This is a preprint of an article published in *Proceedings of the American Society for Information Science and Technology 2007 Annual Meeting*, October 19 - 24, 2007, Milwaukee, Wisconsin, USA

All-author vs. first-author co-citation analysis of the Information Science field using Scopus

Dangzhi Zhao
School of Library and Information Studies, University of Alberta, Edmonton, AB, Canada T6G 2J4, Email: dzhao@ualberta.ca.

Andreas Strotmann
School of Business, University of Alberta, Edmonton, AB, Canada T6G 2J4, Email: andreas.strotmann@ualberta.ca.

Abstract

Although many studies on the various ways of allocating credit among co-authors have brought into general recognition that different citation counting methods can result in quite different author rankings, studies on different author co-citation counting methods are still largely missing. This paper examines whether different co-citation counting methods produce different results in author co-citation analysis studies of the intellectual structure of research fields, and if so, in what ways they differ. Our results indicate that, with respect to the major specialties and how they relate to each other, the intellectual structures of the Information Science field identified through author co-citation analyses based on different co-citation counting methods are largely equivalent, but when it comes to detailed structure, results differ in a number of ways. In particular, classic first-author co-citation analysis appears to better represent the theoretical and methodological aspects of the field whereas all-author co-citation analysis favors more recent empirical studies, and picks out some tightly collaborative research groups or projects. We experiment with using meaningful diagonal values in a co-citation matrix rather than using statistically generated values, and observe favorable results. We also employ a new visualization technique for reporting the results of a classic author co-citation analysis, using a bipartite graph that represents all the information in the factor matrix of specialties and author loadings as author and factor vertices connected by edges with loadings as similarity-measure line values, laid out algorithmically in two dimensions.

1. Introduction

Since its introduction by White & Griffith (1981), author co-citation analysis (ACA) has gained great popularity among those who study the intellectual structures of scholarly fields and the social structures of the corresponding communities that these imply. While most studies to date have applied the general steps and techniques of classic ACA to different research fields with little or no modification, some studies have proposed new techniques for mapping author clusters (White, 2003), different statistical methods for processing co-citation counts (Ahlgren, Jarneving & Rousseau, 2003), or replacements for classic statistical procedures of ACA such as MDS (Leydesdorff & Vaugh, 2006). However, few studies have examined the way that the co-citation counts themselves are defined – one of the fundamental aspects of ACA which provides the raw data on which all subsequent statistical analyses and mappings in ACA are based.

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 co-citation analyses.

2. The problem and research questions

Classic ACA only takes into account first authors in the definition of author co-citation. Specifically, two authors are considered as being co-cited when at least one document from each author's oeuvre occurs in the same reference list, an author's oeuvre being defined as all the works with the author as the first author (McCain, 1990). We have termed this "first-author co-citation".

This definition has rarely been challenged, partly due to the constraints imposed by the main data source that has so far been used for ACA studies – the set of databases developed by the Institute for Scientific Information (ISI), now Thompson Scientific. These databases generally only index the first authors of cited documents. The full set of authors of a cited document can only be found in these databases through matching it with the corresponding source paper if that document also happens to have been indexed as a source paper (citing paper).

As a result, it has been very difficulty if not impossible to go beyond first-author co-citation counting using these databases, especially with the method used for obtaining author co-citation counts in classic ACA.

A study by Persson (2001) is the only one we are aware of that attempts to compare first-author and all-author co-citation analysis. It took all the 7001 citing articles in the LIS field identified in one of the ISI databases between 1986 and 1996, and looked at how these articles had been co-cited by each other, thereby disregarding the more than 90% of references to papers not indexed as source papers by the ISI databases. Persson's brief article provided two multi-dimensional scaling (MDS) maps based on first-author and all-author co-citation counts respectively, but little detail on how co-citation counts were defined and calculated. It observed similar subfield structures on the two maps, and noted the fact that many well-known authors on the all-author co-citation map were excluded from the map based on first-author co-citation counts.

