Information-Seeking Behavior of Chemists: A Transaction Log Analysis of Referral URLs

Philip M. Davis
Albert R. Mann Library, Cornell University, Ithaca, NY 14853-4301. E-mail: pmd8@cornell.edu

This study reports an analysis of referral URL data by the Cornell University IP address from the American Chemical Society servers. The goal of this work is to better understand the tools used and pathways taken when scientists connect to electronic journals. While various methods of referral were identified in this study, most individuals were referred infrequently and followed few and consistent pathways each time they connected. The relationship between the number and types of referrals followed an inverse-square law. Whereas the majority of referrals came from established finding tools (library catalog, library e-journal list, and bibliographic databases), a substantial number of referrals originated from generic Web searches. Scientists are also relying on local alternatives or substitutes such as departmental or personal Web pages with lists of linked publications. The use of electronic mail as a method to refer scientists directly to online articles may be greatly underestimated. Implications for the development of redundant library services such as e-journal lists and the practice of publishers to allow linking from other resources are discussed.

Introduction

The networked environment now provides scientists with many pathways to published journal literature, from online bibliographic databases to the informal e-mail distribution of article links between colleagues. While there is a substantial body of research that has focused on the information-seeking behavior of scientists, this body of knowledge is incomplete, as the information environment continues to evolve.

Understanding the information-seeking behavior of scientists has great significance not only to libraries, which spend considerable acquisition funds purchasing finding tools for the literature, but also to publishers, who invest in the technological infrastructure to make their electronic journals available.

Received July 1, 2003; revised August 29, 2003; accepted August 29, 2003

This study investigates the information-seeking behavior of scientists by analyzing the transaction logs of the American Chemical Society (ACS) for members of the Cornell University community. Specifically, it focuses on referral URLs—the location from which an individual was referred to their site.

Literature Review

Importance of Journal Literature to Chemists

Whereas the time spent reading and the number of articles read varies considerably across subject disciplines, chemists are known to be heavy users of journal literature (C. H. Brown, 1956), spending more time reading than any other group of scientists (Tenopir & King, 2002). Because of their heavy dependence on journal literature, chemists are an ideal group to study with regard to their information-seeking behavior.

In their study of the use of computer networks by scientists, John Walsh and Todd Bayma found that scholars in fields that are tightly linked to commercial markets, like chemistry, tend to limit their use of informal use of communication by computer networks (i.e., use of e-mail and preprint servers) and rely more on the formal communication of published articles (Walsh & Bayma, 1996). A survey of scientists in nine disciplines confirmed that chemists are the least likely to rely on e-print servers. The most frequent response was that the posting of preprints was against the policy of many chemistry publishers (Lawal, 2002). Only 6% of editors of top chemistry journals will publish articles that have appeared as e-prints (C. M. Brown, 2003). An informal survey of preprints authors using ChemWeb illustrated that broad dissemination of the research was the most important reason for using the system (Warr, 2003). It is ACS policy not to accept articles that have been previously published on preprint servers, although Chemical Abstracts has begun to index articles on preprint servers.

Scientists in general employ a number of different methods while searching for information. Tenopir and others describe the changing behavior of scientists during a period

of electronic migration of scholarly literature, from the early 1990s to the early 2000s (Tenopir et al., 2003). The authors report that browsing for information has been steadily declining, from about 58% of the time in the early 1990s to 21% at present. Online searching of bibliographic databases has risen quite substantially from 8.5% to 39%. At the same time the scientists' reliance on citations to find relevant literature has gone up from 5.6% to 16%.

Transaction Log Analysis

Transaction log analysis is a non-intrusive method for collecting data from a large number of individuals for the purpose of understanding online-user behavior. It has been employed to better understand what individuals do when they visit a library Web site (Rozic-Hristovski, Hristovski, & Todorovski, 2002), and to make successive improvements to a library's catalog (Blecic, Dorsch, Koenig, & Bangalore, 1999). It has been employed to track the online behavior of Web users (Thelwall, 2001), their use of particular bibliographic databases (Cooke, Kopelev, Schofield, Boyce, & Dunne, 2002), or the use of full-text journal packages (Institute for the Future, 2002; Ke, Kwakkelaar, Tai, & Chen, 2002). It was most notably used in the superbly documented SuperJournal project in order to understand individuals' use of online journals (Eason, Richardson, & Yu, 2000; Pullinger, 1994). Because individuals using the SuperJournal system were required to register and provide basic demographic information about themselves, researchers could directly track individual behavior.

