

Cite Me, Cite My References? (Scholarly Use of the ACM SIGIR Proceedings Based on Two Citation Indexes)

Elana Broch
Princeton University
Donald E. Stokes Library, Wallace Hall
Princeton NJ 08544
ebroch@princeton.edu

ABSTRACT

A three-part study was designed to document Internet use in scholarly research, using the Annual SIGIR Conference Proceedings from 1997 through 1999. The results suggest an increasing trend toward electronic self-publishing. Furthermore, while electronic availability did not insure that one would be cited, the most highly cited articles were available on the "free" web. The study also found that electronic availability has not, in most cases, decreased the length of time between publication and citation.

Keywords

Citation analysis, ResearchIndex, Science Citation Index, scholarly communication, web publishing

1. METHOD

The first part of the study determined which of the proceedings articles were highly cited. The second part of the study focused on the availability of papers on the Web. The third part of the study was an in-depth analysis of two of the most highly cited papers.

1.1 Documents analyzed

This study focused on the papers in the 1997, 1998, and 1999 ACM SIGIR proceedings.

1.2 Citation indexes

Two citation indexes were used to determine citation counts: Research Index [4], a digital library of computer and information science articles (<http://citeseer.nj.nec.com/cs>) and Science Citation Index [3] via Dialog (<http://www.dialogweb.com>). Citation counts were obtained from September through November 2000.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR '01, September 9-12, 2001, New Orleans, Louisiana, USA.
Copyright 2001 ACM 1-58113-331-6/01/0009...\$5.00.

1.3 Locating articles on the Web

In addition to their availability through the ACM Digital Library, many of the articles from the conference are available on the Web. Articles were located using ResearchIndex (RI), Google, or Inquirus, a metasearch engine [5], during the 4th quarter of 2000.

2. RESULTS

2.1 Citation Counts

Tables 1 and 2 summarize the citation counts for each of the SIGIR Proceedings articles, using the two citation indexes. The counts exclude self-citations. As seen in Table 1, for example, one article was cited 24 times in RI (see the rightmost side of the middle row of the table), while 15 articles from the 1999 Proceedings were never cited. Table 2 has the corresponding data for SCI.

The most frequently cited articles in RI tended to be frequently cited articles in SCI. In the 1998 data, Bharat and Henzinger was the most frequently cited article with 24 citations in RI and 12 in SCI. Zamir and Etzioni was the second most widely cited in both indexes: 19 times in RI and 9 times in SCI. In the 1997 proceedings, Ballesteros & Croft and Hearst were the most highly cited articles based on both indexes. Baumgarten was cited as often as Hearst based on RI ($f=13$), but was not heavily cited in SCI.

The citation rates for 1999 seemed very low. In order to see if SIGIR 1999 articles were cited, but the citation had not been documented by a citation index, the reference lists from the SIGIR 2000 articles were examined to see how many of the articles from the preceding proceedings they contained.

The 39 articles in the 2000 Proceedings included eight citations to 1997 SIGIR papers and 37 to 1998 papers. Among the 21 citations to 1999 articles, 10 were self-citations. Thus, it appears that the low citation rates for 1999 articles are not solely a function of the length of time it takes a citation index to locate a citation. There still is, in most cases, a two-year delay between an article being presented and cited, confirming Harter and Kim's [2] findings.

2.2 Electronic Availability

The majority of Proceedings articles are available on the Web. The values for 1997, 1998, and 1999, were 52, 65, and 71 percent. Almost all of the highly cited articles were available on the "free" web.

Table 1. Number of citations in ResearchIndex for a given article in the SIGIR proceedings, by year

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
1997 (n=36)	5	5	3	6	4	1	2	2	1	2	1	1	0	2	1	0	0	0	0	0	0	0	0	0	0	0
1998 (n=40)	5	8	3	3	5	5	2	1	1	1	1	1	0	1	0	0	1	0	0	1	0	0	0	0	0	1
1999 (n=34)	15	9	3	3	2	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2. Number of citations in Science Citation Index for a given article in the SIGIR proceedings, by year

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
1997 (n=36)	6	6	6	9	4	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1998 (n=40)	12	12	5	3	3	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1999 (n=34)	28	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

2.3 Analysis of highly cited articles

The reference lists of the two most highly cited articles from the 1998 Proceedings--Bharat & Henzinger and Zamir & Etzioni--were analyzed. I wanted to determine if, when someone cites an article, they are likely to cite the references in that article.

The 26 articles (including 2 self-citations) that cited Bharat and Henzinger were analyzed by comparing the citations in their reference lists (the children) with the 30 articles (the grandparents) in Bharat and Henzinger's reference list.

Overall, few of the grandparents were cited by the children; 2/3 of the articles were cited once or not at all. At the other end of the spectrum, one article was cited by 24 of the 26 articles that cited Bharat and Henzinger. In fact, many of the articles cited by the children were actually published or presented more recently than Bharat and Henzinger (i.e., 1998). Often a citation would be to a more recent article by one of the author's on Bharat and Henzinger's reference list, reflecting the fast-paced nature of Web IR.

Each of the articles on the Bharat and Henzinger reference list was also checked in SCI. A similar pattern of results was obtained: the majority of articles in their reference list (in this case, 4/5) had 0 or 1 citation(s).

3. CONCLUSIONS

Researchers in the field of information retrieval are using the Web to make their research available. About 2/3 of the papers from the 1998, 1999, and 2000 SIGIR were located on the "free" Web, in addition to being included in the ACM Digital Library. This represents an increase from the 1997 results, which found that about 1/2 of the papers were located on the Web.

Although a large portion of the IR literature is available on the Web, this availability does not assure that researchers will use or cite them. While researchers are increasingly using the Web for exchanging information, most do not routinely include URLs even when they exist in stable locations (e.g., the ACM Digital Library).

Based on the in-depth analysis of two articles, there are definite "authorities" [1]. The two highly cited articles that were studied in detail, both in the field of Web IR, had distinct reference lists with some overlap. Within each article, however,

some of the authorities (e.g. Kleinberg) were referenced by almost every article that cited the parent article (e.g., Bharat & Henzinger).

4. ACKNOWLEDGMENTS

I would like to thank Tefko Saracevic of Rutgers University and Steve Lawrence and Val Tucci of the NEC Research Institute for their support throughout this project.

5. REFERENCES

- [1] Clever Project (1999). Hypersearching the Web. Scientific American. Accessed 12 July 2000 from www.sciam.com/1999/0699issue/0699raghavan.html
- [2] Harter, S. P. and Kim, H.J. (1996). Electronic journals and scholarly communication: A citation and reference study. Proceedings of the Midyear Meeting of the American Society for Information Science, San Diego, CA, May 20-22, 1996, pp. 299-315. Accessed 18 July 2000 from <http://ezinfo.ucs.indiana.edu/~harter/harter-asis96midyear.html>
- [3] Institute for Scientific Information (2000). Science Citation Index. Philadelphia, PA.
- [4] Lawrence, S., Bollacker, K., and Giles, C.L. (1999). Indexing and retrieval of scientific literature. Eighth International Conference on Information and Knowledge Management, CIKM 99, Kansas City, Missouri, November 2-6, pp. 139-146. Accessed 21 November 2000 from <http://www.neci.nec.com/~lawrence/papers.html>
- [5] Lawrence, S. and Giles, C.L. (1998) Context and page analysis for improved Web search. *IEEE Internet Computing*, 2(4), 38-46. Accessed 1 November 2000 from <http://www.neci.nec.com/~lawrence/papers/search-ic98/>