The evolving scholarly communication system has recently made available alternatives to the ISI databases for collecting co-citation data that allow us to go beyond first-author co-citation towards all-author co-citation. We recently took advantage of one of these new databases, namely *CiteSeer*, and conducted a preliminary study that compared first-author and all-author co-citation analyses of the XML research field, a subfield of computer science (Zhao, 2006).

That study found that an all-author co-citation analysis, which takes into account more links between related authors, resulted in a clearer picture of the intellectual structure of the XML research field than the classic first-author co-citation analysis did, but also that fewer specialties were identified if the same number of most highly cited authors was selected based on complete citation counts.

Our preliminary study also introduced two different possible definitions of all-author co-citation, and compared the ACA results from the corresponding methods of determining author co-citation counts.

(a) **Inclusive all-author co-citation** is defined by simply modifying the definition of first-author co-citation by redefining an author's oeuvre as all the works with the author as *one* rather than as the *first* of the authors of each of these works. With this definition, two authors are also considered co-cited when a paper they co-authored is cited. This in fact includes cited co-authorship in co-citation.

(b) **Exclusive all-author co-citation** is defined just like inclusive all-author co-citation except for excluding cited co-authorship from co-citation counts.

The result from our preliminary comparison was in favor of exclusive co-citation counts in the study of intellectual structures, but it also found that inclusive co-citation analysis shows some promise in the study of social relationships in ACA.

The present study will explore the extent to which the patterns that we found in our preliminary study on a computer science field may be found again in a different research field, namely, in the Information Science field, using a different citation index as the data source, namely Elsevier's Scopus. Unlike the previous preliminary study which used a simplified approach to all-author co-citation counting in that at most five authors of a cited reference were taken into account, the present study strictly counts all authors of cited references to eliminate any drawbacks that might exist in the simplified approach. With more than 3000 citing papers, the scale of the present study is also considerably larger than the preliminary study which analyzed just a little more than 300 papers.

3. Methodology

Data collection

The research area we analyze in the present study is that of Information Science. As in White & McCain (1998), we define the IS research field by the 12 journals listed in Table 1 in White & McCain (1998, p. 330). Our citation window is a ten-year period following that studied in White & McCain (1998), i.e., 1996 – 2005.

We used Elsevier's Scopus database to retrieve citing papers in the IS field thus defined, along with their reference lists. Scopus currently indexes journal articles published in 1996 or later. It is similar to the ISI

databases 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 databases do, including their full titles and the names of up to eight authors (the first seven and the last author) for each cited paper.

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 cited references that have more than 8 authors, we manually added the authors that were missing in the files we downloaded from Scopus, by searching for these papers as citing papers in Scopus or in other data sources. This is only feasible for research fields like IS in which large-group collaboration is not the mainstream, as indicated by the small average number of authors per paper (i.e., less than 2), but it could be considerably more difficult in other fields.

We developed a Java program to parse these records, and to store the resulting data fields such as authors, publishing sources and years of both source papers and cited references in a data structure that was convenient for later data analysis such as counting citations and co-citations. The data structure and algorithms of this program are not reported here.

We noticed that the number of papers retrieved from Scopus is smaller than that from the ISI databases across the 12 journals. For example, as publications in 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. In order to ascertain how these two data sources are different as data sources for ACA studies, we will need to examine in detail what articles are indexed in one but not in another, but that will have to wait for another study. In the meantime, we performed a comparison (not reported here in any detail) between the classic first-author-based ACA results using Scopus and Web of Science data, and found that the specialty structures revealed from these two data sources are very similar. We are thus 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.

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; White, 2003; 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.

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 & McCain (1998) as adequate for the purpose of this type of studies.

A Python program was developed to count co-citation frequencies by the three methods discussed above, and to record them in three separate matrixes. These co-citation matrixes were then used as input to the Factor Analysis (FA) procedure in SPSS. As usual, the diagonal values in these matrixes 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 Python program, the code for counting exclusive all-author co-citations is identical for all cells of the matrix, including the diagonal. 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 et al., 1998). The number of factors extracted was determined based on Kaiser's rule of eigenvalue greater than 1 because the resulting model fit was adequate 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 1 along with their model fits.