An excellent review of Web searching studies is summarized by Jansen and Pooch (2001), and an older history of transaction log analysis is provided by Tom Peters (Peters, 1993).

Electronic Journal Use Studies

Much of the published research to date has focused on the journal, the publisher, or the consortium as the level of analysis. Very little is known about the patterns of *individual use* of electronic journals. Using an IP address as a surrogate for individuals, Davis and Solla recently reported an analysis of ACS e-journal full-text downloads for Cornell University (Davis & Solla, 2003). Their results are similar to the SuperJournal study in that the majority of users limited themselves to a small number of both journal and article downloads, and a small minority of heavy users was responsible for the majority of total journal downloads. A study by Stanford University Libraries and HighWire Press tracked individuals' use of 14 biomedical journals for a single day (Institute for the Future, 2002). What was most revealing about their study was individuals' sequence of events. Individuals who downloaded the PDF version of the article were very likely to have downloaded the same article in HTML. The results of their study will have great implications for the interpretation of usage statistics.

Data Definitions

Referral URL

A referral URL is a Web address that directs (or refers) a browser to another address. In practical terms within this study, a user may be referred to another address when using:

- Any Web page containing a link to the ACS e-journal server;
- A library catalog;
- A bibliographic database containing links to the full-text article;
- A full-text article containing a link to another article; or
- A Web-based e-mail program containing an embedded URL.

In general, referral addresses are not provided when:

- A user connects by using a browser's bookmark;
- A user is making referrals within the same domain; or
- The referral comes from a non-Web application (e.g., client e-mail software).

Based on the last definition, the dataset under investigation will only indicate when a user makes the first connection to the ACS e-journal server, and will ignore further connections that are made within the site. This study is not designed to investigate the total number of e-journal downloads, previously investigated by the author (Davis & Solla, 2003), but attempts to answer how scientists *locate* published articles.

Dataset

The dataset represents three months of referral data (Dec. 2002–Feb. 2003), for all Cornell IP addresses connecting to the ACS servers. The main server (pubs.acs.org), which hosts the publisher's e-journals, represents about 90% of the total visits to ACS servers. Based on a previous analysis of ACS data (Davis & Solla, 2003),this time period should be considered a proportional sample of use at Cornell, although somewhat smaller than samples taken at other times during the year. During this data collection period, there were 15,876 Web connections containing referral URLs from 1,630 unique IP addresses. Because they reflected internal ACS referrals from non-journal sites (e.g., ACS ChemJobs, Faculty Directory, Meetings, ACS Style Manual, etc.), 5,927 referrals from 39 unique IPs were discarded from the dataset, leaving 9,949 valid referrals from 1,591 unique IP addresses.

User

The goal of this study is to better understand individual user behavior. While this sounds simple in theory, in practice, it is very difficult to measure in the current online environment. Librarians, on principle, have defended their patrons' right to confidentiality and have successfully ar-

TABLE 1. Frequency of referrals by type.

Referral type	Total referrals	% Total referrals	Unique IPs	Referrals per IP
Library catalog	2,482	24.9	552	4.5
Bib database	2,372	23.8	324	7.3
E-journal list	1,813	18.2	405	4.5
Web page	1,108	11.1	190	5.8
Web search	996	10.0	491	2.0
E-mail (Web-based)	592	6.0	79	7.5
Article link	571	5.7	204	2.8
Other	15	0.2	9	1.7
Total Referrals	9,949	100.0	1,591	6.3

gued against required logins that would enable researchers to directly measure individual behavior.

