PERGAMON

Information Processing and Management 37 (2001) 661–675

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Scholarly publishing in the Internet age: a citation analysis of computer science literature

Abby A. Goodrum [a,*], Katherine W. McCain [a], Steve Lawrence [b], C. Lee Giles [b]

[a] *College of Information Science and Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104-2875, USA*

[b] *NEC Research Institute, 4 Independence Way, Princeton, NJ 08540-6634, USA*

Accepted 8 September 2000

Abstract

The Web is revolutionizing the entire scholarly communication process and changing the way that researchers exchange information. In this paper, we analyze two views of information production and use in computer-related research based on citation analysis of PDF and Postcript formatted publications on the Web using autonomous citation indexing (ACI), and a parallel citation analysis of the journal literature indexed by the Institute for Scientific Information (ISI) in SCISEARCH. Our goal is to establish a baseline profile of computer science "literature" as it appears in the published journals and as it appears on the publicly available Web. From this starting point, we hope to identify additional research areas dealing with information dissemination and citation practices in computer science and the utility of autonomous citation indexing on the Web as an adjunct to commercial indexing © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Citation analysis; Computer science; Scholarly publishing; World Wide Web

1. Introduction

For some years now, there has been a great deal of discussion regarding the role and relative importance of conference proceedings, scholarly journals and monographs as significant channels for published research (Garvey, Lin, & Tomita, 1972a,b; Drott, 1995; Lindholm-Romantschuk & Warner, 1996). This topic has been of particular interest to scholars in various computer-related fields, where some conference proceedings are seen as more timely, more cutting-edge, and more

* Corresponding author. Tel.: +1-215-895-6627; fax: +1-215-895-2494.
E-mail address: goodruaa@drexel.edu (A.A. Goodrum).

PII: S0306-4573(00)00047-9

strictly refereed than some journals. Some journals, by contrast, are considered useful primarily as archival documents and often not even that. Books are often not considered to be sources of timely information to further R&D. These issues are particularly troublesome at tenure and promotion time in interdisciplinary departments when candidates offer their publications for committee scrutiny (heavy on conference presentations and published proceedings) and provide citation counts to demonstrate the impact of their research. There appears to be a lack of correspondence between "traditional" academic expectations of scholarly performance and current practice in many computer-related research communities.

Communication and publication patterns in various areas of the sciences, including computer-related fields, are also being redirected in response to the opportunities for digital archiving and online distribution offered by the Internet/World Wide Web. The Web has provided a new communication channel for traditional publication of scholarly research as well as the dissemination of informal research discussion. Journal publishers are taking advantage of this new outlet and are publishing electronic versions of their traditional print titles as well as new "Internet only" e-journals. The latter, if they meet the standards for inclusion, are indexed by the Institute for Scientific Information (ISI), and thus provide additional data for citation analyses in addition to those generated by the selected lists of print journals they have always covered.

However, more importantly, the Web is revolutionizing the entire scholarly communication process and changing the way that researchers exchange information. Scholarly communication in the sciences – particularly physics, mathematics, and computer science – is moving increasingly toward a new publishing model that emphasizes conference papers, preprint archives, and the online availability of articles. Although much of this communication will eventually make its way into the traditional published literature as journal articles, the time required for publication and citation indexing may be too slow for the progress of research and development in the sciences (Crawford, Hurd, & Weller, 1996; Ginsparg, 1997). Authors, institutions, and archives are making formal research publicly available on their websites in PDF, Postscript, and other formats. In fields such as computer science, significant research now often appears on the Web before it is published in conference proceedings, journals and books.

With the exception of material in preprint archives, whose existence is well known to the relevant research communities, this new electronic "gray literature" is much less visible and accessible than the print and electronic journal literature. It is not being indexed by ISI or other bibliographic databases and can be discovered only through searching the Web relying on the relative power and coverage of search engines such as AltaVista and HotBot, or metaengines like MetaCrawler. It has been difficult to retrieve and even more difficult to examine systematically due to the limitations of Web search engines (Lawrence & Giles, 1998, 1999). Furthermore, the major Web search engines do not index the contents of Postscript and PDF files.

Autonomous citation indexing (ACI) provides a new tool to access the literature on the Web. An ACI system automatically locates articles (on the Web or in other electronic databases) based on keyword lists. It extracts citations from documents, identifies citations to the same article occurring in different citation formats, and provides the context of citations within the body of documents. It should be pointed out that we are referring to articles available on personal or institutional Web pages whether or not the works are also available as formally published articles, book chapters, conference papers, etc. The ACI system discussed in this paper does not currently index documents from e-journals or other sources that are not freely available to the public (e.g.,

those that require passwords or IP confirmation for access or, like the Ginsparg archives at LANL, exclude robots).