Table 1: 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%)

* Percentage of non-redundant residuals with absolute values greater than 0.05 where residuals are computed between observed and reproduced correlations.

These numbers 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.

The inter-factor correlations matrixes produced by the oblique rotation are not presented here due to space limitations, but they will be referred to in the discussion where necessary.

Visualization of factor structures

We introduce here a novel way of reporting and visualizing ACA results which allows us to circumvent the space limitations that traditionally would make it impossible to report the results of this kind of study here.

Unlike previous ACA studies, which have presented ACA factor structures in large tables and cluster structures on Multi-Dimensional Scaling (MDS) maps (McCain, 1990; White & McCain, 1998), the factor

structures are visualized here as two-dimensional graphs as shown in Figures 1 to 4. To prepare these figures, the factor labels were assigned, as usual, upon examining the frequently cited articles written by authors in those factors that they load highly on. In these visualizations, authors are represented by square nodes with labels in a smaller, regular font, and factors are represented by circular nodes with labels in a larger, italic font. The thickness of a line that connects an author with a factor is proportional to the value of the loading of this author on this factor, as is its grayscale value. Throughout, only sufficiently high loadings of 0.3 or higher are considered. In this way, the map preserves all the relevant information that is usually contained in a large table with author loadings on specialty factors.

The layout of our maps is produced using Pajek's implementation of the Kamada-Kawai graph layout algorithm, using loadings as similarity measure between author and factor nodes. Since this algorithm sometimes gets stuck in a local optimum, we produced several layouts for each matrix and picked one for presentation here. The result is a map of the field that is visually informative and true to the factorization it represents, even if it is not equivalent to the traditional ACA's MDS author cluster map.

The sizes and colors of nodes in these maps carry auxiliary information. An author node is proportional in size to the author's total loading on all factors combined, and the size of a circle node representing a factor corresponds to the sum of the loadings on this factor by all authors who load sufficiently on it. In this way, we try to use the size of a node as a visual indicator of its overall significance in the map. In addition, node colors in this map represent the number of links to or from a node to other nodes in the map (i.e., its degree). The color of factor circles therefore represents the number of authors that load sufficiently on it, while the color of a square node representing an author indicates the number of factors that this author loads on with a value of at least 0.3 each: yellow for authors who only load sufficiently on a single factor, green for those who co-load on two factors, and red, for those who co-load on three factors.

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 (Figures 1) so that we can compare it with the picture revealed in White & McCain (1998) to discuss the development of the IS field over the past ten years that our study identifies. 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 (Figure 2), those from an inclusive all-author co-citation analysis (Figure 3), and those from an exclusive all-author co-citation analysis with missing diagonal values (Figure 4).

Intellectual structure of the IS field

Ten years on from White & McCain's classic (1998) study, which performed a factor analysis of the 120 most highly cited authors in IS based on a first-author co-citation analysis of articles published in 12 journals chosen to represent the field during the years up to 1995, 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. The four specialties that were most active in the past ten years are Experimental retrieval, Information behavior, Scientometrics/citation analysis, and Webometrics. More than half (56%) of the authors analyzed represent 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, Situational Relevance, Interface design, Knowledge Management, and Research methods.

Comparing with the situation of the field 10 years earlier as described in White & 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 & McCain (1998) has split into two groups: IS theories (e.g., Shannon) and Interface design (e.g., Bush, Shneiderman), both joined by additional authors.

In addition to Webometrics, two other new specialties have emerged: Situational relevance, which studies users' criteria for 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.