Because of this limitation, IP addresses are used as a surrogate measure. While an IP address denotes an individual *computer* and not necessarily an individual *user*, it will be used to gain a better understanding of individual online behavior beyond what is currently known. Some of the computers located on a university campus in libraries, computer labs, and departments are shared by several people. In addition, the Cornell Library Proxy Server allows individuals connecting from outside the Cornell network to gain access to secure resources. These types of computers may be regarded as aggregate users. Attempts will be made to identify aggregate data points in this study when they invoke a large degree of leverage on the statistical models. The confidentiality of individual users was maintained at all stages of research.

Categorization of Referrals

Each URL was categorized based on type of referral (article link, bibliographic database, electronic journal list, e-mail, library catalog, Web page, Web search, and other). An electronic journal list is defined as a Web page (static or dynamic) that provides users with either selective or definitive lists of e-journals with links directly to the publishers' sites. Whereas many online catalogs provide the same type of linking, the creation of e-journal lists was accomplished primarily for the function of title browsing and quick lookup. Since the e-journal list is also a relatively new development for libraries compared to the online catalog, it was also important to gain a sense of popularity and preference for this service over searching the comprehensive holdings of the library catalog.

Within each of the categories, individuals may be referred by more than one source. For example, during the three-month observation a patron may be referred from both Chemical Abstracts *and* Medline—both bibliographic databases. In order to get a sense of the total number of different pathways employed by scientists in the referral process, it was also necessary to measure the number of referrals by domain name.

Based on the methodology described, it is impossible to discern the full information-seeking pathway a scientist followed to the published literature. The only part of the pathway to which we were privy was the last referral to the ACS server.

Observations

By Type of Referral

The frequency of referral is presented in Table 1. The most frequent types of referral documented in this study came from the library catalog and bibliographic databases—two traditional tools used by researchers. Regarding bibliographic databases, 84% of referrals within this category came directly from SciFinder Scholar, a database of chemistry abstracts. PubMed accounted for 15% of the bibliographic referrals.

Referral by e-journal list accounted for 18.2% of all referrals. Cornell University Library's e-journal list accounted for only about 21% within this class. Lists provided by individual libraries within Cornell accounted for almost 72%. Specifically, the Physical Sciences Library (which is the primary library supporting chemistry), accounted for 39% within this category. Individuals were also documented using e-journal lists from other major universities in the United States and around the world (3%). Personal and departmental e-journal lists from within Cornell were also documented (4%).

Within the Web page category (Table 2), the most frequent type of referral came from ACS Journal Web pages. Web-based news sources were also frequently observed, the most frequent referral was from *Chemical and Engineering News*, a magazine produced by the ACS. Departmental, lab, and individual faculty pages were also frequently the source of referral. Specifically, these pages included lists of published articles and links to the full text.

Ten percent of the referrals in this study came from generic Web searches, 81% of them from Google, followed by MSN (8%) and Yahoo (6%). Based on a cursory analysis of the search string included in the referral, these individuals were using these search engines either to locate journal titles or to locate specific articles employing a complex search

TABLE 2. Breakdown of Web page referrals by type.

Web page referral	Frequency	Percent
ACS Journal Web page	366	33.0
News	272	24.5
Department/lab	200	18.1
Faculty	75	6.8
Course Web page	43	3.9
Commercial	31	2.8
Organization	21	1.9
Personal	19	1.7
Other	81	7.3
Total	1,108	100.0

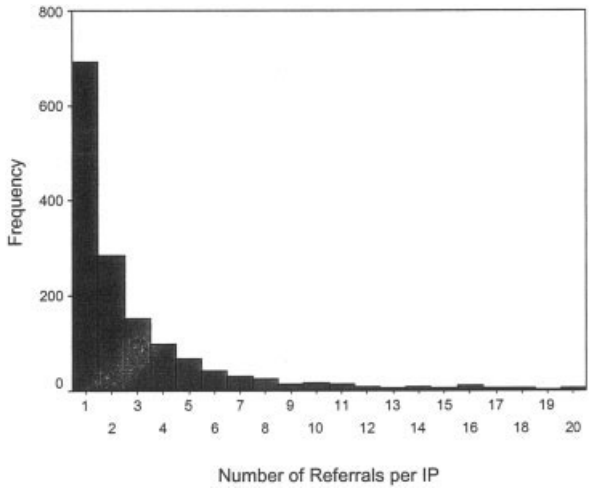

FIG. 1. Histogram of user visits by IP. (N = 1591)

