This is a preprint of an article published in *Scientometrics, 54* (3): 449-472, 2002

Citation analysis using scientific publications on the Web as data source: A case study in the XML research area [1]

Dangzhi Zhao, Elisabeth Logan

School of Information Studies, Florida State University, Tallahassee, Florida 32306-2100 (USA)

e-mail: dzhao@csit.fsu.edu, logan@mailer.fsu.edu

With the primary goal of exploring whether citation analysis using scientific papers found on the Web as a data source is a worthwhile means of studying scholarly communication in the new digital environment, the present case study examines the scholarly communication patterns in XML research revealed by citation analysis of ResearchIndex data and SCI data. Results suggest that citation analysis using scientific papers found on the Web as a data source has both advantages and disadvantages when compared with citation analysis of SCI data, but is nonetheless a valid method for evaluating scholarly contributions and for studying the intellectual structure in XML research.

Introduction

Although the print journal has served as the primary medium of scholarly communication for more than three centuries, the accelerated development of information technology, especially the rapid growth of the Web, is transforming the scholarly communication system by providing new and powerful media for communication. Since the Internet has greatly improved the efficiency of communication, more and more scholars are exchanging scientific information through the Internet, not only by email but also by publishing papers on the Web. In areas such as Physics and Computer Science, "the Web is often the first choice for finding information on current research, for breaking scientific discoveries, and for keeping up with colleagues (and competitors) at other institutions." (Youngen, 1997) The amount of scientific publication on the Web has grown dramatically over the last few years. The Web has become a new and powerful medium for scientific communication.

Since the format of scholarly communication is changing ---- more in some fields than in others ---- examination of these new formats is important to see the types of communication that are taking place and the similarities to what we have come to expect from print based communication. The goal of the present study is to explore these issues by conducting a citation analysis of scientific publications on the Web which address XML research. Results from the study may contribute to understanding the new formats of scholarly communication, to the advance of citation analysis theory and methodology, and to XML research.

Related studies and research questions

Citation analysis is a well-known technique that has long been used to study scholarly communication. In citation analysis studies, citations in research articles, often published in journals, are analyzed as artifacts of scholarly communication representing the citing authors' use of the previously published work. "The choice of works to be cited in a scholarly paper is assumed to reflect the organization of a scientific community and its knowledge base, as perceived by the citing authors, and the value placed by the community (and the author) on previous contributions" (McCain, 1990a, p.195). As a result**,** citation analysis can help understand scholarly communication patterns and identify major contributions and contributors. "The major advantages of citation analysis are its high reliability and unobtrusiveness" (Harter & Kim, 1996, p. 301). According to Borgman (1990), applications of citation analysis in studies of scholarly communication include examining the characteristics and the evolution of scholarly communities and networks (e.g. Ding, 1999; Small, 1990; White & McCain, 1998), evaluating scholarly contributions (e.g. Meho & Sonnenward, 2000), studying the diffusion of ideas (e.g. Rogers & Cottrill, 1990), and investigating scientific collaborations (e.g. Qin et al, 1997). The ISI databases including SCI (Science Citation Index) and SSCI (Social Science Citation Index), which index only journal articles (ISI, 2000), have served as the data source for most citation analysis studies on scholarly communication.

As the Web is becoming a new and powerful medium for scientific communication, citation analysis and other bibliometric techniques have found some applications in studying this new phenomenon in scholarly communication. Studies of this kind roughly fall into two categories. One is to apply, often with modifications, citation analysis and other bibliometric principles and techniques to study the characteristics and link structures of the Web. Examples include studies

[1] This paper is supported in part by a fellowship of the School of Computational Science and Information Technology, Florida State University

on search engines making use of hyperlink structure (Clever, 1999), and so-called "Webometrics" studies (e.g. Almind & Ingwersen, 1997; Cronin et al, 1998, Dahal, 2000; Egghe, 2000; Larson, 1996a, 1996b; Rouseau, 1997; Turnbull, 2000). The other category of studies looks at "electronic ingredients" in citations in journal articles, which essentially belong to traditional citation analysis. For example, McCain (2000) explored the extent of Web publication of electronic research-related information in the sciences by analyzing the citations of this information in abstracts of journal articles. Harter and Kim (1996) examined the impact of e-journals on traditional print journal-based scholarly communication by looking at the degree to which print journal articles were citing electronic information in their reference lists. Lu (1999) looked at the changes associated with the Internet in activities such as citations of some high impact research oriented print journals, to explore the transition to the virtual world from the traditional paper world in formal scholarly communication.

However, a third category of study which seems to be more straightforward has been somehow overlooked: citation analysis using research papers published on the Web as data source. Actually, the scientific research papers increasingly published on the Web open up the possibility of various citation analysis studies, such as studying the patterns of scholarly communication taking place on the Web and the transition of scholarly communication systems from print to electronic, and re-examining and advancing the theory and methodology of citation analysis itself based on newly available data and tools. A few studies, such as Youngen's examination of the citation patterns of the physics preprint literature available electronically (Youngen, 1997), and Goodrum and McCain's comparisons of the distributions of highly cited computer science documents by document type and publishing date identified from the Web and those from SCI (Goodrum & McCain, 2001), produced some interesting results, indicating a good start for such studies. The present study may add to these contributions.

Since such studies use research papers, not just any Web pages, as data sources, and look at citations made in the reference lists, not in inline links, citations are assumed to have the same meanings as those made by journal articles as mentioned above, and citation analysis therefore can be applied in the same way. In addition to the differences in the communication media (the journal vs. the Web) such as the higher speed of communicating and the wider distribution of information afforded by the Web, the major difference here seems to be the types of articles to be analyzed: journal articles only when using SCI in contrast to a wider variety of types on the Web such as degree theses, technical reports, conference papers, preprints, and journal articles which may represent different stages in the scholarly communication process. These differences in media and document type may be contributing factors to differences in communication patterns between Web-based and journal-based publications.

The present study seeks to explore the following research questions by conducting a citation analysis of scientific papers in the research field of XML.

- What can citation analysis of Web publications on XML research tell us about the visibility of scholars in this area, and about its intellectual structure?
- Are there any significant correlations between the rankings of authors identified from the Web and those identified from ISI's Science Citation Index in the areas of XML research?
- Are there any significant differences between the intellectual structure of XML research field identified from the Web publication citation analysis and that reported by ISI's Science Citation Index?
- Is citation analysis of Web publications valid as a method for evaluating the impact of scholars and in studying the intellectual structure in the case of XML research field?
- What is an appropriate methodology for citation analysis of Web publications?