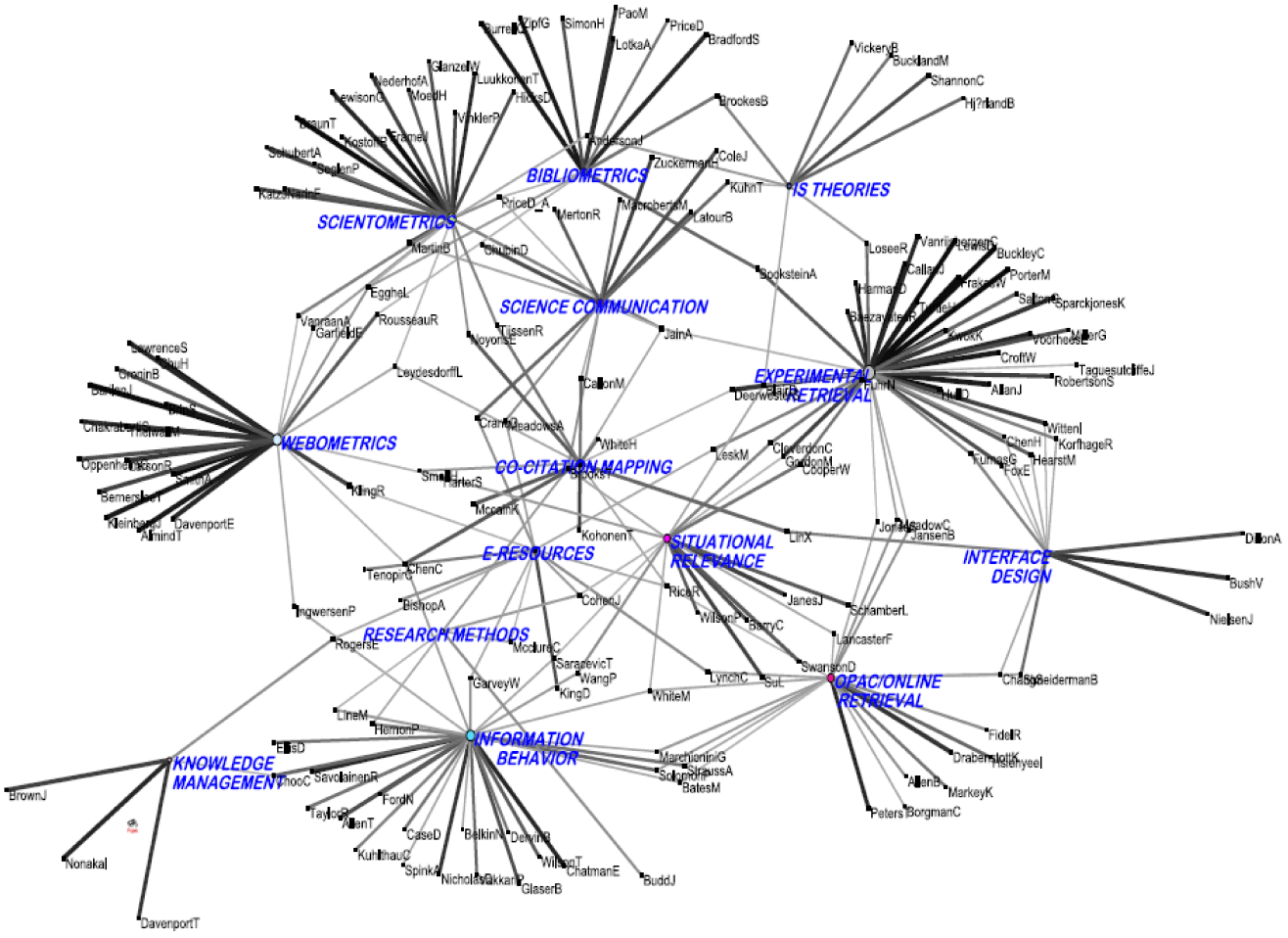


Figure 1: Factor analysis results from first-author co-citation counts

As for the interrelationships between specialties, indicated by the inter-factor correlations matrix produced by the oblique rotation, the largest specialty Experimental retrieval is closely related to Interface design with an

inter-factor correlation of 0.41, as well as with Situational relevance (0.29); the specialty Scientometrics/citation analysis is closely related to Science communication (-0.44), Bibliometrics (-0.37) and Co-citation mapping (0.30), as one would expect; the specialty Information behavior is closely related to Situational relevance (0.35) and Knowledge management (0.33); and the specialty Webometrics is not closely related to any other specialties. Other significant correlations (0.3 or higher) include those of OPACs/online retrieval with Situational relevance (-0.46), Interface design (0.43), Information behavior (0.41) and IS theories (0.32), and that of Situational relevance with IS theories (-0.38).