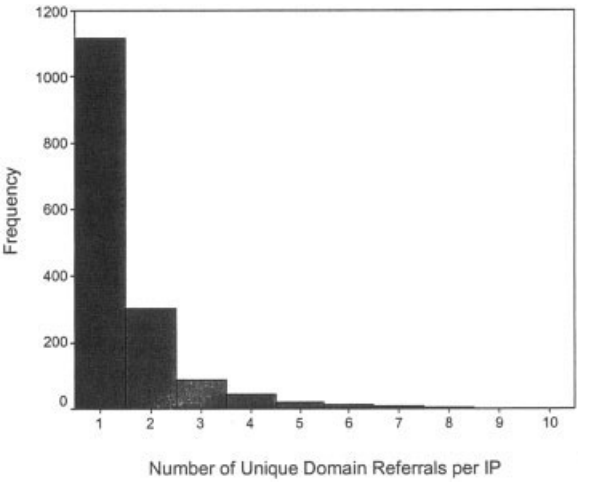

FIG. 2. Histogram of referral domains by IP. (N = 1591)

algorithm that is more typically conducted from bibliographic databases.

Referrals by e-mail amounted to 6% of total referrals. It should be noted that only Web-based e-mail could be tracked. Since most of the academic community at Cornell uses client-based e-mail (Eudora), the majority of e-mail referrals would not be documented.

Article linking comprised 5.7% of total referrals. Referrals from the Digital Object Identifier (DOI) server (dx.doi.org) was the most prevalent method (about 82%). Referrals from individual journals, like *Nature*, *Science*, and *Proceedings of the National Academy of Sciences*, were also identified.

Only 9 (or 0.2%) of the 9,949 referrals could not be determined, either because these came from Web pages that no longer exist, were generated dynamically, or contained languages that were not discernable to the researcher (e.g., Korean).

Most Individuals Were Referred Infrequently

The 9,949 referrals to the ACS site were associated with 1,591 unique IP addresses. A histogram of the number of visits illustrates great skew in the data (Figure 1). Most of the IPs (aka "users") were referred to the ACS site very infrequently over the three-month period, yet a small number of IPs were associated with a high number of referrals. Forty-four percent of IPs were referred to the site only once, 61% were referred to the site two or fewer times, and 71% were referred to the site three or fewer times. The library proxy server, an aggregate of all off-campus users, was associated with 324 referrals.

Individuals Follow Few and Consistent Pathways to Information

Users in general followed few and consistent methods to locate information. The number of different referral domains per IP address is presented in Figure 2. Seventy percent of unique IP addresses were referred from only one domain, 89% from two or fewer domains, and 95% from three or fewer domains. In general, IP addresses that were identified as representing aggregate users were associated with more sources of referrals. The library proxy server, as an example, was referred by 23 different domains.

Relationship Between Domains and Referrals

The relationship between the number of domains and number of referrals is quadratic in nature (Figure 3). Four data outliers were removed from the regression analysis since they had a very high influence (or leverage) on the statistical model. All four outliers were associated with an extremely high number of visits. One of these outliers was the Library Proxy Server, one originated from the Medical College (A), and two were from individual computers

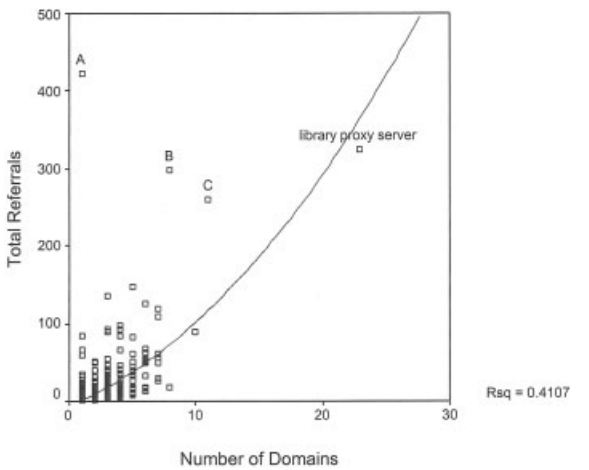

FIG. 3. Relationship between number of domains and number of referrals. Includes four outliers (library proxy server, A, B, C). Each data point represents one IP address (or "user"). (N = 1587)