ACI systems provide opportunities to examine publication and citation patterns in literature on the Web based on the analysis of very large numbers of documents in defined fields of research and scholarship. On its own, it gives entrée to a heretofore inaccessible information resource. CiteSeer, NEC's ACI system, provides access to the full text of source documents on the World Wide Web, and supports retrieval based on keyword or citation links. It can also locate papers related to a given document by using common citation information or word similarity. Given a specific document, CiteSeer can also display the context of how subsequent publications cite that document.

In this paper, we report the results of a dual analysis of a Web-based ACI system and a citation database which indexes the traditionally published literature. We compare two views of information production and use in computer-related research based on citation analysis of PDF- and Postcript-formatted documents on the Web and a parallel citation analysis of the journal literature indexed by ISI in SCISEARCH (the ISI database covering the natural sciences, engineering, and medicine). Our goal is to establish a baseline profile of computer science "literature" as it appears in published journals and as it appears on the publicly indexable Web. From this starting point, we hope to identify additional research areas dealing with information dissemination and citation practices in computer science and the utility of autonomous citation indexing on the Web as an adjunct to commercial indexing services.

There is surprisingly little in the previously published literature with which to compare our findings. Analyses of print sources include Salton and Bergmark's (1979) citation study of computer science literature, Hirst and Talent's (1977) study of computer science journals, Subramanyam's (1976) analysis of core journals in computer science, Culnan's (1978) study of citations in national computer science conference proceedings, and McCain and Whitney's (1994) examination of citation patterns in an emerging area of computer science.

There are several studies from relevant domains not directly related to computer science that deserve mention. The first is an examination of the highly cited items in the area of basic and applied mathematics (Garfield, 1977). Garfield also investigated highly cited items in engineering (Garfield, 1978) and published a two-part analysis of the most highly cited items in the CompuMath Citation Index (Garfield, 1984a,b). Finally, Drott's (1995) examination of the role of conference papers in the scholarly communication of information science provided a framework for the discussion of our results.

2. Methods

2.1. Citation sources

2.1.1. CiteSeer

CiteSeer is a prototype ACI system developed by NEC that uses Web search engines to locate documents on a particular topic. It then downloads Postscript and PDF files and converts these to text. Research papers are automatically identified by the presence of a reference or bibliography section, from which citation information is extracted. Each citation is parsed into fields such as

title, author, year of publication, and page number. The citations are then normalized to detect variant spellings and varied citations to the same article. The design and operation of the CiteSeer program is described in greater detail in Lawrence, Giles, and Bollacker (1999).

CiteSeer's computer science database was created by searching the Web on a list of over 200 keywords and phrases representing research topics in computer science and related fields such as computer engineering, software engineering, and information systems. At the time of this study, the database consisted of 200,314 posted source documents (sources of citations to be counted and analyzed) containing 2,829,529 cited references. The CiteSeer database represents computer science publications that are freely available on the Web.

Our Web-based analysis is based on three datasets taken from the CiteSeer computer science database. We examine the top 500 most highly cited works regardless of whether the works appear in the CiteSeer database. We then examine the 200 most highly cited works which are also part of the CiteSeer database of source documents (that is, they are highly cited by Web authors and are themselves available on the Web). This sample demonstrates the availability of highly cited documents via the publicly available Web. We also examine a random sample of 500 source documents contained in the CiteSeer CS database in order to give a broader profile of the database as a whole.

2.1.2. SCISEARCH

SCISEARCH, the ISI citation index that covers computer science and related areas, began print publication in 1961 and has been published online since 1974. Coverage includes roughly 3500 source journals, over 16 million source documents and over 400 million cited references. The database is available on Dialog in two files, File 434 (1974–1989) and File 34 (1990 to date). The literature of computer science is represented primarily by a number of relevant subdivisions within the broader computer science subject category including: artificial intelligence; cybernetics; hardware and architecture; information systems; interdisciplinary applications; software, graphics and programming; and theory and methods. Due to its size, however, we could not process the entire file to identify the most highly cited documents in computer science and related fields. The RANK command on Dialog creates a frequency-ranked list of values on fields such as DT (document type), JN (source journal name), and CR (cited reference). Unfortunately, there is an upper limit on the number of items it will rank According to Dialog, "A maximum of 50,000 terms can be ranked. This means that if there is only one term in the desired field per record, you can RANK up to 50,000 records. However, in many databases there are multiple terms in a field, therefore, the number of records you are able to RANK is likely to be much less than 50,000." (Dialog Corporation, 1999).

