Journal of Informetrics 2 (2008) 229–239

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Comparing all-author and first-author co-citation analyses of information science

Dangzhi Zhao[a,*], Andreas Strotmann[b]

[a] School of Library and Information Studies, University of Alberta, 3-20 Rutherford South, Edmonton, AB, Canada T6G 2J4
[b] School of Business, University of Alberta, Edmonton, AB, Canada T6G 2R6

ARTICLE INFO

Article history:
Received 12 March 2008
Received in revised form 23 May 2008
Accepted 27 May 2008

Keywords:
Author co-citation analysis
Bibliometrics
Information science
Scopus
Citation analysis

ABSTRACT

Although it is generally understood that different citation counting methods can produce quite different author rankings, and although "optimal" author co-citation counting methods have been identified theoretically, studies that compare author co-citation counting methods in author co-citation analysis (ACA) studies are still rare. The present study applies *strict* all-author-based ACA to the Information Science (IS) field, in that *all* authors of *all* cited references in a classic IS dataset are counted, and in that even the diagonal values of the co-citation matrix are computed in their theoretically optimal form. Using Scopus instead of SSCI as the data source, we find that results from a theoretically optimal all-author ACA appear to be excellent in practice, too, although in a field like IS where co-authorship levels are relatively low, its advantages over classic first-author ACA appear considerably smaller than in the more highly collaborative ones targeted before. Nevertheless, we do find some differences between the two approaches, in that first-author ACA appears to favor theorists who presumably tend to work alone, while all-author ACA appears to paint a somewhat more recent picture of the field, and to pick out some collaborative author clusters.

1. Introduction

Ever since its introduction by White and Griffith (1981), author co-citation analysis (ACA) has been a primary research tool for the study of the intellectual structure of research fields, and of the social structure of the underlying communities of researchers. Studies have since then usually applied, to a variety of scholarly fields, the general steps and techniques of classic ACA as defined there, although a few studies have proposed new techniques for mapping author clusters (White, 2003b), different statistical methods for processing co-citation counts (Ahlgren, Jarneving, & Rousseau, 2003), or variations on the statistical procedures used in ACA such as MDS (Leydesdorff & Vaughan, 2006). Few studies have dealt with the problematic definition of co-citation counts themselves—a much more fundamental aspect of ACA as it defines the raw data from which all statistical author co-citation analyses and mappings are derived.

We recently reported on a preliminary study that aimed to contribute to filling this gap (Zhao, 2006). The present study was conducted as a follow-up verification study that seeks to further contribute to understanding the role of different co-citation counts in ACA.

* Corresponding author. Tel.: +1 780 4922814; fax: +1 780 4922430.
E-mail addresses: dzhao@ualberta.ca (D. Zhao), andreas.strotmann@ualberta.ca (A. Strotmann).

1751-1577/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2008.05.004

2. Problem statement and research questions

ACA defines two authors as co-cited when at least one document from each author's oeuvre occurs in the same reference list, and their co-citation count as the number of different publications that co-cite them in this sense. Classic ACA defines an author's oeuvre as all the works with the author as the first author (McCain, 1990), and thus only uses information on the first author of a publication when calculating author co-citation counts.

This definition stems to a considerable degree from the relative ease with which these *first-author co-citation counts* can be retrieved directly from the main data source for ACA studies—the citation indexes of the Institute for Scientific Information (ISI) (now Thomson Scientific), because these citation indexes only index the first authors in cited references. The only way to find the full set of authors of a cited document in these citation indexes is by matching the reference code to a corresponding source paper—an operation that is laborious and error-prone, and only possible if the cited reference refers to a document that also happens to have been indexed as a source paper (citing paper). As a result, it has been practically impossible to go beyond first-author co-citation counting using these databases.

Rousseau and Zuccala (2004) explored different ways of defining author co-citation counts and strategies for retrieving these counts from the ISI databases, but did not use them in actual ACA studies. A brief study by Persson (2001) is the only study we know of that attempts to compare first-author and all-author co-citation analysis using pure ISI data—all others, including our own, have had to rely on other sources of citation data.

Persson (2001) retrieved all the 7001 articles in the LIS field that were indexed in one of the ISI databases between 1986 and 1996, and calculated how these articles co-cited *each other*. It observed similar subfield structures derived by first- or all-author co-citation analysis, and noticed that many well-known authors on the all-author co-citation map were excluded from the map based on first-author co-citation counts. Unfortunately, more than 90% of cited references in that dataset refer to documents *outside* it, whose authors the study therefore disregards, which severely limits the reliability of the results of his courageous first attempt. In addition, this brief communication provided little detail on how co-citation counts were defined and calculated.

Recently, alternatives to the ISI citation indexes have become available that provide better support for collecting co-citation data that allows us to perform true all-author co-citation analysis studies, i.e., studies that rely on a dataset that includes all authors of all cited references.

Zhao (2006) used one of these new databases, namely *CiteSeer*, for this purpose, and conducted a preliminary study that compared first-author and all-author[1] co-citation analyses of the XML research field, a subfield of computer science, the research field covered by that citation index. It was found that, at least in the XML research field, the all-author ACA resulted in a clearer picture of the intellectual structure than the classic first-author ACA did, at the price of identifying fewer specialties if the same number of most highly cited authors was selected.

Two possible definitions were identified in Zhao (2006) that extend first-author to all-author co-citation counts, and results of the corresponding ACAs were compared.

(a) *Inclusive all-author co-citation* derives from classic first-author co-citation by simply redefining an author's oeuvre as everything that lists the author as *one of* its authors rather than as its *first* author, and by retaining the definition of co-citedness between two authors as the number of papers that reference both authors' oeuvres. This definition therefore *includes* cited co-authorship in co-citation, because two authors are also considered co-cited when a paper they co-authored is cited.
(b) *Exclusive all-author co-citation* derives from inclusive all-author co-citation by *excluding* pure cited co-authorship and re-defining co-citedness between two authors as the number of papers that contain *distinct* references to both authors' oeuvres, i.e., references to these authors' oeuvres that do not both refer to the same document co-authored by these two authors.