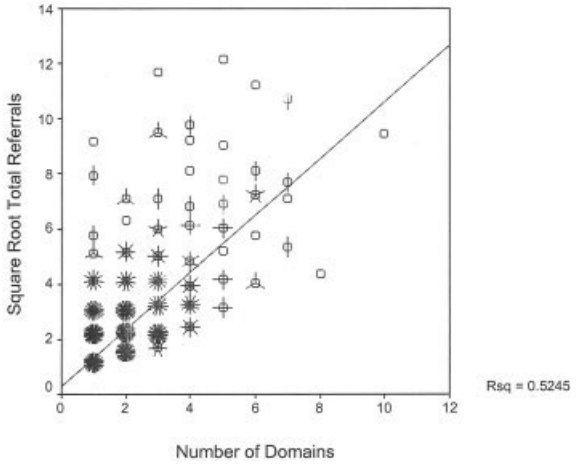

FIG. 4. Relationship between number of domains and number of referrals fits inverse-square law. Outliers removed. Each data point represents one IP address (or "user"). (N = 1587)

within the Engineering Materials Research Department (B and C).

The best fit of the data to a linear model was achieved by taking the square root of the number of total referrals (Rsq = 0.52), which resulted in no serious violations of the regression model (Figure 4). The fit of the data to this model suggests that an inverse-square law is in effect.

Discussion

How Chemists Learn About the Articles They Read

Because of the size of the field of chemistry and the sheer amount of literature published each year, chemists employ a number of techniques to survey the literature, including browsing articles by title, abstract, graphics, and captions (Olsen, 1994). The heavy use of bibliographic sources (especially SciFinder Scholar) to identify and summarize the chemical literature seems consistent with what is known about the information-seeking behavior of chemists.

The relatively high referrals from generic Web searches and Web pages in this study may indicate that these sources are providing substitutes for traditional tools such as the library catalog and bibliographic databases. A graduate student may find a more comprehensive list of prospective faculty publications by visiting his/her site rather than trying to piece together publications from various indexing databases. A librarian trying to verify a poorly documented citation may conceive of doing a generic Web search hoping to find additional information from a departmental Web page. A researcher may search Google for a known journal title rather than go through the process of searching the library catalog and having to sift through the results that contain paper and online holdings. The very popularity of the e-journal lists in comparison to the library catalog may indicate that a single finding tool is not sufficient and efficient for all types of needs. Pirolli and Card (1999) describe the strategies of how individuals search for information in the same terms as how animals forage for food. Based on their theory of "information foraging" individuals will modify their strategies to maximize their rate of gathering valuable information. As there is no optimal foraging technique for all animals, information foragers will adapt to maximize their gathering within individual niches. In other words, the model of information foraging assumes that individuals will gravitate toward different forms of searching.

Electronic mail was a substantial source of referral in this study. It should be noted again that only Web-based e-mail messages would have been recorded in this study. From the late 1980s, the use of e-mail by academics has been increasing and substituting for other forms of communication (e.g., surface mail, telephone, etc.) (Schaefermeyer & Sewell, 1988). The success of e-mail within the research community is believed to have flourished because it extends the concept of the invisible college and reduces proximity between colleagues (Carley & Wendt, 1988). A future survey of how scientists were referred to the journal literature may indicate that e-mail referral is more common than indicated in this study.

This study could also not determine the extent of use of Web browser bookmarks, since a referral URL was not provided in the connection. A recent survey on the use of electronic journals at the University of Edinburgh reported that bookmarking of electronic journals was a principal mode of access to the literature for academic staff (Bonthron et al., 2003).