Classic first-author-based vs. 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 (Figure 2). The four specialties identified in this analysis that were most active in the past ten years are the same as those from first-author-based ACA as shown in Figure 1. In this analysis, about 60% of the authors analyzed 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, Situational Relevance, and Knowledge Management specialties.

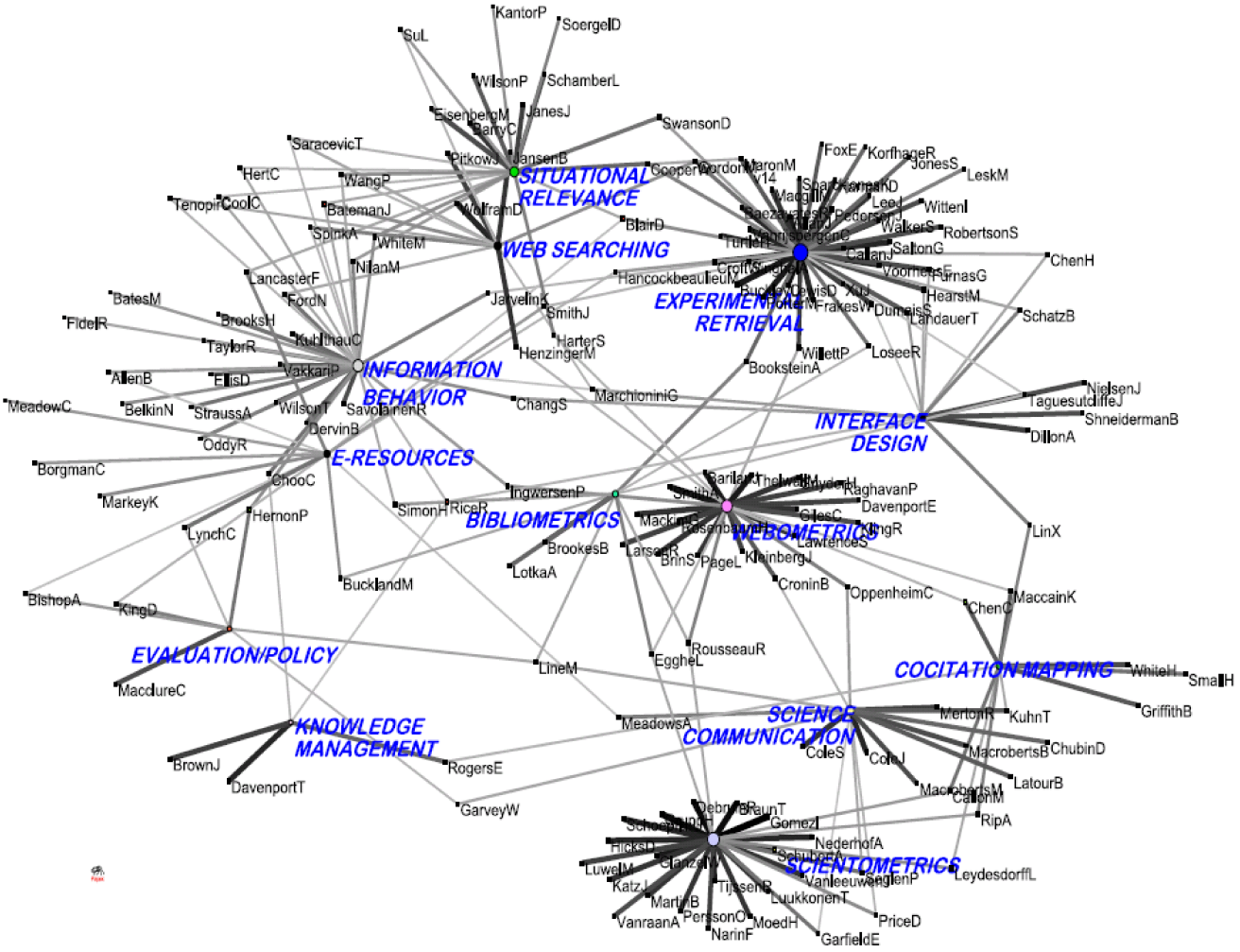


Figure 2: Factor analysis results from exclusive all-author co-citation counts (with meaningful diagonal values)