While various citation analysis techniques may be used to explore these and other interesting issues in the context of the Web, this study is essentially an author co-citation analysis (ACA). Since 1981 when it was introduced by White and Griffith (1981), ACA has been developed into a well-known literature-based technique of studying the intellectual structure of scholarly fields and the characteristics of scholarly communities (Ding, 1999; White, 1990; White & McCain, 1998). This study applies this technique to the research field of XML as represented by a slice of its literature found on the Web. It first determines a set of the most visible scholars in this field by comparing author rankings according to the number of citations they received. It then analyzes the interrelationships of these scholars, or more precisely, their oeuvres based on co-citation data to examine the intellectual structure and characteristics of the scholarly community represented by these scholars. Comparisons are made between the print world and the Web to see whether there are any differences in scholarly communication patterns.

Methodology

Why XML for the case study?

XML has been chosen for this case study for several reasons.

1. There are enough data available for conducting citation analysis studies.

The core of the XML field of study belongs to computer science. As mentioned above, computer science is one of the areas where researchers are publishing heavily on the Web. Therefore the number of research papers published on the Web in this field is likely to be big enough for applying citation analysis, a method known to achieve a macro perspective on the scholarly communication structure based on a large number of publications (Cronin, 1984). In addition, there is a tool available, namely ResearchIndex.com, for browsing and searching citation data in computer science. ResearchIndex, developed by the NEC Corporation Research Institute, is a SCI-like tool available on the Web which automatically indexes research papers found on the Web, and in fact provides more information on cited papers than SCI: titles, all authors, and abstracts or full text papers for those available on the Web.

2. New models for scholarly communication, if any, should be more easily identified.

Due to the commonly existing tendency of resistance to change, it is likely to be more difficult for a well-established field to adopt new technology, such as new formats of communication, than a field that was born in the new technology. "Born digitally" in 1996, XML, a fairly young but fast-growing field of study, has been growing with the exploding Web technology. In such a field, the difference between the Web-based and the print journal based scholarly communication, if any, should be more pronounced, although this might also introduce biases to this study which would limit the generalization of findings. A separate study may explore these issues in detail by comparing this field with a long-established field, say, SGML in terms of the differences in communication patterns identified from the Web and from the journal.

3. Although the core of the XML research field belongs to computer science, XML technology has applications in a wide range of areas, making it a broadly interdisciplinary field of study. Citation analysis is recognized as a good approach to studying the interdisciplinary structure of a research area.

Data collection

ResearchIndex was used to identify papers on XML published on the Web and papers cited by these papers. (The actual search was conducted on November 21, 2000.) The corresponding data have also been collected from ISI's Science Citation Index. (The actual search was done on November 30, 2000.) No years of publication are specified as XML research is a fairly young field of study, and a five-year period**,** which is commonly used in citation analysis studies, can cover almost all publications in this field. For example, searches using Science Citation Index and choosing "all years" resulted in only 3 more papers than choosing years of 1996 through 2000.

"XML" and "extensible markup language" were used in both tools to search papers (citing papers) on XML. The resulting paper entries were retrieved from the databases and downloaded into a local machine. Since the existence of duplicates was found to be one of the major differences between traditional databases and the Web, the papers from ResearchIndex were examined by a program and then manually to remove possible duplicates. Programs were then developed in Java to parse the descriptions of these papers, to store the resulting citation information such as titles, authors, publishing sources and years in a data structure that is convenient for counting citations and co-citations, and to obtain the rankings, matrixes and distributions needed for subsequent data analyses.

As the procedures described above reveal, the present study defines the field of XML research by all papers indexed under terms "XML" or "extensible markup language" by certain tools, namely ResearchIndex or SCI, rather than operationalizing the field in terms of its journals as some citation analysis studies have done. Also unlike some studies that obtained citation counts based on the entire database of SCI or SSCI, this study counts only citations made by those papers meeting the searching criteria. Although they are thus subsets of the authors or papers' total counts, and therefore may not reflect the authors or papers' whole influences, these counts fully represent how the authors or papers were perceived by scholars doing XML research, which is sufficient for the purpose of this study. (White & McCain, 1998)

Data analysis

1. Identifying the most visible scholars.

Authors are ranked by the number of publications and the number of citations they received respectively based on each of the two data sets: the data set from ResearchIndex and that from SCI. As is well-known, SCI indexes only first authors of cited papers, which has raised discussions on whether this is a serious problem with citation analysis using SCI as data source especially when used in evaluating authors and contributions (Garfield, 1979; Lindsey, 1980; Long et al, 1980; MacRoberts & MacRoberts, 1989; McCain, 1988; Smith, 1981; Stokes & Hartley, 1989). ResearchIndex that indexes all authors gives us the opportunity to provide some empirical data on this issue. In order to do so, rankings of authors by

citations received are obtained and compared from both ResearchIndex and SCI by considering only the first authors and then from ResearchIndex by considering the first five authors.

When considering the first five authors of cited papers in counting authors' citations, the five authors are treated equally, meaning that the number of citations received by each of the five authors would increase by one when the paper co-authored by them is cited. Although this may favor authors who often publish as co-authors, there are several reasons justifying this approach. First, there is a well-known citing practice in some fields that authors of a paper are listed alphabetically, resulting in that the first author is not always the one who contributed the most to the article. Second, it is nearly impossible to assess the relative contributions to co-authored papers based solely on the publicly available data such as the sequence of authors on the paper (Lindsey, 1982).

2. Author co-citation analysis.

In order to analyze the intellectual structure of the field, a set of representative authors are identified by the number of citations they received from each of the two citation indexes: 53 authors who received no less than 30 citations in ResearchIndex when considering the first five authors and 47 authors who received no less than 5 citations in SCI. The threshold 5 was chosen for SCI data so that the authors selected have the possibility of having enough co-citations with other authors, assuming that authors who are co-cited with too few other authors are not good representatives of the field. The threshold 30 was chosen for ResearchIndex data to obtain roughly the same number of authors as from SCI data so that the ACA results are more comparable. Since there are no strict rules regarding thresholds in ACA studies (McCain, 1990b), changing the thresholds to see what would happen would result in another interesting study. For example, the threshold 30 is pretty high. It can be anticipated that more authors would be included in the analysis if a lower threshold were used and a more comprehensive map of the research area might be produced as a consequence.

Co-citation matrixes for these authors were first obtained. Then, based on the assumption mentioned above, 2 authors from ResearchIndex and 7 from SCI who were co-cited with only one author in the same author set were deleted. The matrixes were converted to Pearson r correlation matrixes using the FACTOR procedure in SPSS 10 which are in turn used as input of the factor and cluster analysis. Factor analysis was used to explore the underlying structure of the authors' oeuvres reflecting various aspects of the domain of XML research. Factors were extracted by Principal Component Analysis (PCA) with an oblique rotation (SPSS Direct OBLIMIN) because of the theoretical expectation that the resulting factors (specialties) would in reality be correlated. The number of factors extracted was determined based on Kaiser's rule of eigenvalue greater than 1 (Hair et al, 1998). The internal structures of these matrixes were also explored using cluster analysis (SPSS CLUSTER: complete linkage) and multidimensional scaling (SPSS ALSCAL).