Note that inclusive and exclusive author co-citation counts are identical when calculated for distinct authors who both only ever wrote single-authored papers. Same-author co-citation counts, however, will differ—inclusive counts would result in an author's number of citations, while exclusive counts strictly count the number of times that that author is co-cited with him or herself. The latter therefore provides a meaningful diagonal even for a first-author co-citation matrix. Zhao (2006) found in favor of exclusive co-citation counts when studying intellectual structures, although inclusive co-citation analysis showed some promise for studying social relationships using ACA.

Schneider, Larsen, and Ingwersen (2007) built on Zhao (2006), and further explored inclusive all-author co-citation analysis using a large citation index generated from a XML version of the IEEE Computer Society journals for the years 1995–2004. They found that all-author-based ACA can help produce MDS maps that better fit the underlying data, and may lead to stronger concentration in the maps.

Eom (2008) used a hand-made citation database that covered 692 citing papers in the decision support systems area during 1971–1990, in an attempt to compare first-author- and all-author-based ACA results. Instead of controlling for the

[1] Actually, that study limited itself to up to five authors, but only a small fraction of cited references in that dataset had more than five authors.

Table 1
Journals used to define information science[a]

Information science	Library automation
Annual Review of Information Science and Technology	Electronic Library
Information Processing & Management (and Information Storage & Retrieval)	Information Technology and Libraries (and Journal of Library Automation)
Journal of the American Society for Information Science and Technology	Library Resources & Technical Services
Journal of Documentation	Program—Automated Library and Information Systems
Journal of Information Science	
Library & Information Science Research (and Library Research)	
Proceedings of the American Society for Information Science and Technology (and Proceedings of the ASIST Annual Meeting)	
Scientometrics	

[a] Taken from White and McCain (1998, p. 330) with only minor updates.

effect of the total number of authors included in these two types of ACA, which, experience tells us, tends to be correlated significantly with the number of specialties identified as a result of an ACA, Eom's study used the same absolute citedness threshold for both types of ACA, resulting, of course, in a significantly smaller number of authors in his first-author ACA study than in his all-author ACA study. By failing to control for such a significant variable, his study was unfortunately not able to actually show the differences resulting from different author co-citation counting methods.

The present study revisits the question of first- vs. all-author co-citation analysis and of different types of all-author co-citation counting in a different, less collaborative, field (i.e., Information Science) than those targeted by previous studies, using yet another citation index to collect data, namely, Elsevier's Scopus. The research fields studied previously, namely subfields of computer science and engineering, have a considerably higher level of collaboration as measured by the average number of coauthors in their literature than the IS field studied here. Unlike our previous preliminary study, which only counted up to five authors of a cited reference, and possibly like some of the other previous studies reviewed above, the present study strictly counts every author of every cited reference in order to eliminate any potential problems with previous simplified approaches. Based upon more than 3000 citing papers, the scale of the present study is also considerably larger than the preliminary study and most of the previous studies. In addition, we test in practice a way of treating diagonal values of author co-citation matrices in ACA that has been identified as theoretically optimal but has not been empirically studied.

Specifically, the research questions addressed in this study are

(a) Does all-author-based ACA produce a clearer picture of the intellectual structure of the IS field with higher concentration?
(b) What are the differences between different ways of calculating author co-citation counts in terms of the intellectual structure of the IS field revealed?

Addressing these research questions will contribute to a better understanding of ACA in general, and to all-author-based ACA in particular. The latter is particularly important because all-author-based ACA, on the one hand, is theoretically optimal, and on the other hand, can be expected to be necessary to do justice to researchers in the highly collaborative "big sciences", which are of particular interest to science and technology policy makers.

3. Methodology

3.1. Data collection

The research area we analyse in the present study is information science. As in White and McCain (1998), we define the IS research field by its 12 core journals listed in Table 1 (White & McCain, 1998, p. 330). Our citation window is a 10-year period (i.e., 1996–2005) following that studied in White and McCain (1998).

We use Elsevier's Scopus database to retrieve citing papers in the IS field thus defined, along with their reference lists. At the time of data collection, Scopus indexed journal articles published in 1996 or later, and thus was able to provide us with the intended dataset very well. It is similar to the ISI citation indexes in that it indexes cited references of the (citing) papers it covers, but it also provides, directly, more information on cited references than the ISI citation indexes do, including their full titles and the names of up to eight authors (the first seven and the last author) in each cited reference.

Journal by journal, we retrieved this information for all papers published in the 12 IS journals during 1996–2005 that are indexed in Scopus as "articles", and exported them in "RIS format" as "Full Documents" to a local computer. This way, we collected 3828 records of citing papers that have references. These papers included 110,785 references altogether, i.e., 29 references per citing paper on average.