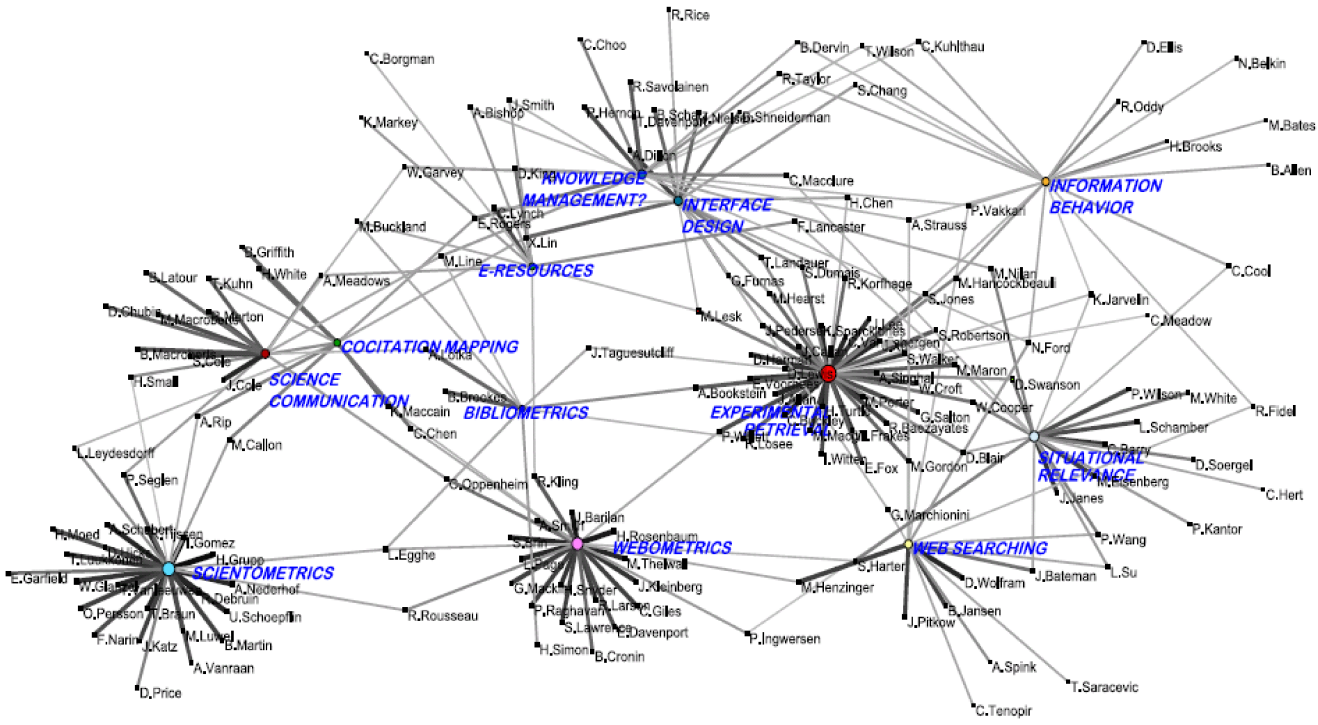
The IS theories group and the small and vague group on Research methods identified in the results from first-author-based ACA do not show up here in the results from exclusive all-author-based ACA. This indicates that first-author-based ACA may better represent the theoretical and methodological aspects 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 OPAC/online/ Web searching specialty seen in the results from first-author-based ACA has switched focus in the results from exclusive all-author-based ACA, from OPACs/online retrieval systems to Web search engines, resulting in the distinct Web searching specialty. This suggests that first-author-based ACA represents older research better whereas all-author-based ACA reflects current studies more clearly.

An examination of the highly cited publications by authors in the Web searching specialty in the results from all-author-based ACA 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.