There is one thing worth mentioning here. In SCI-based citation analysis studies, two authors are considered to be co-cited when at least one document in each author's oeuvre (defined as all works with the author as the first author) occurs in the same reference list. In the present study, since not only the first authors are considered, an author's oeuvre is defined as all works with the author as one of the authors, and two authors are also considered to be co-cited when the paper co-authored by them is cited. This seems to us a valid way of counting author co-citations when not only the first authors are available because, just like co-citations, co-authorship indicates authors being related to each other in some sense, and is actually a closer relationship between authors than that formed by co-citations. This may also make it easier to identify the interrelationships among authors because it results in higher co-citation rates. Although this approach is not applicable to SCI data where there are no second authors, we use this approach anyway for ResearchIndex data because it seems an authentic measure of Web-based scholar communication.

Results and discussion

General results

A search on "XML" or "eXtensible Markup Language" resulted in, after removing duplicates, 686 papers using ResearchIndex.com and 165 papers using SCI. The papers from ResearchIndex contain 12020 citations, and those from SCI contain 2511 citations. Among cited papers from ResearchIndex, 24.4% are proceedings, 18.5% are Web publications (8.2% from the World Wide Web Consortium (W3C)), 4.5% are technical reports and 1.5% are degree theses. Among cited papers from SCI, 15.7% are proceedings, 1.5% are degree theses and 1.4% are W3C documents. An examination of these papers revealed some interesting things:

First, only approximately 10% of the papers indexed by SCI are also indexed by ResearchIndex. This low percentage of papers shared by the two data sources seems to make sense since SCI indexes papers only in "the most important" journals while ResearchIndex collects papers published on the Web as completely as possible. It also suggests the significance of the present study because of the lack of studies on the scholarly communication patterns reflected from the huge amount of literature not indexed by SCI.

Second, papers indexed by ResearchIndex cited many more proceedings but only a few more degree theses than those indexed by SCI. For Web publications, although from the limited information about cited papers given by SCI it is difficult to distinguish Web publications from other papers, the difference between the percentages of W3C documents implies that papers indexed by ResearchIndex cited many more Web publications than those by SCI. These seem to suggest that papers published on the Web are more sensitive to other Web publications and the timeliness of information.

Visibility of scholars

This section explores what citation analysis of papers found on the Web can tell us about the visibility of scholars, whether there are any significant correlations between the rankings of authors identified from the Web and those identified from SCI, and whether it is valid to use the Web as a data source for citation analysis in evaluating contributions.

Scholars in the XML research area are ranked by number of papers authored and number of citations received respectively based on the two data sources**,** ResearchIndex and SCI. Table 1 and table 2 list top ranked scholars.

Table 1: Authors ranked by number of publications

SCI						ResearchIndex					
Rank	**Authors**	**# pubs**	**Rank**	**Authors**	**# Pubs**	**Rank**	**Authors**	**# Pubs**	**Rank**	**Authors**	**# Pubs**
1	H. S. Rzepa*	5	8	M. Gaedke	2	1	Derick Wood*	12	11	A. Deutsch*	6
1	**D. Suciu***	5	8	J. Hunter*	2	2	A. Brüggemann-Klein*	11	16	R. Studer	5
3	P. Murray-rust*	4	8	H. W. Gellersen	2	3	Y. Papakonstantinou*	10	16	R. Harper	5
4	S. Mcgrath	3	8	H. Liefke	2	4	**D. Florescu***	9	16	P. Buneman*	5
4	S. Ceri*	3	8	G. V. Gkoutos	2	5	N. Ide*	8	16	**M. Fernandez***	5
4	**M. Fernandez***	3	8	F. Vitali	2	6	W. Fan	7	16	Gustaf Neumann	5
4	J. Bosak*	3	8	E. F. Begley	2	6	S. Cluet*	7	16	D. McKelvie	5
8	S. Paraboschi	2	8	E. Damiani	2	6	**D. Suciu***	7	16	A. Isard	5
8	R. Khare*	2	8	**D. Florescu***	2	6	D. Fensel*	7	23	Y. Labrou	4
8	P. Fraternali	2	8	C. P. Sturrock	2	6	A. Levy*	7	23	S. Weinstein	4
8	P. Ciancarini	2	8	C. Baru	2	11	S. Decker*	6	23	S. Abiteboul*	4
8	N. Sundaresan	2	8	A. Kristensen	2	11	M. Erdmann	6	23	R. van Zwol	4
8	M. Wright	2	8	A. Gupta*	2	11	J. Widom*	6	23	Pavel Velikhov	4
8	M. Rezayat*	2	8	A. Dwelly	2	11	Holger Meuss	6	23	J. Shanmugasundaram	4

Table 1 shows that there are just three common entries among the top 28 authors ranked by number of publications based on the two data sources, and these three have very different ranks in the two rankings. The small overlap of active citing authors between the two data sources indicates that two very different groups of scholars are actively publishing on the Web and in the journals. It is interesting to note that there are 36% more citing authors from the RI column (15) than from the SCI column (11) in table 1 (indicated by *) who belong to the highly cited authors. These include the top 60 cited authors from ResearchIndex and the top 59 cited authors from SCI. Assuming that highly cited authors tend to produce high quality papers, this seems to be a challenge to the commonly held belief that the quality of papers published in journals is higher than that of papers on the Web because of the peer review process. This may also suggest that the group of scholars who are actively publishing on the Web have more influence on XML research than the other group of scholars who are publishing in the journals. It would be interesting to compare the characteristics of the two groups by, for example, examining author profiles.

The difference between the two groups of citing authors leads us to expect some differences between their references. Table 2 presents three lists of authors ranked by number of citations received: one is based on SCI data which contain only first authors (list1), the other two are based on ResearchIndex data considering first authors only (list2) in contrast with considering first five authors (list3).