To overcome this, we created a sampling frame that partitioned the entire file sequence into clusters containing 2000 computer-related source documents. From this, we extracted a systematic sample of 15 clusters spanning the period of 1973–1999. Our analysis is thus based on a set of 30,000 source documents and 413,890 cited references.

2.2. Identification of highly cited works

Citations to a given document may have widely varying formats, including variable placement and presentation of publication dates, presence and placement of authors' first names or initials,

and variability in citing publication sources. All fields including author names, title, date of publication, and publication source routinely contain errors and variability. The identification of highly cited works is therefore constrained by the methods used to normalize among variant citations.

2.2.1. Citation normalization in CiteSeer

Although several different algorithms for normalization have been explored, CiteSeer currently normalizes citations based on matching the title and first author of documents. The algorithm depends primarily on extracting the title correctly and uses multiple hypotheses to identify the first author. This allows variations in journal titles and different editions of the same work to be grouped together, enabling researchers to track versions of the same document over time (if the title does not change). Fig. 1 presents a small sample of the 448 citations to Salton and McGill's (1983) book: Introduction to Modern Information Retrieval.

In a small percentage of citations, the title is not extracted correctly, leading to matching errors. On tests covering 1158 citations, about five percent of the automated groupings contained an error in at least one citation. Normalization problems occurring in the set of 500 most highly cited works were detected only during data processing. This reduced the data available for analysis in this study to 488. It is important to note that, while this algorithm works well for computer science literature, it may not work well for certain areas like physics, where citations often do not contain the titles of documents. The CiteSeer project has produced other algorithms that may be used in these cases (Lawrence, Bollacker, & Giles, 1999).

2.2.2. Citation normalization in SCISEARCH

The format for a cited reference in SCISEARCH is an 80 character string that generally gives the last name and initials of the first author of the cited work, the year of publication, volume and/or first page, and a brief cryptic abbreviation of the title. Fig. 2 gives examples of CR strings for a journal article.

There is no standard format beyond these data elements and no attempt to regularize citations at the time of data entry for SCISEARCH (though the journal titles are regularized before the production of the *Journal Citation Reports*, the annual statistical compilation of journal-level citation and publication data). Thus, a particular cited work may have a number of subtly or radically different CR strings, which Howard White has called *allonyms* (White, 2001). Our

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. NY: McGraw-Hill.

G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

Gerard Salton and Michael J. McGill. Introduction to Modern Information Retrieval, chapter 2, pages 24--51. McGraw-Hill Computer Science Series. McGraw-Hill, New York, 1983.

G. Salton and M. McGill. Introduction to Modern Information Retrieval. McGraw Hill Book Co., New York, 1983.

Fig. 1. Sample document normalization grouping from CiteSeer.

Counts	Cited Reference String
6	CR=ZADEH L, 1965, P338, INFORM CONTR
2	CR=ZADEH LA, 1965, FUZZY SETS INFORMATI
2	CR=ZADEH LA, 1965, INF CONTROL
7	CR=ZADEH LA, 1965, INFORMATION CONTROL
7	CR=ZADEH LA, 1965, P338, FUZZY SETS INF CONTR
2	CR=ZADEH LA, 1965, P338, FUZZY SETS INFORM CO
32	CR=ZADEH LA, 1965, P338, INFORM CONTR
2	CR=ZADEH LA, 1965, P338, INFORMATION CONT AUG
3	CR=ZADEH LA, 1965, P338, INFORMATION CONTROL
11	CR=ZADEH LA, 1965, V8, P338, FUZZY SETS INF CONTR
73	CR=ZADEH LA, 1965, V8, P338, FUZZY SETS INFORM CO
70	CR=ZADEH LA, 1965, V8, P338, FUZZY SETS INFORMATI
8	CR=ZADEH LA, 1965, V8, P338, INF CONTROL
2492	CR=ZADEH LA, 1965, V8, P338, INFORM CONTR
18	CR=ZADEH LA, 1965, V8, P338, INFORM CONTROL
793	CR=ZADEH LA, 1965, V8, P338, INFORMATION CONTROL

Fig. 2. Sample CR strings from SCISEARCH.

sample of 413,890 individual cited references included 306,723 unique CR strings. In order to determine the nature and identity of the highly cited documents in the cluster sample, we had to identify all reasonable allonyms for a given cited work and sum the citation counts. This was done manually, by identifying all CR strings with 10 or more citations (over 1400) and grouping each with its probable allonyms by sorting on the cited author subfield. Citations to all pages in the same edition of the same book were combined, as were strings that varied only in journal title abbreviation; editions of books were kept separate, however. We retained for analysis the top 515 cited works.