The interrelationships between specialties in Figure 2 are similar to those in Figure 1 in terms of common specialties, except for a closer relationship between Information behavior and E-resources (-0.29 vs. -0.16), and lower correlations between Experimental retrieval and Situational relevance (0.24 vs. 0.29), and between Scientometrics/citation analysis and Bibliometrics (-0.11 vs. -0.37).

Inclusive all-author-based ACA



Figures 3: Factor analysis results from inclusive all-author co-citation counts

The results from an inclusive all-author-based ACA as shown in Figure 3 are very similar to those from exclusive all-author-based ACA with meaningful diagonal values as shown in Figure 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 Figure 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 the widest 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.

Exclusive all-author-based ACA with diagonal values deleted

The results from an exclusive all-author-based ACA with diagonal values deleted (Figure 4) are very similar to those from inclusive all-author-based ACA shown in Figure 3, except that 11 rather than 12 specialties are identified, with the Information behavior specialty splitting into two groups which merged into the Situational relevance specialty (BrooksH, CoolC, JarvelinK, VakkariP, OddyR, KuhlthauC, BelkinN) and into the Web searching specialty (AllenB, EllisD, BatesM), respectively. The weird 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, BrownJ), Evaluation/Policy (Hermon.P, McClure.C), Information behavior (Savolainen.R, Choo.C, Taylor.R, Dervin.B, WilsonT), and Science communication (Garvey.W).

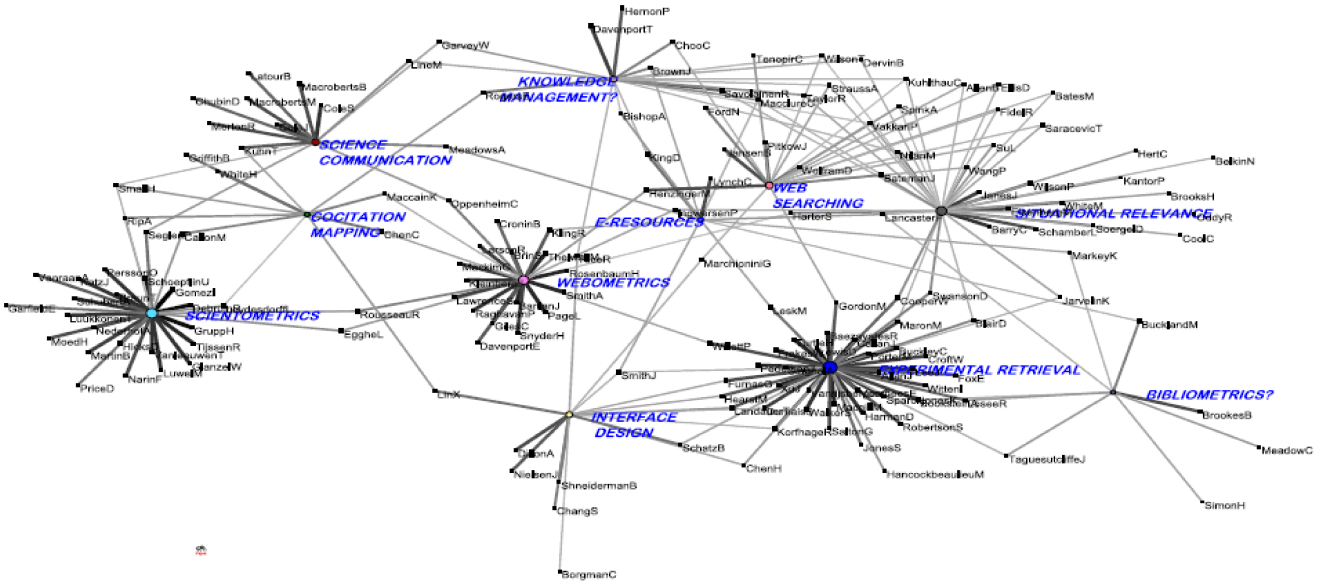


Figure 4: Factor analysis results from exclusive all-author co-citation counts (diagonal values deleted)

Comparing results from the two types of exclusive all-author-based ACA – that with meaningful diagonal values (Figure 2) vs. that without diagonal values (Figure 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 merged into the Knowledge management specialty; the Information behavior specialty in Figure 2 disappears into three other specialties: Situational 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 present study has found that the major specialty structures of the Information Science field identified from first-author-based ACA and from all-author-based ACA are largely the same. First-author-based ACA appears to have a slight advantage representing the theoretical and methodological aspects of the field, and all-author-based ACA may represent current trends (e.g., Web searching) more clearly, and appears to pick out some highly collaborative research groups or projects.