As it can be seen from table 2, all three lists seem to be very different both in terms of the compositions of the top 34 authors ---- 12 common entries between list1 and list2, 14 between list1 and list3, and 19 between list2 and list3 ---- and in terms of the rankings of the common authors. As far as correlations between rankings are concerned, we already can see some interesting patterns just by visual inspection. There is complete chaos between list1 and list3 because there are at least two influential factors involved: the media and the methods of citation counting. When the latter factor is filtered out, the resulted lists (list1 and list2) are much more compatible especially for the top ranked 10 authors when considering that the overlap of the top 10 authors is 60% while the overlap of citing authors in the two data sets is as low as 10%. The

investigation of the remaining differences should shed light on the differences between the two different media in terms of their effect on scholarly communication patterns, which we may address in a later paper. The shifting pattern between list2 and list3 seems to suggest that the alphabetical position of author name may impact author rankings by citations. To test

Table 2: Authors ranked by number of citations

List2: First Author (ResearchIndex)			List3: First Five Authors (ResearchIndex)			List1: First Authors (SCI)			List2: First Authors (ResearchIndex)		
Rank	Name	# C	Rank	Name	# C	Rank	Name	# C	Rank	Name	# C
1	S. Abiteboul	232	1	S. Abiteboul	331	1	T. Bray	47	1	Serge Abiteboul	232
2	P. Buneman	134	2	D. Suciu	307	2	S. Abiteboul	26	2	P. Buneman	134
3	T. Bray	128	3	J. Widom	257	3	J. Clark	24	3	T. Bray	128
4	A. Deutsch	93	4	A. Levy	213	4	P. Murrayrust	20	4	A. Deutsch	93
5	Y. Papakon-stantinou	87	5	D. Florescu	197	5.5	P. Buneman	18	5	Y. Papakon-stantinou	87
6	M. Fernandez	83	6	M. Fernandez	191	5.5	H. S. Rzepa	18	6	M. Fernandez	83
7	R. Goldman	60	7	J. McHugh	158	7.5	J. Bosak	15	7	R. Goldman	60
8	J. Clark	55	8	H. Garcia-Molina	154	7.5	M. Fernandez	15	8	J. Clark	55
9.5	D. Florescu	54	9	Y. Papakon-stantinou	153	9	A. Deutsch	13	9.5	D. Florescu	54
9.5	A. Bruggemann-Klein	54	10	P. Buneman	152	10.5	T. Bernerslee	12	9.5	A. Bruggemann-Klein	54
11	J. McHugh	53	11	T. Bray	141	10.5	Y. Papakon-stantinou	12	11	J. McHugh	53
13	V. Christophides	45	12	Sophie Cluet	139	12	S. J. Derose	11	13	Vassilis Christophides	45
13	S. Cluet	45	13	D. Quass	125	13.5	P. Atzeni	10	13	S. Cluet	45
13	A. Levy	45	14	A. Deutsch	105	13.5	E. Maler	10	13	A. Levy	45
15	J. Robie	36	15	J. Paoli	99	16.5	S. Cluet	8	15	J. Robie	36
16	O. Lassila	32	16	Tova Milo	97	16.5	M. Rezayat	8	16	O. Lassila	32
17.5	S. Chawathe	30	17	R. Goldman	95	16.5	J. Robie	8	17.5	S. Chawathe	30
17.5	N. Ide	30	18.5	S. Davidson	73	16.5	C. F. Goldfarb	8	17.5	N. Ide	30
19.5	J. Goguen	28	18.5	C. M. Sperberg-McQueen	73	21.5	S. Ceri	7	19.5	J. Goguen	28
19.5	F. Neven	28	20	J. Weiner	71	21.5	R. Khare	7	19.5	F. Neven	28
21	E. Maler	27	21	V. Christophides	60	21.5	K. Takanashi	7	21	E. Maler	27
23.5	R. Harper	26	22	A. Rajaraman	59	21.5	D. Florescu	7	23.5	R. Harper	26
23.5	J. Hammer	26	23	J. Clark	58	21.5	C. M. Sperberg-mcqueen	7	23.5	J Hammer	26
23.5	D. Suciu	26	24	J. Ullman	56	21.5	A. Ohori	7	23.5	D. Suciu	26
23.5	A. O. Mendelzon	26	25	A. Bruggemann-Klein	54	29.5	R. N. Shiffman	6	23.5	A. O. Mendelzon	26
26	C. Beeri	23	26.5	D. Wood	52	29.5	R. H. Dolin	6	26	C. Beeri	23
28	R. Milner	21	26.5	C. Knoblock	52	29.5	R. Goldman	6	28	R. Milner	21
28	G. Wiederhold	21	28	J. Hammer	50	29.5	L. Wood	6	28	G. Wiederhold	21
28	D. Calvanese	21	30	J. Robie	47	29.5	K. Gronbaek	6	28	D. Calvanese	21
30.5	N. Ashish	20	30	G. Hillebrand	47	29.5	H. Lie	6	30.5	N. Ashish	20
30.5	L. Cardelli	20	30	E. Maler	47	29.5	D. Raggett	6	30.5	L. Cardelli	20
33	J. Shanmuga-sundaram	19	32.5	V. Vianu	46	29.5	D. Connolly	6	33	J. Shanmuga-sundaram	19
33	H. Garcia-Molina	19	32.5	A. Mendelzon	46	29.5	B. Bos	6	33	H. Garcia-Molina	19
33	G. Salton	19	34	Yehoshua Sagiv	43	29.5	A. Hunter	6	33	G. Salton	19

Striking pattern of alphabetical rearrangement of rankings

Very little correlation

Good correlation of top ranks overlaid by Web-specific trends

this observation, the distribution of the number of the first cited authors by the first letters of their last names was matched against that of the first five cited authors. The assumption here is that if author rankings are not affected by the alphabetical position of author names, the distributions should match. Of course the ideal way of doing this is to use all authors instead of only the first five authors to obtain the natural (true) distribution of author names in this field. However, since the present study only records the first five authors and the percentage of papers with more than four authors does not seem to be high (5.4% of source papers in ResearchIndex and 9.1% of those in SCI), the present approach is chosen with the hope of obtaining a good approximation. Fig. 1 shows that the two distributions are different in that the percentage of authors with last names beginning with letters appearing early in the alphabet are much higher in the case of first authors. This suggests that the first author of a paper in XML research field does not necessarily contribute most to the paper, and that he/she may have been placed as the first alphabetically. Therefore it is unfair and inaccurate to count only the first authors when evaluating contributors using citation analysis in XML research field.