2.3. Source document categorization

Source documents were categorized by type to identify approximate numbers of books and book chapters, journal articles, conference papers, technical reports, and miscellaneous other types of documents such as dissertations and reviews.

Three sets of source documents were categorized: from CiteSeer, we examined the 200 most highly cited source documents, and a set of 500 random source documents contained in the database; from SCISEARCH we examined the set of 30,000 source documents. We found the random sample from CiteSeer to be limited for categorizing the kinds of documents typically found in the CiteSeer database. Of the 500 random documents, 179 provided reasonably complete information, (e.g. Title, Author, Publication Name, and Date) and could be identified as book chapter, journal article, conference paper, etc. CiteSeer currently only provides publication details for articles that are cited within the database (publication details for other articles may be available in other databases, on the Web pages where the articles were found, or by contacting the authors).

3. Results and discussion

3.1. Profiles of source documents in CiteSeer and SCISEARCH

We first examined publication patterns as expressed in the data corpus. Here, we were primarily interested in identifying where the documents come from which make up the databases of CiteSeer and SCISEARCH. We were also interested in the age of the material included as source documents in these databases. We examined the 200 most highly cited works, which are also source documents in CiteSeer as evidence of the availability of highly cited documents via the publicly available Web. We also examined a random sample of 500 source documents contained in the CiteSeer database in order to give a broader view of the database as a whole.

Table 1 presents a profile of document types, and Table 2 shows publication ranges for the CiteSeer 200 highly cited source documents and for the CiteSeer 500 document random sample. In the latter case, because full publication details were not available for many of the papers, the date listed with the file may not reflect the actual date of publication.

One interesting finding is the large number of conference proceedings and recent publications, which appear among the most highly cited source documents as well as the source documents in general. Setting aside documents whose publication type is unidentifiable, we find that nearly half (45%) of the attributable source documents emerge from conference proceedings. Similarly, setting aside documents whose publication date is unknown, we find that the majority (91%) of all CiteSeer source documents were published in the last ten years. It is also interesting to note that a few of the source documents predate the Web and its widespread use of the Postscript format. This may indicate that some scholars at least are retrospectively converting their older materials for increased dissemination via the Web.

Table 1
CiteSeer source documents by type

	Top 200 highly cited source documents	500 Random source documents
Books/book chapters	17	15
Journal articles	77	42
Conference proceedings	87	90
Technical reports	10	22
Miscellaneous	9	16
Unidentifiable	0	315

Table 2
CiteSeer source documents by year

	Top 200 highly cited source documents	500 Random source documents
1970–1979	3	0
1980–1989	25	2
1990–1999	168	188
Unknown	4	310

SCISEARCH includes a field which identifies the document type of materials they routinely include. For our analysis of these source documents, we have expanded our categories of document types to reflect these fields. Table 3 presents the document types for the 30,000 SCISEARCH source documents. Analysis of the SCISEARCH documents by decade is presented in Table 4.

What is immediately apparent in examining the SCISEARCH source documents is the high proportion of journal articles compared to all other document types, but as the reader can see, other document types taken from journals are also present.

It is important to note that ISI selection policies for SCISEARCH focus almost exclusively on the scholarly journal literature. Conference proceedings that are published as part of an issue of one of the selected journals may also be included. These conference papers are coded not as proceedings, however, but as journal articles by ISI.

While the striking growth in document counts reflects the growth of the journal literature in computer science, it may also reflect expanded coverage of the computer science literature by ISI. It must be noted that these numbers may have been affected by our systematic sample. In creating our sampling frame, we sought a broad spectrum of representation rather than the specific profiling of literature by years or decades.