For the few cited references that had more than eight authors, we manually completed the author lists, by searching for these papers as citing papers in a number of data sources. This is only feasible in a research field like IS where large-group collaboration is uncommon, of course, but it does eliminate completely any potential side effects of disregarding any authors

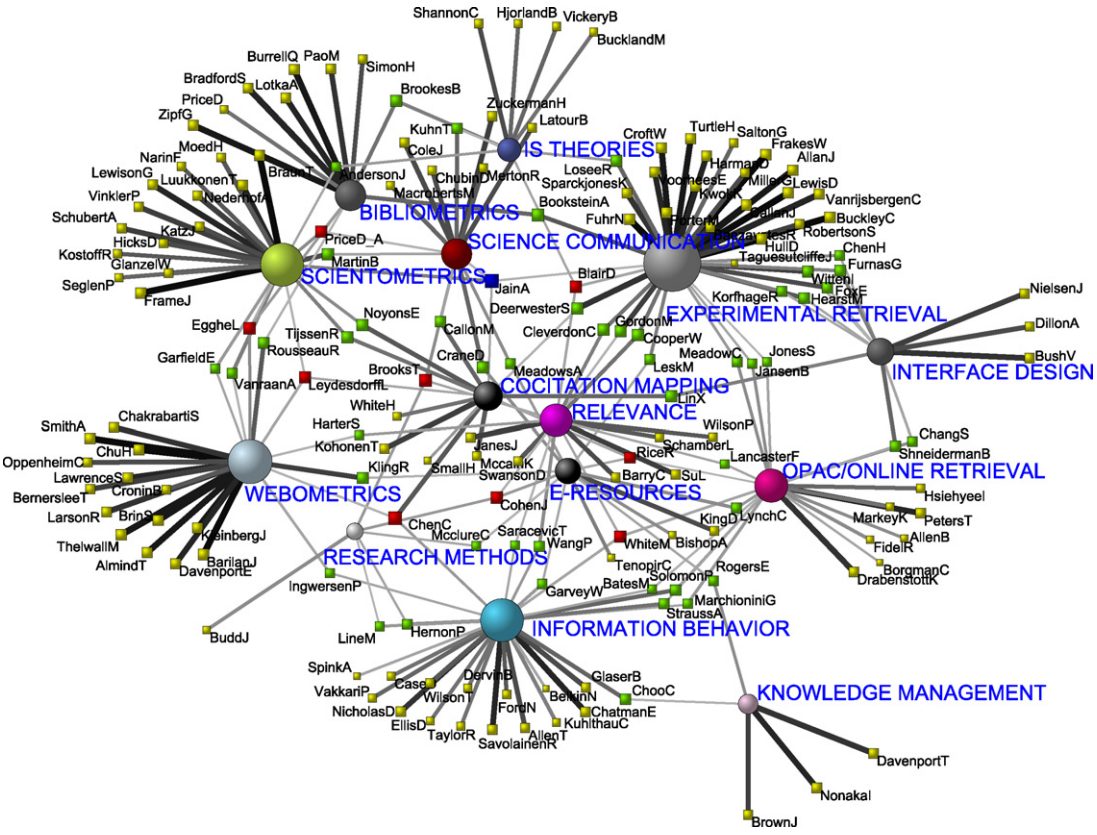

Fig. 1. Factor analysis results from first-author co-citation counts.

at all in cited references. In more intensely collaborative research fields, the limitations imposed by Scopus may either be entirely acceptable to those who try to study them using ACA, or it may be necessary to develop additional methodology to match cited references to metadata in the same or other bibliographical databases that do include all their authors.

We developed software to parse the downloaded and hand-completed records, and to store the resulting data fields such as author names, publishing sources, and years of publication of both source papers and cited references in a structure (essentially, a pair of tables, one for citing papers, one for cited references) that simplified the subsequent data analysis procedures, e.g., counting citations and co-citations.

We noticed that the number of papers retrieved from Scopus is slightly smaller than that from the ISI databases across the 12 journals. For example, in the journal *Information Processing & Management* we retrieved 479 and 506 papers, respectively, from Scopus and from the ISI databases (through Web of Science); for JASIST, we retrieved 976 and 1022, respectively, from these sources. There was no immediately discernible pattern as to which articles are indexed in one but not in another, but the differences appeared minor. In addition, a comparison between the classic first-author-based ACA results using Scopus on the one hand and Web of Science data on the other, as represented in Fig. 1 in this paper and Fig. 2 in Zhao and Strotmann (2008), respectively, indicates that the specialty structures revealed from these two data sources are very similar. We are therefore confident that using Scopus for ACA studies can provide us with a view of the intellectual structure of the IS field that matches closely results that the ISI databases might have provided.

3.2. Data analysis

We conducted ACAs using factor analysis based on first-author, inclusive all-author, and exclusive all-author co-citation counts, as perceived by the authors of these 3828 publications as citers.

We followed commonly accepted steps and techniques of ACA (McCain, 1990; White & McCain, 1998; Zhao, 2003) except for the different definitions of co-citation as discussed earlier. Core sets of authors were selected based on "citedness"—the number of citations they received. Two sets of highly visible authors were thus selected using two different citation counting methods—first-author counts and complete counts. Simply put, when a paper with N authors is cited, with first-author counts, only the number of citations of the first author of this paper increases by 1, and with complete counts, full credit is given to all authors of the paper, i.e., the number of citations of each of its N authors increases by 1.