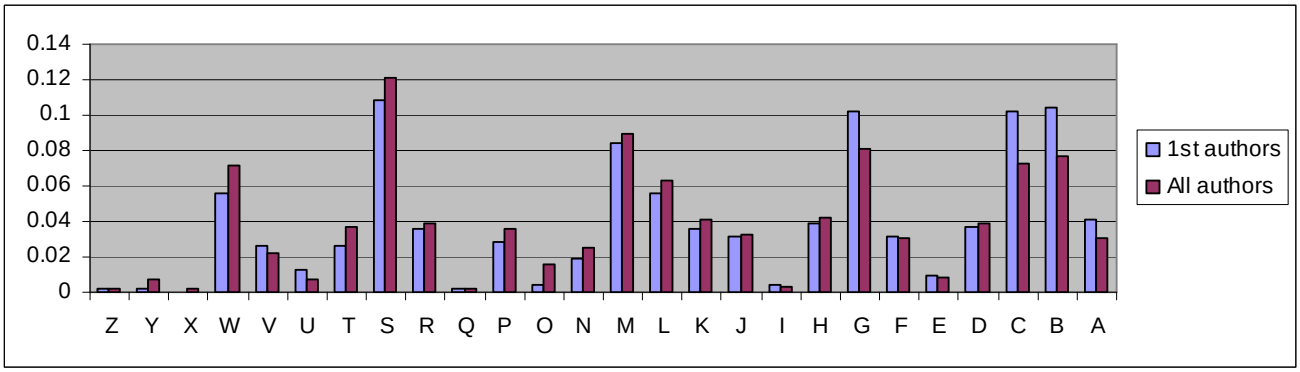

Fig. 1: The distributions of 1st letters of author last names
(The vertical axis is the fraction of the corresponding numbers in the total numbers)

These observations seem to suggest that the Web might be an alternative to SCI as the data source for citation analysis in evaluating contributors and contributions in XML research area. However, more studies are needed to test whether this can be generalized to other fields. The above observations also confirm that "straight counts" (counting only the first authors) and "complete counts" (counting all co-authors) of citations produce very different results regarding the visibility of scholars. Since it has long been suggested that "complete counts" be used (Lindsey, 1980; Long et al, 1980) to acknowledge all authors because most citation analysis studies used "straight counts" "mainly for reasons of economy" (Stokes & Hartley, 1989, p. 106), it seems that the Web as an alternative data source for citation analysis is not only valid but also has the advantage of using "complete counts" without additional expense as the data about all authors are electronically available.

Although table 1 and 2 suggest differences and perhaps have even more importance for scholars with sufficient knowledge of XML research, the discussions above have been limited to the overall patterns due to time and other constraints. As noted above, it would be interesting to examine the remaining differences in detail and to explore the causes, especially when combining the data here with data obtained by other methods such as content analysis and sociometrics.

Intellectual structure

This section explores what ACA (author co-citation analysis) based on data from the Web can tell us about the intellectual structure of XML research area, whether there are any significant differences between the structure identified from the Web publication citation analysis and that reported by SCI, and whether ACA based on data from the Web is valid for studying the intellectual structures of disciplines.

Based on the considerations mentioned above, we chose to use "complete counts" to select authors for the ACA of Web publications. An inspection of the co-citation matrix tells us something about the integration level of the field (McCain, 1990b). The percentage of cells with value of zero is as low as 18.4% and decreased to 12.2% when the two extreme cases were deleted. Thirty out of 51 authors were co-cited or had co-authored with more than 90% of the authors and ten of them were even co-cited or had co-authored with everybody. This suggests to us that the XML research field is well integrated. However, we do not regard the resulting 51 authors as "wholly definitive" of the XML research field although citedness above some threshold seems a good criterion for selecting authors in ACA (White & McCain, 1998).

Table 3: Factor Analysis of 51 authors in XML research area (ResearchIndex)								
Name	Mgt .of XML data (Semi-structured databases)	Middleware: mediators, distributed cooperative info systems	XML standards and specifications	Logical foundations	Logic of formal representations	Knowledge representation, semantics for the Web	Tree / graph transformers (Algorithmic foundations) or SGML	Type theory and Standard ML
J. Weiner	0.989							
J. McHugh	0.954							
D. Quass	0.934							
A. Deutsch	0.909							
G. Hillebrand	0.894							
S. Davidson	0.891							
J. Robie	0.850							
M. Fernandez	0.849							
J. Kang	0.849							
S. Cluet	0.840							
V. Christophides	0.838							
P. Buneman	0.835							
R. Goldman	0.835							
S. Abiteboul	0.807							
J. Widom	0.800							
M. Scholl	0.783							
D. Florescu	0.780							
D. Suciu	0.763							
T. Milo	0.721							
A. Levy	0.691							
V. Vianu	0.688							
A. Mendelzon	0.636							
F. Neven	0.581						0.385	
Y. Sagiv	0.489	-0.464						
J. Hammer	0.318	-0.378		-0.305				-0.304
A. Gupta		-0.757						
L. Raschid		-0.688						
J. Ullman		-0.651						
J. Ordille		-0.621		-0.507				
A. Rajaraman		-0.617		-0.394				
Y. Papakonstantinou	0.416	-0.574						
H. Garcia-Molina	0.436	-0.518						
S. Chawathe	0.437	-0.493						
A. Sheth		-0.398						
J. Paoli			0.947					
T. Bray			0.918					
C. M. Sperberg-McQueen			0.912					
E. Maler			0.839					
S. DeRose			0.789					
J. Clark			0.699					
D. Weld				-0.892				
N. Ashish				-0.881				
C. Knoblock				-0.840				
M. Lenzerini					-0.806			
N. Ide					0.418			
S. Decker						0.906		
D. Fensel						0.890		
O. Lassila			0.439			0.692		
A. Bruggemann-Klein							0.847	
D. Wood							0.820	
R. Harper								0.613

Although we conducted Factor Analysis (FA), Cluster Analysis and Multidimentional Scaling, we only report the results of the factor analysis here, both because of space constraint and because Factor Analysis when applied in ACA has shown to provide clear and revealing results as to the nature of the discipline (White & McCain, 1998). However, we will integrate results from other analyses wherever appropriate.

If factors are interpreted as specialties, the results of the factor analysis presented in Table 3 reveal a structure of specialties within the XML research field and the associated authors' memberships in one or more specialties. Kaiser's rule of eigenvalue greater than 1 resulted in an eight factor model which accounts for 76.53% of the total variance, and the differences between observed and implied correlations are for the most part (90%) smaller than 0.05. The factor names shown in the column headings were given based on the examination of the cited articles written by authors in the corresponding factors. Following White and McCain's example, authors are ranked in the factor on which they load most highly and their loadings on other factors that are above 0.3, if any, are also presented, indicating their contributions to more than one specialty (White & McCain, 1998).

The biggest specialty is obviously *management of XML data or semi-structured databases*. Almost half of the authors belong to this area. The other two big specialties are *middleware* including mediators and distributed cooperative information systems, and *XML standards and specification documents*. All the remaining factors are very small but capture some interesting aspects of XML research. For example, the factor consisting of Decker, Fensel and Lassila represent the studies on knowledge representation and semantics for the Web, which is one of the most promising research areas in XML and is attracting more and more attention. While the factor represented by Weld, Ashish and Knoblock capture the logical foundations of XML research, the factor represented by Bruggemann-
Klein and Wood capture the algorithmic foundations of XML research including SGML and tree/graph transformers. The other two factors pick up some interesting isolates (Lenzerini, Ide, and Harper).