Table 3
SCISEARCH source documents by type

Total articles	23 520
Total letter	1732
Total note	1462
Total editorial	1424
Total editorial material	508
Total software review	292
Total news item	274
Total correction, addition	201
Total review	140
Total hardware review	132
Total book review	84
Total meeting abstract	76
Total item about an individual	42
Total review, bibliography	42
Total dicussion	34
Total database review	20
Total bibliography	9
Total reprint	8

Table 4
SCISEARCH source documents by decade

# of documents	Decade
2203	1970–1979
7739	1980–1989
20 058	1990–1999

Most notable in comparing the publication patterns of the source documents from these two databases is the disjoint between the inclusion of papers from conference proceedings. SCISEARCH emerges in this instance as representative of the traditionally published scholarly literature; the source material consists almost entirely of works appearing in journals. CiteSeer source material, on the other hand, includes publications representing scholarly communication at additional points in the R&D timeline (Crawford et al., 1996). Fig. 3 shows points of participation in the R&D timeline by CiteSeer and SCISEARCH.

The CiteSeer database, as represented in the random 500 document sample, displays a large number of technical reports and conference proceedings which may not be retrievable by scholars in their literature searches of traditional indexes. Many conference and journal publications in CiteSeer may be available before the respective proceedings or journal issue. This provides access to scholarly communication in its earliest stages, before formal presentation or publication.

3.2. Citation profiles

We next turned our attention to an examination of the cited references contained within source documents. We examined the approximately 500 most highly cited works in the CiteSeer database and SCISEARCH cluster sample.

It is important to note that the 200 most highly cited source documents in CiteSeer do not all appear within the ~500 most highly cited works. There is some overlap occurring, but not all of the ~500 most highly cited works actually appear as source material for the CiteSeer database. This discrepancy occurs because the majority of highly cited documents are not accessible to CiteSeer. Not all works are freely distributed on the World Wide Web, in PDF or Postscript format. In particular, the most highly cited documents tend to be older documents, which are less likely to be available on the Web. The most highly cited document in the top 500 highly cited works was cited 2109 times. The most highly cited source document is cited 542 times and appears at number 34 in the top 500 most highly cited documents. Overall, only 48 (24%) of the highly cited source documents also appear in the list of the 500 most highly cited documents.

Having observed patterns of difference between the types of source documents included in each database, we sought to understand the types of documents that were being cited most highly by scholars in computer-related fields. Of particular interest to us was the question of whether journal articles or conference proceedings would prove to be more often cited. The document type categorization of the most highly cited works in both databases is presented in Table 5.

Given the differences between the source documents contained in the two databases, it was interesting to see such similarity in the types of works being cited. Documents in both databases overwhelmingly cite books and book chapters, followed closely by journal articles. This is

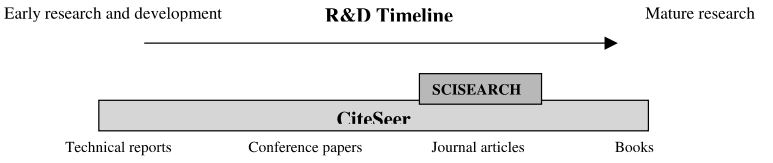

Fig. 3. R&D timeline location of source materials from CiteSeer and SCISEARCH.

Table 5
Most highly cited works by document type

	CiteSeer – 488 most highly cited works (% rounded)	SCISEARCH – 515 most highly cited works (% rounded)
Journal articles	182 (37%)	205 (39%)
Books/book chapters	207 (42%)	290 (56%)
Conference proceedings	77 (15%)	18 (3%)
Technical reports	8	0
Computer documentation	2	1
Miscellaneous	12	1

reminiscent of citation patterns among mathematicians and engineers (Garfield, 1977, 1978). It is also worth noting that both CiteSeer and SCISEARCH contain similar proportions of journal articles amongst the most highly cited articles. The major distinction between the two databases occurs in citations to conference proceedings. There is a higher proportion of conference proceedings amongst the most highly cited documents in CiteSeer (15%) than in SCISEARCH (3%).

Next, we examined the age of the highly cited works. Table 6 presents the publication dates of the most highly cited works in the literature of computer-related fields.

CiteSeer referenced many more recent documents than did SCISEARCH. This may be expected because the source documents in CiteSeer are newer and because of the publication lag in scholarly journal publication. As CiteSeer draws source documents from earlier stages of the R&D timeline, it is not unreasonable to think that these source documents are capable of referencing more recent research.

We noted earlier that both databases demonstrated a high proportion of citations to books, and we wondered if both databases were referencing the same books. We were also curious if other factors might distinguish the citation patterns represented in the two databases. Tables 7 and 8 present author/title citation lists of the top 25 cited works in computer related fields taken from the SCISEARCH and the CiteSeer databases. Citations in bold appeared in the top 25 lists of both.

