. (n.d.). Retrieved November 12, 2000 from http://www.pbs.org/internet/timeline.

Moore, A., & Murray, B. H. (2000). *Sizing the Internet.* Retrieved November 14, 2000 from http://www.cyveillance.com/web/us/downloads/Sizing_the_Internet.pdf.

Oppenheim, C.; Morris, A.; McKnight, C; & Lowley, S. (2000). The evaluation of WWW search engines. *Journal of Documentation, 56*(2), 190–211.

Page, L., & Brin, S. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30,* 107–117 (Proceedings of the 7th International World Wide Web Conference, April 1998). Retrieved November 15, 2000 from http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm.

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.

Piper, A. (1998). Conducting social science laboratory experiments on the World Wide Web. *Library and Information Science Research, 20*(1), 5–21.

Rosenbaum, H. (1998). Web-based community networks: A study of information organization and access. In *Proceedings of the ASIS Annual Meeting, 35,* 516–530. Medford: Information Today, Inc.

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), Retrieved November 20, 2000 from http://www.cindoc.es/cybermetrics/articles/v1i1p1.htm.

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and Northern Light. *Cybermetrics, 2/3*(1), paper 2, Retrieved August 20, 2000 from http://www.cindoc.csis.es/cybermetrics/articles/v2i1p2.html.

Savolainen, R. (1998). Use studies of electronic networks: A review of empirical research approaches and challenges for their development. *Journal of Documentation, 54*(3), 332–351.

Schoch, N. A., & White, M. D. (1997). A study of the communication patterns of participants in consumer health electronic discussion groups. *Proceedings of the ASIS Annual Meeting, 34,* 280–292. Medford: Information Today, Inc.

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

Spink, A.; Bateman, J.; & Jansen, B. J. (1999). Searching the Web: A survey of Excite users. *Internet Research: Electronic Networking Applications and Policy, 9*(2), 117–128.

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.

Watson, J. S. (1998) "If you don't have it, you can't find it." A close look at students' perceptions of using technology. *Journal of the American Society for Information Science, 49*(11), 1024–1036.

Woodruff, A.; Aoki, P. M.; Brewer, E.; Gauthier, P.; & Rowe, L. A. (1996). An investigation of documents from the World Wide Web. *Computer Networks and ISDN Systems, 28,* 963–980 (Proceedings of the 5th International World Wide Web Conference, May 1996). Retrieved November 14, 2000 from http://www5conf.inria.fr/fich_html/papers/P7/Overview.html.

Zhang, Y. (1998). The impact of Internet-based electronic resources on formal scholarly communication in the area of library and information science: A citation analysis. *Journal of Information Science, 24*(4), 241–254.

Zhang, Y. (2000). Using the Internet for survey research: A case study. *Journal of the American Society for Information Science, 51*(1), 57–68.