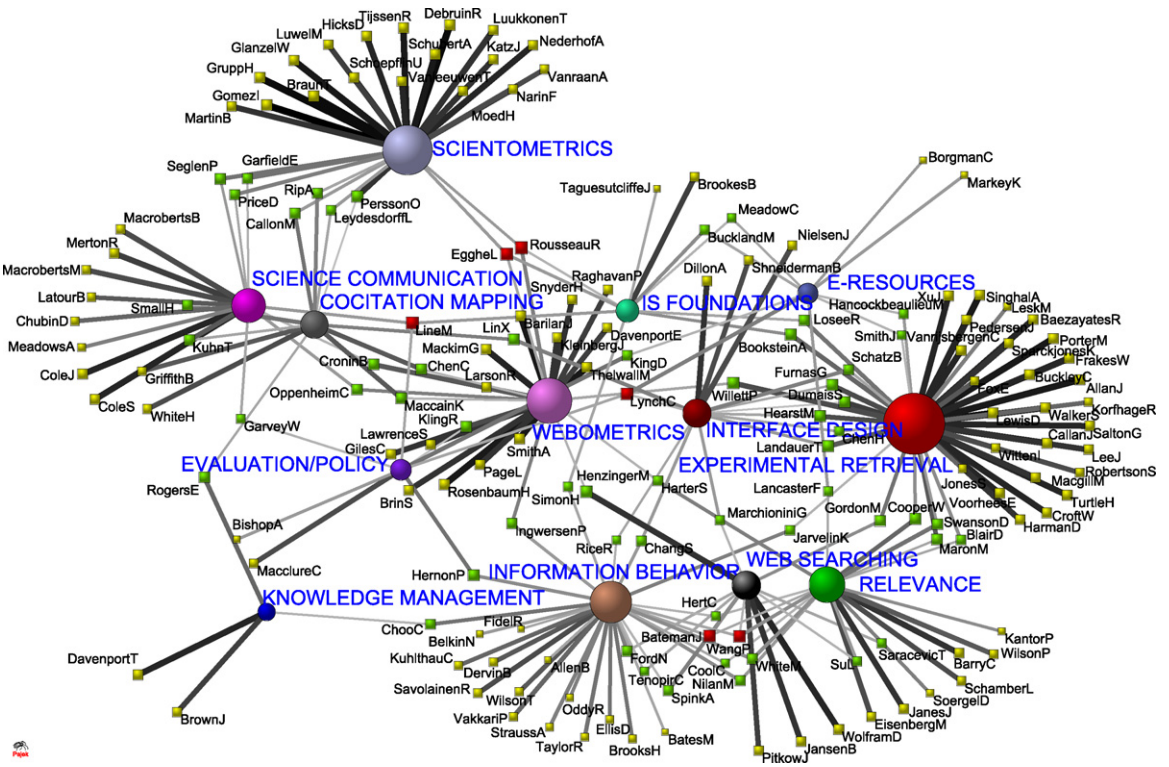

Fig. 2. Factor analysis results from exclusive all-author co-citation counts (with meaningful diagonal values).

There are no strict rules regarding thresholds for citation-based author selection in author co-citation analysis studies (McCain, 1990). Assuming that the more authors the better a research field is represented, the present study used an arbitrary number of 165 as the number of authors to be included in the final factor analyses, a number that is larger than the 120 considered in White and McCain (1998) as adequate for the purpose of this type of studies.

Software was developed to count co-citation frequencies by the three methods discussed above, and to record them in three separate matrices. These co-citation matrices were then used as input to the Factor Analysis (FA) procedure in SPSS. As usual, the diagonal values in these matrices were deleted from the input files to the FA routine in SPSS, and were treated as missing values and replaced by the mean in SPSS.

In addition, we performed a successful experiment with more meaningful diagonal values of the co-citation matrix in the case of exclusive all-author co-citation counts. In this case, instead of treating the diagonal as missing values, its values were determined as follows: the diagonal value for author A (i.e., author A's auto-co-citation count) increases by 1 when *at least two different* articles with author A as one of the authors appear in the same reference list. This definition is consistent with the off-diagonal values because, according to the definition of exclusive all-author co-citation counts, the co-citation count of author A and author B increases by 1 when a citing article's reference list contains at least one article with author A as one of the authors and at least one *additional* article with author B as one of the authors. Otherwise, authors A and B would just be co-authors of a single cited paper, which per our definition does not count. In fact, in our computer program, the code for counting exclusive all-author co-citations is identical for all cells of the matrix, including the diagonal. This way of treating diagonal values of co-citation matrices in ACA is ideal in theory (Rousseau & Zuccala, 2004), but has not before been empirically studied as to whether it is indeed better than other ways of dealing with diagonal values in the study of the intellectual structure of research fields.

As usual, we used the resulting exclusive all-author co-citation matrix with meaningful diagonal values as input to the FA routine in SPSS. As shown below, this matrix of co-citation counts produces good results.

In all cases, factors were extracted by Principal Component Analysis (PCA) with an oblique rotation (SPSS Direct OBLIMIN). An oblique rotation was chosen because it is often more appropriate than orthogonal rotations when it can be expected theoretically that the resulting factors (in this case, specialties) would in reality be correlated (Hair, Anderson, Tatham, & Black, 1998). The number of factors extracted was determined based on Kaiser's rule of eigenvalue greater than 1 because the resulting model fit was good in all four cases as represented by total variance explained, communalities, and correlation residuals (Hair et al., 1998). The factor models produced this way are shown in Table 2 along with their model fits. For example, a factor analysis of the first-author co-citation matrix resulted in a 14-factor model which explains 84% of the total variance, and the differences between observed and implied correlations were smaller than 0.05 for the most part (almost

Table 2
Factor models and their model fits

Input co-citation matrix	Factor model	Total variance explained (%)	% |nonredundant residuals| > 0.05 (%)	Communalities		
				Range	<0.7	<0.8
First-author	14-factor	84	0	0.60–0.94	6 (4%)	28 (17%)
Exclusive all-author with meaningful diagonal values	13-factor	91	0	0.68–0.99	1 (0.6%)	7 (4%)
Inclusive all-author	12-factor	84	0	0.55–0.95	5 (3%)	38 (23%)
Exclusive all-author with diagonal values deleted	11-factor	87	0	0.58–0.95	3 (2%)	14 (8%)

100%). The communalities ranged from 0.60 to 0.94, only 6 (or 4%) of which were below 0.7 and 28 (or 17%) of which below 0.8.

The numbers in Table 2 indicate that exclusive all-author-based ACA appears to produce statistically more significant results in terms of the amount of variance explained by factor models extracted according to identical criteria, especially when meaningful diagonal values are used for the factor analysis.