The oblique rotation procedure used here allows us to examine the relationships between the specialties by providing correlations between factors. The correlations between the eight factors are low in general, indicating that the factors do represent different specialties within the field. Factor 2 (*middleware*) has relatively high correlations with factor 1 (*management of XML data*, -0.395) and with factor 4 (*logical foundations*, 0.413), which also can be seen from the relatively heavy overlaps of memberships between these factors in Table 3. This suggests that the study of middleware is closely related to the study of management of XML data or semi-structured databases and draws a lot on the logical foundations of XML research. Actually, some authors such as Sagiv and Chawathe are almost equally recognized as researchers in both of the first two specialties. It is interesting to see how the factor analysis "bestows the primary identification" of these authors (White & McCain, 1998). The results from cluster analysis confirm and add to these observations. While the number of clusters decreases, the two clusters corresponding to factor 2 and factor 4 merge first, then the factor represented by Lenzerini joins them, and then the algorithmic foundations group merges into the management of XML data group. Harper and the Web semantics group keep being separated from the three big groups, indicating that they are very different. Indeed, the study of knowledge representation and the semantics for the Web is obviously different from the rest of XML research which essentially deals with the syntax of XML. Harper's work that was frequently cited focuses on type theory and Standard ML (Meta-Language) which is a programming language developed independently from XML. Harper's example exemplifies the interdisciplinary nature of XML research.

The secondary loadings of Lassila and Neven agree with White and McCain's observation that factor analysis technique is both accurate and sensitive to nuance (White & McCain, 1998). As the author of the RDF (Resource Description Framework) specification document, Lassila's contribution to XML related standards is well recognized although he is primarily perceived as a core figure in the study of the semantic Web. An analysis of Neven's work gives us the same impression.

Almost all authors have fixed citation images because they load high either on a single factor or on two of the three closely related specialties: *management of XML data*, *middleware*, and *logical foundations*. Hammer is the only exception who loads on more than two factors. Authors of this kind who have lower loadings on several factors either write on more topics than those with high loading on a single factor or are just perceived by citers as being related to a greater variety of other oeuvres irrespective of content. Hammer seems to be the former case.

In summary, it seems that scholars in the field of XML research are addressing five major areas of studies: (1) management of XML data or semi-structured databases, (2) middleware including mediators and distributed cooperative information systems, (3) XML standards and specification documents, (4) knowledge representation or semantics for the Web, and (5)algorithmic or logical foundations such as tree / graph transformers. The first two areas of study are heavily overlapped.

It is interesting to observe that these wholly automatic outcomes correspond closely to areas of research identified in W3C draft recommendations and standard documents.[2] This seems to suggest that ACA of scientific publications on the Web is a valid approach to the study of intellectual structure of disciplines, at least in the case of XML research.

ACA was also conducted for SCI data. Factor analysis and cluster analysis of these data give us the impression that the results yielded by SCI data are more complicated compared with those produced from ResearchIndex data. For example, Kaiser's rule of eigenvalues greater than 1 resulted in 3 more factors than ResearchIndex. Although these factors jointly accounted for a fairly high percentage of total variance (73.7%) and all of the individual variables were well explained (having communalities ranging from 0.517 to 0.875), there are more (21% in contrast to 10% in the case of ResearchIndex data) non-redundant residuals with absolute values greater than 0.05, that is, for a larger number of the authors the differences between observed and implied correlations were significantly large.

The results of the factor analysis are presented in table 4 in the same way as table 3 except that the factors are not labeled. Since SCI data do not contain information about the content of the cited papers, we were unable to identify factor topics. Instead, we limit our examinations to the overall structure and the positioning of the common authors between the SCI data set and the ResearchIndex data set analyzed. However, detailed examination of the results presented in table 4 including labeling the factors may yield some very interesting findings.

Table 4: Factor Analysis of 40 authors in XML research area (SCI)

Y. Papakonstantino	0.926										
S. Cluet	0.887										
M. F. Fernandez	0.875										
J. Mchugh	0.869										
P. Buneman	0.826										
D. Florescu	0.781							0.409			
S. Abiteboul	0.746										
J. Robie	0.708									0.339	
A. Deutsch	0.683									0.375	
R. Goldman	0.538							-0.455			
E. Maler		0.785									
V. Apparao		0.740	0.385								
J. Suzuki		0.730									
J. Clark		0.655									
B. Bos		0.618		0.445							
C. F. Goldfarb		0.548				0.329	0.368				
L. Wood		0.501	0.306								
T. Bray	0.325	0.457		0.356	-0.354						
O. Lassila			0.834								
R. Khare			0.752								
T. Bernerslee			0.636		-0.316						
D. Connolly			0.588						-0.528		
J. Bosak			0.533			0.345					
M. Bieber				0.883							
C. M. Sperbergmcqueen				0.854							
H. Lie				0.756							
K. Gronbaek					-0.738						
P. Hoschka					-0.723						
D. Raggett				0.427	-0.657						
R. H. Dolin						0.812					
C. E. Kahn						0.726					
P. Murrayrust							0.779				
H. S. Rzepa							0.765				
A. Ohori								-0.571			
D. Calvanese	0.477							-0.551			
P. Atzeni	0.325							0.524			
L. Alschuler									-0.834		
S. J. Derose				0.470						0.711	
S. Ceri	0.411							0.435		0.479	
A. Hunter											0.900

[2] Thanks to the reviewer of our extended abstract for the ISSI 2001 conference who pointed out the correspondence**.**

Table 4 is obviously more complicated than table 3 not only in the sense that there are three more factors but also because of the loading structures. Three groups in table 3 seem to be recognized here: *management of XML data*, *XML standards and specifications*, and *knowledge representation and the semantics for the Web*, corresponding to the first 3 columns in table 4. Another big group in table 3, namely the study of *middleware*, does not occur separately here. Instead, Papakonstantino's position seems to suggest that this group is merged into the *management of XML data* group. Unlike in table 3, Lassila does not load any on the factor corresponding to *XML standards and specifications*. However, the *Web semantics* group to which Lassila belongs is significantly larger than that in table 3. All the remaining 9 groups have no more than 3 authors. They are likely to have captured some interesting aspects of XML research, especially those interdisciplinary aspects just as Harper in table 3. An examination of the corresponding oeuvres can tell if this educated guessing is correct. However, it would not be unexpected if it were confirmed that more such interdisciplinary aspects are revealed here than in table 3 because SCI data in principle include articles on XML in all related sciences while ResearchIndex data are more restricted to the core of XML research, namely the computer science area.