Most of the material was published between the late 1970s and the mid-1980s. The citation to the most recent publication (1995) appears in the CiteSeer database. The citation to the oldest

Table 6
Age of highly cited works

	CiterSeer (% rounded)	SCISEARCH (% rounded)
1930–1939	0	1 (0.2%)
1940–1949	2 (0.4%)	6 (1%)
1950–1959	2 (0.4%)	13 (2%)
1960–1969	15 (3%)	64 (12%)
1970–1979	63 (13%)	189 (37%)
1980–1989	203 (42%)	215 (42%)
1990–1999	203 (42%)	27 (5%)
Totals	488	515

Table 7
Top 25 cited works in computer-related journals – ISI sample

ISI CITES	
300	**M. R. Garey, D. S. Johnson, *Computers and Intractability. A Guide to the Theory of NP-completeness*, W. H. Freeman and Company, 1979**
254	Knuth, Donald Ervin, *The Art of Computer Programming*. [3 vol] 2d ed. Reading, Mass: Addison-Wesley Pub. Co. 1973
220	**A. Aho, J. Hopcroft, and J. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.**
206	L.A. Zadeh, Fuzzy Sets, *Information and Control*, Vol. 8, pp. 338–353, 1965
Oldest	
164	**Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York, 1973.**
136	**Goldberg, D.E. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, Reading, MA, 1989.**
106	Rumelhart, David E., McClelland, James. PDP Research Group. Parallel Distributed Processing Research Group. *Parallel distributed processing: explorations in the microstructure of cognition*. 2 vol_Cambridge, Ma.; London: MIT Press, 1986
103	**S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.**
100	Adele Goldberg and David Robson. *Smalltalk-80: The Language and its Implementation*. Addison-Wesley, 1983
88	**F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction.* New York: Springer Verlag, 1985.**
87	Kailath, Thomas. *Linear systems.* Englewood Cliffs, N.J.: Prentice-Hall, 1980
86	**Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, California, 1988.**
84	**C. A. R. Hoare. *Communicating Sequential Processes.* Prentice Hall International, 1985**.
82	James Rumbaugh, Michael Blaha, William Premerlani, Frederick Eddy, and William Lorensen. *Object-Oriented Modeling and Design*. Prentice Hall, New Jersey, 1991.
82	L. Kleinrock. *Queueing Systems, Theory*, volume I. John Wiley, 1975.
80	Rosenfeld, Azriel and Avinash Kak, *Digital Picture Processing* 2nd ed. New York: Academic Press, 1982
79	Ronald L. Rivest, Adi Shamir, and Leonard M. Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 21(2):120–126, 1978.
79	Shafer, Glenn. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.
75	Kwakernaak, Huibert and R. Sivan, *Linear Optimal Control Systems.* New York: Wiley Interscience 1972
74	**J. Holland. *Adaptation in Natural and Artificial Systems.* The University of Michigan Press, Ann Arbour, 1975**.
73	**A. V. Aho, R. Sethi, and J. Ullman. *Compilers: Principles, Techniques, and Tools.* Addison-Wesley, Reading, MA, second edition, 1986.**
73	E. W. Dijkstra, *A Discipline of Programming*. Prentice Hall, 1976.
73	W. Diffie, and M. E. Hellman, New Direction in Cryptography, *IEEE Transactions on Information Theory*, Vol. IT-22, NO.6, Nov.1976, pp. 644–654.
72	**R. Milner. *Communication and Concurrency*. Prentice-Hall, 1989**
72	J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Oxford University Press, Oxford, 1965.