3.3. Visualization of factor structures

Recently a novel way of visualizing factor analysis results in ACA was introduced in Zhao and Strotmann (2008) that provides very informative two-dimensional visual maps to aid the interpretation of results, and that allows us to report results of multiple ACA studies in a single paper as we do here. To prepare these maps, the factor labels were assigned, as usual, upon examining the frequently cited articles written by authors in each factor that load highly on them. In these graphs, authors are represented by square nodes, and factors are represented by circular nodes. Factor and author nodes are connected by lines if the author loads sufficiently on the factor (i.e., 0.3 or higher in this study). The width of such a line is proportional to the value of the author's loading on this factor, as is its gray-scale value. Node sizes are accumulated from loadings, and are intended to increase with the importance of a node in the factor structure map.

The layout of these maps is an automatically generated Kamada–Kawai graph layout using loadings as similarity measures between author and factor nodes, produced by Pajek (Batagelj & Mrvar, 2007). The result is an intellectual structure map of the field that is visually informative and true to the factorization it represents, but it is also quite different from the MDS maps with author clusters from cluster analysis in a traditional ACA.

4. Results

We will first discuss the intellectual structure of the Information Science field based on the factor analysis results from classic first-author-based ACA (Fig. 1). Then, we will examine how this structure compares with the intellectual structures shown from other types of ACA, i.e., results from an exclusive all-author co-citation analysis with meaningful diagonal values (Fig. 2), those from an inclusive all-author co-citation analysis (Fig. 3), and those from an exclusive all-author co-citation analysis with missing diagonal values (Fig. 4). All these maps are visualizations of the pattern matrix of the corresponding factor analysis with an oblique rotation.

4.1. Classic first-author-based ACA

A factor analysis of the first-author co-citation matrix of the 165 most highly cited authors in the IS field in the years 1996–2005 identifies 14 specialties (Fig. 1). The four specialties that were most active during this time period are experimental retrieval, information behavior, scientometrics/citation analysis, and webometrics. More than half (56%) of the authors analysed belong to these four specialties. The smaller specialties identified are OPACs and online retrieval, bibliometrics, science communication, co-citation mapping, e-resources organization and retrieval, IS theories, relevance, interface design, knowledge management, and research methods.

Comparing with the situation of the field 10 years earlier as described in White and McCain (1998) using equivalent methodology, we can see some trends in the development of the IS field. Both the experimental retrieval specialty and the scientometrics/citation analysis specialty remain active research areas. Research on information behavior or user theory has grown dramatically, and webometrics has emerged as a new, active and distinct specialty of study.

The bibliometrics and science communication specialties remain largely unchanged. OPACs and online retrieval have merged into a single specialty which has been joined by authors who study web searching, indicating perhaps the trend of integrating OPACs and bibliographic databases into library web portals. By contrast, research on co-citation mapping has been separated out from the general scientometrics/citation analysis specialty as a new research focus that applies co-citation analyses to the mapping of science.

It appears that the imported ideas group of White and McCain (1998) has split into two groups: IS theories (e.g., Shannon) and interface design (e.g., Bush, Shneiderman), both joined by additional authors.

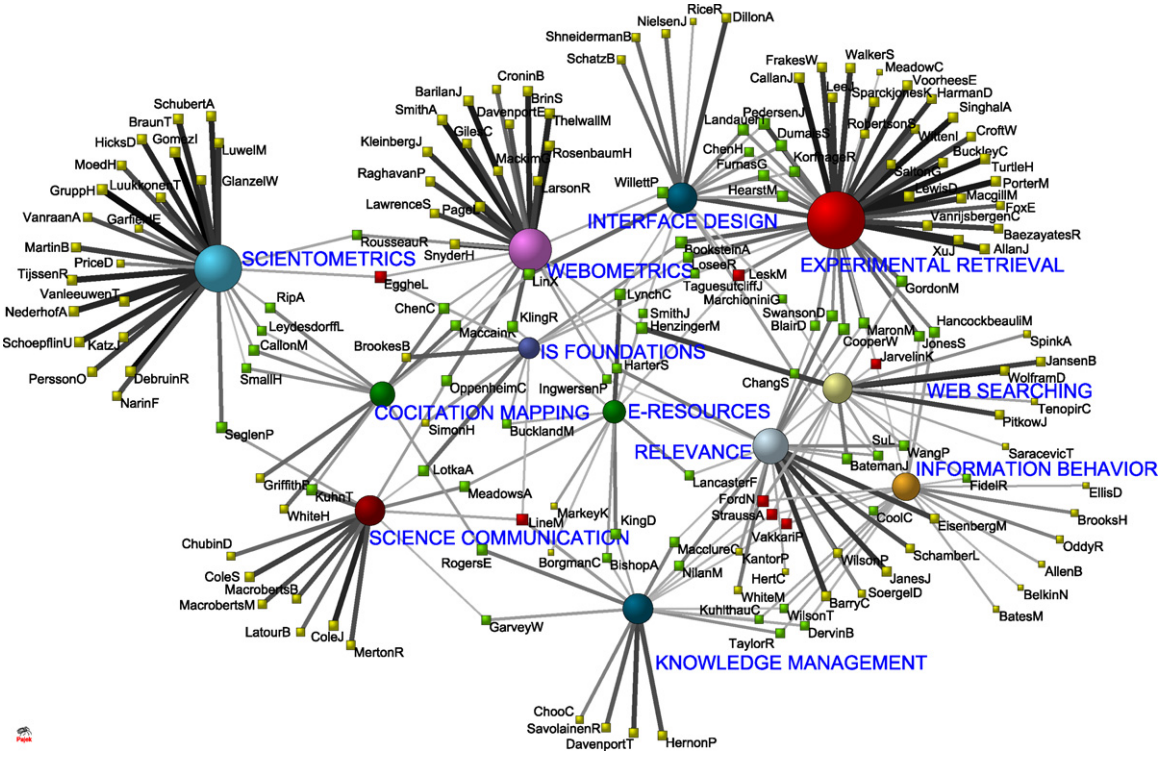

Fig. 3. Factor analysis results from inclusive all-author co-citation counts.