Inverse-Square Law (Lotka's Law)

Whereas the population as a whole employed various pathways to reach the ACS journal information, most individuals relied on few and consistent methods. There was however a quadratic relationship (specifically an inverse-square relationship) between the number and types of referrals. It is not entirely understood why heavier users of ACS journals would rely on proportionally more methods of referral. This same inverse-square law was discovered in the author's analysis of the relationship of the number of journals read and the number of articles downloaded in a previous study of ACS e-journals (Davis & Solla, 2003).

The first account of the inverse-square relationship in bibliometrics was described by Alfred Lotka, who counted the frequency of authors indexed in Chemical Abstracts. The number of authors contributing n articles is approximately $1/n^2$ (Lotka, 1926). This relationship is commonly referred to as Lotka's Law. Derek de Solla Price later wrote that several laws in information science (Lotka's Law, Bradford's Law, Pareto's Law, and Zipf's Law) may all be described by what he termed a "cumulative advantage distribution"—a distribution similar to the negative binomial (Price, 1976).

Article Linking

The only other bibliographic database found in the transaction logs was PubMed, the public version of Medline. During this study, the ACS only had linking agreements with the Chemical Abstract Service, products using its ChemPort linking service, and PubMed. This fact explains why other relevant bibliographic databases (e.g., Web of Science, Biosis, Agricola, CAB, Food Science Technology Abstracts, etc.) were not present in the transaction log. Cornell researchers using other databases would be directed back to the library's catalog before linking to the ACS server. This reason alone may explain the very high proportional use of the catalog.

Lastly, because the server logs from the ACS do not include internal referrals, it is not known how frequently scientists link to other ACS journals from within their site.

Generalizability to Other Publishers

Although the basic findings of this research may be generalized to other scientific publishers that provide electronic access to their journal literature, the American Chemical Society is different in some respects. It publishes a small number of journals compared to other STM publishers, and has established prestige and brand name–recognition for its titles. This may be partially responsible for a high number of referrals from general Web searches. Many ACS titles are also considered core reading within fields of chemistry and related disciplines, and so the type of reading done by these scientists may include a much higher proportion of browsing and current awareness. This may partially explain the high number of referrals from the ACS Journal pages.

In comparison, other publishers whose collections include less prestigious titles may find proportionally higher referrals from bibliographic databases, or library catalogs than referrals from within the site, e-mail, or article links.

Implications for Libraries

This study demonstrates that scientists follow a number of different pathways to scholarly information, but individuals depend on very few and consistent methods. These findings suggest that the library should create redundancy in the tools that guide its patrons to the literature. For example, the Cornell University Library catalog includes URLs to electronic journals, which are also duplicated in a Cornell Library e-journal list. It was argued that the creation of the e-journal list represented an inefficient use of time and resources since the information was already contained in the catalog. The results of this study demonstrate the popularity of both of these tools—in essence, demonstrating two redundant but complementary services.

What was also intriguing from the results was that labs and departments created and relied upon their own list of relevant e-journals, in spite of the fact that individual campus libraries have created their own subject-based lists. This demonstrates that scientists will create local and personal tools to increase efficient connections to what they consider to be their core literature.

Implications for Publishers

From the perspective of scientists, it is in their interest to have the electronic literature linked to as many types of information referral as possible. A publisher's rationale for limiting direct linking from other databases and full-text products may be as much political as technical, and for that reason, the reasons ACS limits linking beyond its own CAS products and Medline will not be explored. Chemical Abstracts is one of the primary finding tools for the chemical literature, and for that reason, the core users of ACS titles may feel sufficiently supported. Scientists using other finding tools are not prevented from using ACS e-journals; it will just take them more time.

Adoption of standards that ensure direct linking from other resources will help provide seamless access to a publisher's content, reduce the number of steps it takes a user to get to the desired content, and ultimately save the reader time.

Acknowledgements

Sincere appreciation goes to Al Funk and Cheryl Mathews at the American Chemical Society for preparing and providing the data involved in this study, and several colleagues for providing substantial feedback: Leah Solla, Cornell University; Bill Walters, St. Lawrence University; and Carol Tenopir, University of Tennessee.

References

Blecic, D.D., Dorsch, J., Koenig, M., & Bangalore, N. (1999). A longitudinal study of the effects of OPAC screen changes on searching behavior and searcher success. College & Research Libraries, 60, 515–530.