There seem to be two intertwining factors that contribute to the observed higher complexity of the structure revealed by SCI data compared with that produced from ResearchIndex data, the multidisciplinary nature of SCI data and the limited coverage of SCI in terms of the number of publications in each research field. The wide coverage of disciplines allows most, if not all, aspects of the intellectual structure of the research field under scrutiny to be revealed to some extent whereas the narrow coverage of publications in each field limits the revealed structure to a vague and unclear picture. As suggested by the structure revealed by ResearchIndex data, a clearer picture requires more authors be included in the analysis as well as higher author co-citation rates which in turn require larger number of citing publications and more than first authors involved in co-citation counting. Although ISI may change its "first authors only" indexing policy in the future as demand from users increases, the limited number of publications in each research field has been ISI's approach to subject coverage and reflects a basic conviction that the essence of a field's work is included in selected journals and is therefore not likely to be changed.

It also might be that the differences between the results in table 3 and table 4 to some extent reflect the differences between the two groups of scholars doing XML research, for example, the group mainly publishing on the Web is more closely knit while the group mainly publishing in the journal more diffuse. We may test this in a later paper addressing the characteristics of the different groups in XML research by limiting citing papers from SCI to computer science area and considering only first authors when counting co-citations from ResearchIndex data.

Methodological considerations of Web publication citation analysis

As discussed above, citation data obtained from ResearchIndex have some advantages compared with ISI data. Some of them are highlighted below.

1. They contain many more citing papers, 686 vs. 165 in the case of XML research, which allows citation analysis studies to use larger sample size. This is important to a method such as citation analysis which "is not meant for small population statistics" (Garfield, 1998, p. 1) and whose validity largely lies in the use of voluminous datasets in building macro-level views of phenomena studied. (Borgman, 1990; White, 1990)

2. They contain a wider variety of document types such as conference papers, technical reports and degree theses, which may facilitate various comparison studies.

3. They contain more information about cited papers such as titles, all authors and full source names, which may overcome some problems with SCI data such as being limited to "straight counts" and may facilitate more sophisticated and a larger variety of citation analysis studies such as context analysis and studies applying more sophisticated algorithms.

4. Data collection and analysis of citation analysis studies using ResearchIndex data may be automated more extensively, which might lead to wider use and influence of citation analysis. Back in 1990, one of the major practitioners of citation analysis has seen the "labor-intensive and time-consuming" aspect of citation analysis, and expressed the wish for a citation analysis tool that integrates the separate steps involved in citation analysis into "one smooth-flowing, economical machine process" as well as his anticipation of wider applications of citation analysis in information retrieval and mapping scholarly fields resulting from such a tool (White, 1990, p. 104). Data available on the Web not only makes it possible to develop such citation analysis tools but also to integrate such tools into other Web services and vice versa. Citation analysis would then become an integrated part of other Web services such as digital libraries and search engines to help the user determine the relevance and quality of scientific papers encountered.

However, some of the problems of SCI data, such as different names for the same authors or different authors with the same names, and citation errors produced by citing authors for various reasons (MacRoberts & MacRoberts, 1989;

Smith, 1981), remain with Web data, although the easier access to the original papers and authors' profile (e.g. homepages) helps to correct the data to some extent.

Web data also have some disadvantages.

1. A large portion of papers published on the Web do not have explicit information about date of publishing. Therefore, while citation data obtained from the Web may well facilitate studies of scholarly contributions in a general sense and of overall structures of disciplines, it is difficult to carry out studies based on these data on the evolution of scholarly communities or diffusion of ideas over time.

2. Unlike SCI, citation indexing tools on the Web like ResearchIndex are fully automatic. It is difficult for them to recognize the various referencing formats that are likely to occur in a divergent environment whereas it may be very easy for human beings to make these distinctions. Therefore, these tools tend to produce errors by mixing up information about authors, titles and sources when uncommon or non-standard referencing formats are encountered.

3. While the picture of the structure of specialties within XML research field revealed by ResearchIndex data is clearer, that yielded by SCI data may capture more interdisciplinary aspects of XML research. This calls for a SCI-like citation indexing tool on the Web which covers all disciplines. ResearchIndex is an important contribution to this because, although it is currently limited to broadly defined computer science, it provides citation data of sufficient accuracy and its technology can be adapted to other fields (Lawrence et al, 1999).

4. ResearchIndex provides information about cited papers in an HTML format and information about citing papers in both HTML and BibTeX format. The accuracy of parsing data in HTML format depends heavily on features (e.g. HTML tags) provided for distinguishing different data segments such as authors, title and source. In the case of ResearchIndex, the only such feature for cited papers is that titles are italic. This feature together with the fixed sequence of presenting data, that is authors go first, then title followed by source information, makes it possible to distinguish the basic data segments needed for citation analysis, that is authors, title and source. But it is very difficult to go any further. However, this problem would disappear if ResearchIndex could give the option of saving citation data in a standardized XML format, say. Future citation indexing tools on the Web should provide data in XML format to facilitate data sharing, which is especially important in the Web environment.

Since both SCI data and data available on the Web, specifically ResearchIndex data, have advantages and disadvantages, the combination of these two data sources seems appropriate for citation analysis studies to obtain a larger and richer characterization of scholarly communication structure and process.

Conclusion

With the primary goal of exploring whether citation analysis using scientific papers published on the Web as a data source is a worthwhile means of studying scholarly communication in the new digital environment, the present case study examined the scholarly communication patterns in XML research as revealed by citation analysis of ResearchIndex data and SCI data. Results suggest that citation analysis using scientific papers published on the Web as a data source has both advantages and disadvantages when compared with citation analysis of SCI data, but is nonetheless a valid method for evaluating scholarly contributions and for studying the intellectual structure in XML research. Citation analysis studies should therefore combine SCI data with data available on the Web, whenever possible, in order to obtain a larger and richer characterization of scholarly communication structure and process.

Further studies are needed to examine the differences between the scholars who are actively publishing on the Web and those who are publishing in journals, and to explore the differences between the three groups of scholars who are either top-ranked by both citers on the Web and in journals or by one group of citers but not by the other. Studies are also needed to explore whether results from this study can be generalized to other fields, especially those that were not "digitally born".