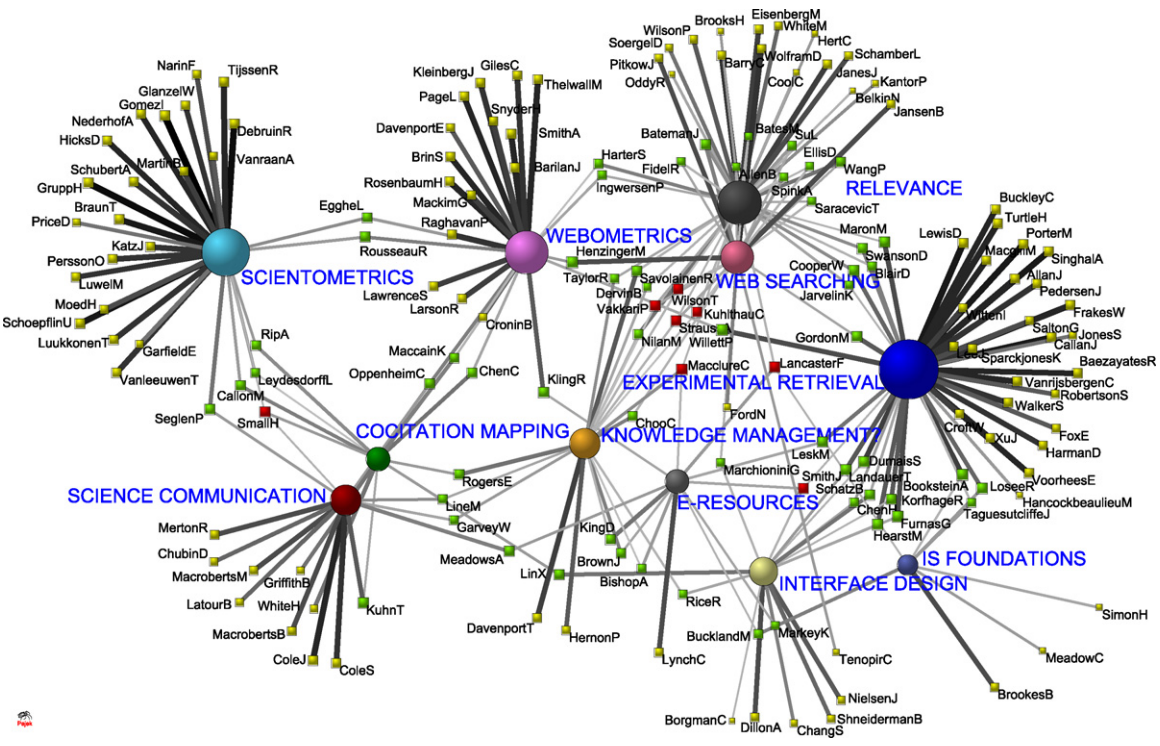

Fig. 4. Factor analysis results from exclusive all-author co-citation counts (diagonal values deleted).

In addition to webometrics, two other new specialties have emerged: relevance and knowledge management. The knowledge management specialty, although small, is quite clear, reflecting the research reality in the IS field quite well. Another small group composed of Budd, Cohen and McClure is quite vague, but appears to be about research methods.

It appears that the ACA results reflect the development of the IS field over the past 10 years quite well. It would be interesting to examine how individual authors have changed their specialty concentrations. We will leave it to readers who are familiar with the authors on the map to see how well the ACA results match their perceptions in this regard, as our focus here is the comparison between different types of ACA.

4.2. Exclusive all-author-based ACA with meaningful diagonal values

13 specialties are identified from a factor analysis of an exclusive all-author co-citation matrix of 165 authors with meaningful diagonal values (Fig. 2). The four specialties identified in this analysis that were most influential in the past 10 years are the same as those from first-author-based ACA as shown in Fig. 1. In this analysis, about 60% of the authors analysed represent these four specialties. Most of the smaller specialties identified are also the same as in the first-author co-citation analysis, including the co-citation mapping, interface design, science communication, bibliometrics, relevance, and knowledge management specialties.

Here in the results from exclusive all-author-based ACA, only one of the five authors who primarily loaded on the IS theories specialty in the results from first-author-based ACA made the list of top 165 authors (i.e., Buckland), and many authors in the bibliometrics specialty did not show up either (e.g., Zipf, Bradford, Lotka, Burrell). As a result, these two groups merged into a single group led by Brookes who had done significant research in both areas. It appears that this group of authors represent foundations and history of IS, especially of its quantitative aspects. We thus tentatively label it as "IS foundations".

This and the disappearance of the small and vague group on research methods identified by the first-author-based ACA indicate that first-author-based ACA may better represent the theoretical and methodological aspects of the field. It is possible that this is due in part to the use of full citation counts for ranking highly cited authors in the field. This method disadvantages authors who publish alone, as theorists are wont to do. It remains to be seen in later studies if fractional citation count rankings would retain these fields.

The focus of the OPAC/online/web searching specialty seen in the results from first-author-based ACA has shifted from OPACs/online retrieval systems to web search engines in the results from exclusive all-author ACA, resulting in the distinct web searching specialty. This suggests that first-author ACA represents older research better whereas all-author ACA reflects current studies more clearly.

An examination of the highly cited publications by authors in the web searching shows that several authors in this specialty co-authored papers, such as Wolfram, Jansen, Spink, Bateman, and Saracevic. A similar observation applies to the evaluation/policy specialty where McClure and Hernon co-authored highly cited works. It therefore appears that all-author-based ACA can pick out some tightly connected research groups or projects.