Table 8
Top 25 cited works in CiteSeer computer-related database

NEC CITES	
2109	**M. R. Garey, D. S. Johnson, *Computers and Intractability. A Guide to the Theory of NP-completeness*, W. H. Freeman and Company, 1979**
1139	**Goldberg, D.E. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley, Reading, MA, 1989. 4210**
1116	**C. A. R. Hoare. *Communicating Sequential Processes*. Prentice Hall International, 1985.**
1018	G. Golub and F. Van Loan, *Matrix Computations*, 2nd edition, Johns Hopkins University Press 1989.
1011	T. Cormen, C. Leiserson, and R. Rivest. *Introduction to Algorithms*. The MIT Press, Cambridge, MA, 1990
991	**J. Holland. *Adaptation In Natural and Artificial Systems*. The University of Michigan Press, Ann Arbour, 1975.**
980	**A. V. Aho, R. Sethi, and J. Ullman. *Compilers: Principles, Techniques, and Tools*. Addison-Wesley, Reading, MA, second edition, 1986.**
907	**Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, California, 1988.**
863	**Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.**
850	J. W. Lloyd. *Foundations of Logic Programming*. Springer Verlag, 1984.
820	**R. Milner. *Communication and Concurrency*. Prentice-Hall, 1989**
798	D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing*, chapter 8, pages 318–362. MIT Press, 1986
748	Hopcroft, J. E., and Ullman, J. D., *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Reading, Massachusetts, 1979.
741	Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.
737	Knuth, D. E. *Fundamental Algorithms*, vol. 1 of *The Art of Computer Programming*. Addison-Wesley, New York, 1968
725	**F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*. New York: Springer Verlag, 1985.**
700	**S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.**
680	Rodney A. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2:14–23, April 1986.
664	L. Lamport, Time, clocks, and the ordering of events in a distributed system, *Communications of the ACM* (July 1978), pp. 558–565.
661	Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* B, 39:1–38.
649	J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986
636	C. L. Liu, and J. W. Layland, Scheduling algorithms for multiprogramming in a hard-real-time environment, *J. of the ACM*, 1973, pp. 46–61.
628	**A. Aho, J. Hopcroft, and J. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.**
623	Gamma, E., Helm, R., Johnson, R., and Vlissides, J., *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley Pub. Co., 1995.
Most recent	
607	I. Daubechies, *Ten Lectures on Wavelets* (SIAM, Philadelphia, 1992)

publication (1965) occurs in the SCISEARCH database. With one exception, all of the overlapping citations between the two databases are to books and book chapters. We compared these lists to the 100 most-cited books in the CompuMath Citation Index, 1976–1980, compiled by Garfield (1984b). Of the top 25 works highly cited in SCISEARCH, five books (three of the top five items) and the most highly ranked article were also in Garfield's list even though the study was conducted 16 years ago. Of the top 25 works highly cited in CiteSeer, three books were also present in Garfield's list.

4. Discussion and conclusion

While there can be no doubt that the World Wide Web provides a fast and efficient means of disseminating and accessing scientific information, it may be premature to ring the death knell for traditional publishing and databases that provide access to the traditionally published literature. Currently, it seems that scholars in computer-related fields prefer to cite books and journal articles. It is unclear whether or not citations to conference papers will increase over time due to the existence of ACI systems like CiteSeer, which makes these papers easier to locate. It is interesting to note that, while the majority of papers on the Web are from conference proceedings, the most highly cited works overall are from books and book chapters which seldom make their way as full text into electronic databases or on the Web.

With respect to the role of conference proceedings, Drott (1995) found that only 13% of a sample of ASIS Proceedings papers ever made their way into the journal literature. Drott speculated that conference papers function in one of three roles:

1. Self-improvement – the paper is offered for scrutiny to peers for discussion and improvement as it evolves to journal form. Depending on the criticism received, the paper may not move to journal publication.
2. Group contribution – the paper is presented as a means of sharing information within a discourse community with no further aim.
3. Final product – the conference paper is an end in itself and is a different type of final product than the journal paper. Hence the conference paper may contain information that editors prefer not to publish (techniques rather than results), present limited results, or pose untested hypotheses. This view of the communication value of conference proceedings in computer science has also been noted by Kling and McKim (1999).

Discussion with colleagues in computer science suggests a fourth role for conference papers: as a substitute for the journal article. In this role, the conference paper is an alternative which offers the same functions as the journal article. It represents the intended end product of research rather than a stepping stone to future journal publication. Within this context, the provision of access to this material for promotion and tenure decisions assumes increased importance. It is unclear what role is being played by scholars making their conference papers available on the Web. CiteSeer provides both a greater number of conference papers as source materials as well as a greater proportion of citations to conference papers than SCISEARCH. An interesting area of future research would be to examine more closely the citations coming from these conference papers in order to determine if they contain a greater proportion of citations to other conference papers. Similarly, we could seek to understand the impact of Web availability on scholarly

communication by examining the number and types of sites which link to or are linked from a paper posted on the Web. These "hubs and authorities" (Chakrabarti et al., 1999) may function as a form of citation practice within certain communities of discourse who routinely seek out and publish scholarly research on the Web.

In this article, we compared two views of information production and use in computer-related research. We based this on citation analysis of PDF- and Postcript-formatted publications on the publicly available Web using autonomous citation indexing, and a parallel citation analysis of the journal literature indexed by ISI in SCISEARCH. From this starting point, we identified additional research areas dealing with information dissemination and citation practices in computer science and the utility of autonomous citation indexing on the Web as an adjunct to commercial indexing services. A closer integration of these tools will be of great value in tracking scholarly communication and exploring scholarly disciplines. In order to create such a federated system, we must take into consideration the differences in source document structures, citation normalization, and access to gray literature.