Bonthron, K., Urquhart, C., Thomas, R., Armstrong, C., Ellis, D., Everitt, J., et al. (2003). Trends in use of electronic journals in higher education in the UK—Views of academic staff and Students. D-Lib Magazine, 9(6). Available: http://www.dlib.org/dlib/june03/urquhart/06urquhart.html

Brown, C.H. (1956). Scientific serials: Characteristics and lists of most cited publications in mathematics, chemistry, geology, physiology, botany, zoology, and entomology (Vol. ACRL Monograph no. 16). Chicago: Association of College and Reference Libraries.

Brown, C.M. (2003). The role of electronic preprints in chemical communication: Analysis of citation, usage, and acceptance in the journal literature. Journal of the American Society for Information Science and Technology, 54, 262–271.

Carley, K., & Wendt, K. (1988, Aug 24–28). Electronic mail and the diffusion of scientific information: The study of SOAR and its dominant users. Paper presented at the 83rd Annual Meeting of the American Sociological Association, Atlanta, GA.

Cooke, F., Kopelev, N., Schofield, H., Boyce, G., & Dunne, S. (2002). Approaches to understanding the searching behavior of CrossFire users. Journal of Chemical Information and Computer Sciences, 42, 1016–1027.

Davis, P.M., & Solla, L. (2003). An IP-level analysis of usage statistics for electronic journals in chemistry: Making inferences about user behavior.

Journal of the American Society for Information Science and Technology, 54, 1062–1068.
Eason, K., Richardson, S., & Yu, L. (2000). Patterns of use of electronic journals. Journal of Documentation, 56, 477–504.
Institute for the Future. (2002). E-journal user study. Report of Web log data mining (SR-786). Menlo Park, CA: Stanford University. 19p. Available: http://ejust.stanford.edu/logdata.html
Jansen, B.J., & Pooch, U. (2001). A review of Web searching studies and a framework for future research. Journal of the American Society for Information Science and Technology, 52, 235–246.
Ke, H.-R., Kwakkelaar, R., Tai, Y.-M., & Chen, L.-C. (2002). Exploring behavior of e-journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan. Library & Information Science Research, 24, 265–291.
Lawal, I. (2002). Scholarly communication: The use and non-use of e-print archives for the dissemination of scientific information. Issues in Science and Technology Librarianship, 36. Available: http://www.istl.org/02-fall/article03.html
Lotka, A.J. (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences, 16, 317–323.
Olsen, J. (1994). Electronic journal literature: Implications for scholars. Westport, CT: Mecklermedia.
Peters, T.A. (1993). The history and development of transaction log analysis. Library Hi Tech, 11(2), 41–58.
Pirolli, P., & Card, S. (1999). Information Foraging. Psychological Review, 106, 643–675.
Price, D.J. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science, 27, 292–306.
Pullinger, D.J. (1994). The SuperJournal project. Philadelphia, PA: Institute of Physics Publishing.
Rozic-Hristovski, A., Hristovski, D., & Todorovski, L. (2002). Users' information-seeking behavior on a medical library Web site. Journal of the Medical Library Association, 90, 210–217.
Schaefermeyer, M.J., & Sewell, E.H. (1988). Communicating by electronic mail. American Behavioral Scientist, 32, 112–123.
Tenopir, C., & King, D. (2002). Reading behaviour and electronic journals. Learned Publishing, 15, 259–265.
Tenopir, C., King, D.W., Boyce, P., Grayson, M., Zhang, Y., & Ebuen, M. (2003). Patterns of journal use by scientists through three evolutionary phases. D-Lib Magazine, 9(5). Available: http://www.dlib.org/dlib/may03/king/05king.html
Thelwall, M. (2001). Web log file analysis: Backlinks and queries. ASLIB Proceedings, 53, 217–223.
Walsh, J., & Bayma, T. (1996). Computer networks and scientific work. Social Studies of Science, 26, 661–703.
Warr, W.A. (2003). Evaluation of an experimental chemistry preprint server. Journal of Chemical Information and Computer Sciences, 43, 357–361.