4.3. Inclusive all-author-based ACA

The results from an inclusive all-author-based ACA as shown in Fig. 3 are very similar to those from exclusive all-author-based ACA with meaningful diagonal values as shown in Fig. 2, except that 12 rather than 13 specialties are identified and that the information behavior specialty is much smaller after several authors moved to a specialty that is a mix of authors from the evaluation/policy group, the knowledge management specialty, and science communication in Fig. 2. The resulting mix does not make much sense to us at first glance. An examination of their highly cited works suggests that this group of authors appear to have been cited as relevant to the study of knowledge management in a wider sense, such as Davenport and Rogers' works on innovation, Hernon and McClure's studies on evaluation of library services, Savolainen's everyday life information seeking, Choo's environment scanning, Taylor's information use environment, Garvey's science communication, and Strauss' grounded theory. Readers who have deeper knowledge of these authors' research areas may make better sense of it. We thus tentatively label this group knowledge management and list its authors here: Davenport T, Hermon P, Savolainen R, Rogers E, McClure C, Choo C, Taylor R, Garvey W, Line M, Dervin B, and Strauss A.

4.4. Exclusive all-author-based ACA with diagonal values deleted

The results from an exclusive all-author-based ACA with missing diagonal values (Fig. 4) are very similar to those from inclusive all-author-based ACA shown in Fig. 3, except that 11 rather than 12 specialties are identified, with the information behavior specialty splitting into two groups which merge into the relevance specialty (Brooks H, Cool C, Jarvelin K, Vakkari P, Oddy R, Kuhlthau C, Belkin N) and into the web searching specialty (Allen B, Ellis D, Bates M), respectively. The group tentatively labeled knowledge management appears here again as a mix of authors from several of the specialties identified in the results from exclusive all-author-based ACA with meaningful diagonal values: knowledge management (Davenport T, Rogers E, Brown J), evaluation/policy (Hermon P, McClure C), information behavior (Savolainen R, Choo C, Taylor R, Dervin B, Wilson T), and science communication (Garvey W).

4.5. Missing vs. meaningful diagonal values in ACA

Comparing results from the two types of exclusive all-author-based ACA, that with meaningful diagonal values (Fig. 2) vs. that without diagonal values (Fig. 4), we find the factor analysis results become less informative as information is discarded with the deletion of diagonal values in the co-citation matrix. The evaluation/policy group merges into the knowledge management specialty; the information behavior specialty in Fig. 2 disappears into three other specialties: relevance, web searching, and knowledge management. As most of these authors load low (less than 0.5) on the factors they merge into, they do not add much information to the picture. Deleting the diagonal values and letting SPSS replace them with the mean therefore appears to cause a noticeable loss of information from the co-citation matrix.

5. Discussion

The major specialty structure of the Information Science field produced by author co-citation analysis is largely the same whether it is first-author- or all-author-based, confirming findings from Persson (2001)'s simplified study. This is due most likely to the relatively low number of co-authored papers in this field, which generally makes for relatively small differences between first- and all-author co-citation counts. Nevertheless, even in this field with a low level of collaboration, there are noticeable differences in results from different types of ACA. First-author ACA appears to highlight the theoretical and methodological aspects of the field, while all-author ACA results appear to be clearer when it comes to representations of current trends (e.g., web searching), and appear to pick out highly collaborative research groups.

In our preliminary all-author ACA study of the XML field, we noticed that all-author ACA tended to result in a smaller number of specialties than first-author ACA when the same number of most highly cited authors was selected for analysis. Although this is confirmed here for the IS field, the strength of this effect depends heavily on the particular choice of all-author ACA methodology: compared to the 14 IS specialties identified through a classic first-author ACA with missing diagonal values in the co-citation matrix, exclusive all-author ACA yielded a mere 11 in the case of *missing* diagonal values and 13 when meaningful diagonal values were used, while inclusive all-author ACA identified 12 specialties.

All in all, classic first-author ACA results in a picture of the intellectual structure of the IS research field that is of similar structure and comparable clarity to that found by exclusive all-author co-citation analysis with meaningful diagonal values. In the case of inclusive all-author co-citation analysis and exclusive all-author ACA with missing diagonal values, on the other hand, it is first-author ACA that appears to produce a clearer picture, unless we correctly identified the common interest of a somewhat mixed group of cited authors that we tentatively labeled knowledge management.

Given the low level of collaborative research and publishing in the IS field (less than two citing authors on average), especially with respect to the cited references that we actually analyse in ACA (close to 1.5 cited authors per reference on average), we have retrospectively and independently validated the use of first-author instead of all-author ACA in fields like IS, in that it appears to produce quite similar results to an ACA that *fully* takes into account *all* authors of *all* referenced articles. This assumes that, in general, it is preferable to include *all* available information in a statistical analysis rather than biasing it by arbitrarily ignoring some of the data, and that therefore, in our case, the results from an all-author co-citation analysis that uses meaningful values in the diagonal of the co-citation count matrix are as close to the "true" intellectual structure of a field as an ACA can get. Indeed, Rousseau and Zuccala (2004) identified all-author ACA with the meaningful diagonal values we used here as theoretically optimal.

Conversely, the close correspondence between the results of the classic first-author ACA and the all-author ACA that uses an exclusive all-author co-citation matrix with meaningful diagonal values provides evidence that the latter is an excellent candidate for valid co-citation analyses that do take into account all authors of all cited papers. This is because we verified that, in a situation where similar results are to be expected from both, this alternative method works just as well as the classic first-author ACA whose validity has been amply demonstrated in a range of research fields like IS.