There will undoubtedly continue to be limitations to what is accessible on the Web or within structured databases such as SCISEARCH, and individuals interested in tracking all citations to a given work may still need to go beyond a single database. It is conceivable that developments in XML and other document content markers may make for greater access to citations appearing in Web based documents. For the time being, Cameron's (1997) notion of a universal citation database linking every scholarly work ever written remains an elusive dream.

References

Cameron, R. D. (1997). A universal citation database as a catalyst for reform in scholarly communication. *First Monday*, April, http://www.firstmonday.dk/issues/issue2_4/cameron/index.html.

Chakrabarti, S., Dom, B., Kumar, S., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Hypersearching the Web. *Scientific American, June*, 54–60.

Crawford, S., Hurd, J., & Weller, A. (1996). *From print to electronic: The transformation of scientific communication*. Medford, NJ: Learned Information.

Culnan, M. (1978). An analysis of the information usage patterns of academics and practitioners in the computer field: A citation analysis of a national conference proceedings. *Information Processing and Management*, *14*(6), 395–404.

Dialog Corporation (1999). *Dialog lab workbook*. Cary, NC: The Dialog Corporation.

Drott, M. C. (1995). Reexamination of the role of conference papers in scholarly communication. *Journal of the American Society for Information Science*, *46*(4), 299–305.

Garfield, E. (1977). Highly cited works in mathematics. Part 2: Applied mathematics. In *Essays of an information scientist: Vol. 1* (pp. 509–513). *Current Contents, #48*, 5–9 (28 November 1973) (reprint).

Garfield, E. (1978). Characteristics of highly cited publications in engineering sciences. In *Essays of an information scientist: Vol. 2* (pp. 441–446). *Current Contents, #12*, 5–10 (22 March 1976) (reprint).

Garfield, E. (1984a). The multidisciplinary impact of math and computer science as reflected in the 100 most-cited articles in CompuMath Citation Index, 1976–1980. In *Essays of an information scientist: Vol. 7* (pp. 232–239). *Current Contents, #31*, 3–10 (30 July 1984) (reprint).

Garfield, E. (1984b). The 100 most-cited books in the CompuMath Citation Index, 1976–1980. In *Essays of an information scientist: Vol. 7* (pp. 264–269). *Current Contents, #34*, 3–8 (20 August 1984) (reprint).

Garvey, W. D., Lin, N., & Tomita, K. (1972a). Research studies in patterns of scientific communications – II. The role of the national meeting in scientific and technical communication. *Information Storage and Retrieval*, *8*, 159–196.

Garvey, W. D., Lin, N., & Tomita, K. (1972b). Research studies in patterns of scientific communications – III. Information exchange processes associated with the production of journal articles. *Information Storage and Retrieval*, *8*, 207.

Ginsparg, P. (1997). Winners and losers in the global research village. *Serials Librarian*, *30*(3/4), 83–95.

Hirst, G., & Talent, N. (1977). Computer science journals – an iterated citation analysis. *IEEE Transactions on Professional Communication*, *PC-20*(4), 233–238.

Kling, R., & McKim, G. (1999). Scholarly communication and the continuum of electronic publishing. *Journal of the American Society for Information Science*, *50*(10), 890–896.

Lawrence, S., Bollacker, K., & Giles, L. (1999). Autonomous citation matching. In *Proceedings of the third international conference on autonomous agents* (p. 393). New York: ACM Press.

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

Lawrence, S., & Giles, L. (1999). Accessibility of information on the Web. *Nature*, *400*, 107–108.

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

Lindholm-Romantschuk, Y., & Warner, J. (1996). The role of monographs in scholarly communication: An empirical study of philosophy, sociology, and economics. *Journal of Documentation*, *52*(4), 389–404.

McCain, K., & Whitney, J. (1994). Contrasting assessments of interdisciplinarity in emerging specialities: The case of neural networks research. *Knowledge*, *15*(3), 285–306.

Salton, G., & Bergmark, D. (1979). A citation study of computer science literature. *IEEE Transactions on Professional Communication*, *PC-22*(3), 146–158.

Subramanyam, K. (1976). Core journals in computer science. *IEEE Transactions on Professional Communication*, *PC-19*(2), 22–25.

White, H. D. (2001). Authors as citers over time. *Journal of the American Society for Information Science*, *52*(2), 87–108.