In our preliminary study of the XML field, it was observed that all-author-based ACA tended to represent a smaller number of specialties than first-author-based ACA when the same number of authors was selected by citedness. This is also seen here in the IS field but not as clearly: exclusive all-author-based ACA with missing diagonal values identified 11 rather than 14 specialties, the one with meaningful diagonal values found 13, and the inclusive all-author-based ACA identified 12 specialties.

Unless the mixed group tentatively labeled Knowledge management makes good sense, the classic first-author-based ACA results in a picture of the intellectual structure of the IS research field that is clearer than the inclusive all-author co-citation analysis and the exclusive all-author-based ACA with diagonal values deleted, and it is comparable to that revealed from the exclusive all-author co-citation analysis with meaningful diagonal values.

At first blush, this appears to contradict our earlier observation in the XML field that all-author-based ACA tends to produce a clearer picture of the intellectual structure of a research field, and that it is therefore important to count all authors of cited papers in co-citation analysis studies.

However, if we take into account the differences in collaboration levels between the two fields that we have studied, we can offer an alternative interpretation of these results. In the IS data set we used, citing papers had less than two authors on average, and the average number of authors of all cited references in that data set is about 1.5. The latter number is not that much different from the average number of first authors of all cited references, which is 1.0. In the XML dataset, by contrast, we find about three authors per citing paper on average, and about 2.5 when we average across all cited references instead of citing papers. Unlike in the IS field, the average number of cited authors per reference for the XML field is significantly larger than 1.0, so that we would expect to see a much larger difference (by a factor of three, in fact) between all-author and first-author analyses in the XML field than in the IS field.

With hindsight, we may thus interpret our results as a retrospective validation of the classic ACA's simplification of the author co-citation analysis methodology to that of counting first authors only, at least in the Information Science field that has often been studied using an ACA approach. The results we obtained from an analysis of this field which does take into account all authors of cited references and does so in a thoroughly consistent fashion, are very similar to the results of an analysis that uses the classic first-author-only ACA methodology. We suspect, however, that it is the low level of collaboration in this field that has contributed significantly to this high similarity, and that extending ACA to the study of highly collaborative "big science" areas does indeed require moving into all-author co-citation analysis.

Conversely, we have also identified in this paper a good candidate for valid co-citation analyses that do take into account all authors of cited papers. The analysis that is based on what we termed in this paper an exclusive all-author co-citation matrix with meaningful diagonal values produced a multivariate factorization that closely matches the one produced by classical first-author co-citation analysis. We have thus verified that this alternative method works just as well as the classical method in a situation where similar results are to be expected from both.

Further study is however needed in order to clarify findings from our two studies and to test if, in an increasingly collaborative academic environment, citation and co-citation analyses need to switch from

first-author to all-author citation counting methods if they are to be true to their goal of providing realistic scholarly communication indicators. A research field from the bio-medical area, in which large-group collaboration are even more prevalent than in computer science, should be a good choice for such a follow-up study.

6. Conclusion

In this study, we find from several ACA analyses that the IS field has remained remarkably stable compared to an equivalent study of this field a decade earlier (White & McCain, 1998), but we also find 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, if 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 (e.g., Web searching in the case of the IS field). 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 requires both types of ACA.

The present study also shows that, for exclusive all-author co-citation count matrices, computing meaningful diagonal values appears to produce 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 field in which large-group collaboration is commonplace such as in the bio-medical area should be a good choice as the research field to be studied using an ACA approach for this purpose.

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

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

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

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.

White, H. D. (2003). 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.

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. Retrieved Jan. 20, 2005, from http://etd.lib.fsu.edu/theses/available/etd-09232003-012028/unrestricted/DangzhiZhao_dissertation_summer03.pdf

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