In the XML field that we investigated in our preliminary study, we find about three authors per citing paper on average, and about 2.5 authors on average per cited reference. Unlike in the IS field, the average number of cited authors per reference for the XML field is thus significantly larger than 1.0 (the corresponding value for first-author ACA), so that we can expect to see a much larger difference (by a factor of three, in fact[2]) between all-author and first-author analyses in the XML field than in the IS field, and that is indeed what we found. In the XML field, we found that even a "lesser" type of all-author ACA tended to produce a clearer picture of the intellectual structure, and that it is therefore important to count all authors of cited papers in co-citation analysis studies. In this paper, we find that we need to amend this statement slightly: the higher the level of collaborative research is in a field, the more important it is to find ways to include the names of all authors of all referenced papers in an all-author ACA. In order to further test (and strengthen) this statement, a true highly collaborative research field should be chosen.

[2] $(2.5 - 1.0) = 3 * (1.5 - 1.0)$: the difference in average co-authorship levels between all- and first-author citation counting is three times higher in XML than in IS.

This underscores suspicions voiced by White (2004), that it may not be entirely straightforward to extend first-author ACA methodology to areas where collaborations are the norm rather than the exception, although he appears to be more concerned about the problem of reliably studying individual author citation identities when practically all their papers are multi-authored. Since "big sciences" of this sort are of particular interest to social scientists and science and technology policy researchers, however, we suspect that it is very important to develop a better understanding of all-author citation analysis methodologies—and this paper hopefully contributed to this.

Further research will need to be done, however, in order to clarify findings from our two studies and those of others that are beginning to appear. In particular, it is important to test if, in an increasingly collaborative academic environment, citation and co-citation analyses do indeed need to upgrade from first-author to all-author citation counting methods if they are to realize the goal of measuring realistic scholarly communication indicators.

6. Conclusion

In this study, we find from several ACA analyses that the information technology revolution in general, and the World Wide Web in particular, have had a significant impact on the field, witness the emergence of the specialty of webometrics, the shift of traditional OPAC research towards web searching, and the budding-off from scientometrics of the compute- and web-resource-intensive co-citation mapping specialty.

Many studies on various ways of allocating credit among co-authors have shown that counting all authors can result in very different author rankings by number of citations or publications compared to counting just first authors. Results from the present study indicate that counting all authors in ACA studies, however, may result in very similar specialty structures compared to classic ACA that only counts first authors, *provided* the level of collaboration in the field studied is sufficiently low.

Nevertheless, all-author- and first-author-based ACAs do appear to pick out a small number of different research foci while maintaining the same major specialty structure, e.g., theoretical and methodological background vs. recent trends. All-author-based ACA also tends to be sensitive to collaborative research projects and groups. Thus, a complete view of the intellectual structure of a research field may require both types of ACA.

The present study also confirms a theoretical prediction (Rousseau & Zuccala, 2004; White, 2003a) that, for exclusive all-author co-citation count matrices, computing meaningful diagonal values produces a statistically more meaningful picture of the intellectual structure of a research field than treating the diagonal as missing values. We suspect that this is true because information in the matrix of co-citation counts is retained and not thrown out by replacing diagonal values with statistically generated values (e.g., the mean or the average of the three highest off-diagonal values).

Further studies are suggested in order to test these findings more thoroughly. A research area in which large-group collaboration is commonplace such as in the biomedical field should be a good choice for this purpose.

Acknowledgements

This study was funded in part by the Social Sciences and Humanities Research Council of Canada (SSHRC), by Genome Canada, and by Genome Prairies. The authors would like to thank Huai-Yang Lim for collecting the data for this study. We are particularly grateful to the anonymous reviewers of this paper for their insightful comments and helpful suggestions.

References

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science*, *54*, 550–560.

Batagelj, V., & Mrvar, A. (2007). *Pajek: Program for analysis and visualization of large networks*. http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/pajekman.pdf Accessed 03.08.07.

Eom, S. (2008). All author cocitation analysis and first author cocitation analysis: A comparative empirical investigation. *Journal of Informetrics*, *2*, 53–64.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. *Journal of the American Society for Information Science and Technology*, *57*(12), 1616–1628.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, *41*, 433–443.

Persson, O. (2001). All author citations versus first author citations. *Scientometrics*, *50*(2), 339–344.

Rousseau, R., & Zuccala, A. (2004). A classification of author co-citations: Definitions and search strategies. *Journal of the American Society for Information Science and Technology*, *55*(6), 513–629.

Schneider, J. W., Larssen, B., & Ingwersen, P. (2007). Comparative study between first and all-author co-citation analysis based on citation indexes generated from XML data. In *Proceedings of the eleventh international conference of the International Society for Scientometrics and Informetrics* (pp. 696–707).

White, H. D. (2004). Reward, persuasion, and the Sokal Hoax: A study in citation identities. *Scientometrics*, *60*(1), 93–120.

White, H. D. (2003a). Author cocitation analysis and Pearson's *r*. *Journal of the American Society for Information Science and Technology*, *54*, 1250–1259.

White, H. D. (2003b). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science*, *54*, 423–434.

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, *32*, 163–171.

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*, 327–355.

Zhao, D. (2003). *A comparative citation analysis study of web-based and print journal-based scholarly communication in the XML research field*. Dissertation, Florida State University. http://etd.lib.fsu.edu/theses/available/etd-09232003-012028/unrestricted/DangzhiZhao_dissertation_summer03.pdf Accessed 20.02.08.

Zhao, D. (2006). Towards all-author co-citation analysis. *Information Processing & Management*, *42*, 1578–1591.

Zhao, D., & Strotmann, A. (2008). Information Science during the first decade of the Web: An enriched author co-citation analysis. *Journal of the American Society for Information Science and Technology*, *59*(6), 916–937.