The Internet is transforming the scholarly communication system by providing powerful communication media. The increasingly available data and tools on the Web are valuable for both scholarly communication and the study of scholarly communication. They not only facilitate traditional bibliometric studies such as citation analysis, but also open up the possibility of reexamining old methods and developing new approaches. For instance, more sophisticated algorithms than simply counting citations, such as that used by the Clever search engine in ranking hit documents (Clever, 1999), may be applied to evaluate scholars' influences and contributions. Advanced visualization techniques may help to study the social networks identified by citation analysis. We look forward to a booming period of scientometric studies made possible now by the role of the Web in scholarly communication.

Acknowledgments

The authors wish to thank Andreas Strotmann of the Department of Computer Science, Florida State University, for his many helpful insights.

References

1. Almind, T. C., Ingwersen, P. (1997), Informetric analyses on the World Wide Web: methodological approaches to "Webometrics", Journal of Documentation, 53(4), 404-426
2. Borgman, C. L. (1990), Editor's introduction, In: Borgman, C. L. (ed.), Scholarly communication and bibliometrics, Newbury Park, CA: Sage Publications Inc., p. 10-27
3. Clever Project. (1999), Hypersearching the Web, Found at http://www.sciam.com/1999/0699issue/0699raghavan.html, 2000
4. Cronin, B. (1984), The citation process: the role and significance of citations in scientific communication, London: Taylor Graham
5. Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., Callahan, E. (1998), Invoked on the Web, Journal of the American Society for Information Science, 49(14), 1319-1328
6. Dahal, T. M., Cybermetics: the use and implications for Scientometrics and Bibliometrics: a study for developing science & technology information system in Nepal". Found at http://www.panasia.org.sg/nepalnet/ronast/cyber.html, 2000
7. Ding, Y. (1999), Mapping the intellectual structure of information retrieval studies: an author co-citation analysis, 1987-1997, Journal of Information Science, 25(1), 67-78
8. Goodrum, A. A., McCain, K. W. (2001), Scholarly publishing in the Internet age: a citation analysis of computer science literature. (in press)
9. Egghe, L. (2000), New informetric aspects of the Internet: some reflections, many problems, Journal of Information Science, 26(5), 329-335
10. Garfield, E. (1979), Citation Indexing: its theory and application in science, technology and humanities, New York: John Wiley & Sons
11. Garfield, E. (1998), Comment by Eugene Garfield, found at http://info.uibk.ac.at/sci-org/voeb/vhau9402.html, 2000
12. Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C. (1998), Multivariate data analysis (Fifth edition), Upper Saddle River, NJ: Prentice Hall
13. Harter, S. P., Kim, H. J. (1996), Electronic Journals and Scholarly Communication: A citation and reference study, Proceedings of the 1996 Midyear Meeting of ASIS, p. 299-315
14. ISI, The ISI Database: the journal selection process, Found at http://www.isinet.com/isi/hot/essays/199701, 2000
15. Larson, R. R. (1996a), Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace, Proceedings of the 59th ASIS Annual Meeting, Baltimore, MD, Oct. 21-24, 1996, p71-78. Medford, NJ: Information Today/ASIS
16. Larson, R. R. (1996b), Co-Citation analysis and the WWW, Found at: http://sherlock.berkeley.edu/docs/asis96/node4.html, 2000
17. Lawrence, S., Giles, C. L., Bollacker, K. (1999), Digital libraries and autonomous citation indexing, IEEE Computer, 32(6), 67-71
18. Lindsey, D. (1980), Production and citation measures in the sociology of science: the problem of multiple authorship, Social Studies of Science, 10, 145-162
19. Long, J. S., McGinnis, R., Allison, P. D. (1980), The problem of junior-authored papers in constructing citation counts, Social Studies of Science, 10, 127-143
20. Lu, S. (1999), The transition to the virtual world in formal scholarly communication: a comparative study of the natural sciences and the social sciences, Dissertation, University of California, Los Angeles, 1999
21. MacRoberts, M. H., MacRoberts, B. R. (1989), Problems of citation analysis: a critical review, Journal of the American Society for Information Science, 40(5), 342-349
22. McCain, K. W. (1988), Evaluating cocited author search performance in a collaborative specialty, Journal of the American Society for Information Science, 39(6), 428-431
23. McCain, K. W. (1990a), Mapping authors in intellectual space: population genetics in the 1980s, In: Borgman, C. L. (ed.), Scholarly communication and bibliometrics, Newbury Park, CA: Sage Publications Inc., p. 194-216
24. McCain, K. W. (1990b), Mapping authors in intellectual space: a technical overview, Journal of the American Society for Information Science, 41(6), 433-443
25. McCain, K. W. (2000), Sharing digitized research-related information on the World Wide Web, Journal of the American Society for Information Science, 51(14), 1321-1327
26. Meho, L. I., Sonnenwald, D. H. (2000), Citation ranking versus peer evaluation of senior faculty research performance: a case study of Kurdish scholarship, Journal of the American Society for Information Science, 51(2), 123-138

27. Qin, J., Lancaster, F. W., Allen, B. (1997), Types and levels of collaboration in interdisciplinary research in the sciences, Journal of the American Society for Information Science, 48(10), 893-916
28. Rogers, E. M., Cottrill, C. A. (1990), An author co-citation analysis of two research traditions: technology transfer and the diffusion of innovations, In: Borgman, C. L. (ed.), Scholarly communication and bibliometrics, Newbury Park, CA: Sage Publications Inc., p. 157-165
29. Rousseau, R. (1997), Sitations: an exploratory study, Cybermetrics, 1(1). Found at: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html , 2001
30. Smith, L. E. (1981), Citation analysis, Library Trends, 30, 83-106
31. Small, H., Greenlee, E. (1990), A co-citation study of AIDS research, In: Borgman, C. L. (ed.), Scholarly communication and bibliometrics, Newbury Park, CA: Sage Publications Inc., p. 166-193
32. Stokes, T. D., Hartley, J. A. (1989), Coauthorship, social structure and influence within specialties, Social Studies of Science, 19, 101-125
33. Turnbull, D., Bibliometrics and the World-Wide Web, Found at http://donturn.fis.utoronto.ca/research/bibweb.html, 2000
34. White, H. D. (1990), Author co-citation analysis: overview and defense, In: Borgman, C. L. (ed.), Scholarly communication and bibliometrics, Newbury Park, CA: Sage Publications Inc., p. 84-106
35. White, H. D., Griffith, B. (1981), Author co-citation: a literature measure of intellectual structure, Journal of the American Society for Information Science, 32, 163-172
36. White, H. D., McCain, K. W. (1998), Visualizing a discipline: an author co-citation analysis of information science, 1972 – 1995. Journal of the American Society for Information Science, 49(4), 327-355
37. Youngen, G. (1997), Citation patterns of the physics preprint literature with special emphasis on the preprints available electronically, Found at http://www.physics.uiuc.edu/library/preprint.html, 2000