Using the Appearance of Citations in Full Text on Author Co-citation Analysis

Yi Bu[1], Binglu Wang[2], Win-bin Huang[2,*], Shangkun Che[2], Yong Huang[3]

[1] *School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN., U.S.A.*

[2] *Department of Information Management, Peking University, Beijing, China*

[3] *Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, Wuhan, Hubei, China*

Corresponding author: Win-bin Huang (Email: huangwb@pku.edu.cn; Address: No.5 Yiheyuan Road, Haidian District, Beijing 100871, P. R. China).

Abstract: As a frequently used method of depicting scientific intellectual structures, author co-citation analysis (ACA) has been applied to many domains. However, only count-based information is involved as the input of ACA, which is not sufficiently informative for knowledge representations. This article catches several metadata in full text of citing papers but not aims at content-level information, which increases the amount of information input to ACA without increasing computational complexity a lot. We propose a new method by involving information including the number of mentioned times in a citing paper and the number of context words in a citing sentence. We combine these pieces of information into the traditional ACA and compare the results between ACA and the proposed approach by using factor analysis, network analysis, and MDS-measurement. The result of our empirical study indicates that compared with the traditional ACA, the proposed method shows a better clustering performance in visualizations and reveals more details in displaying intellectual structures.

Keywords: Author co-citation analysis, co-citation analysis, citation analysis, bibliometrics, scientometrics, mapping knowledge domains

Mathematics Subject Classification: 68T30

JEL Classification: D830

INTRODUCTION

Author co-citation analysis (ACA) is a bibliometric method in knowledge representation and has shown a good performance in depicting scientific intellectual structures and mapping knowledge domains (White & Griffith, 1981; McCain, 1990; Jeong, Song, & Ding, 2014). More than three decades since its born, ACA has been applied to many disciplines, such as library and information science (White & Griffith, 1998; Ding, Chowdhury, & Foo, 2001; Ding, 2011a; Zhao & Strotmann, 2014), cognitive science (Bruer, 2005), management science (Eom, 1999; Chen & Lien, 2011; Zhao, Zhang, & Kwon, 2017), and medical science (Chu, Liu, & Tsai, 2012).

Traditional ACA regards two authors with higher co-citation frequency as higher topical relatedness. Such assumption hints that every author pair with the same co-citation frequency as identical, which simply considers count-based instead of content-based information. As the availability of full-text data nowadays, Jeong *et al.* (2014) firstly proposed content-based ACA method and compared the similarity between citing sentences. However, the computational cost of their content-based method could be high

because of the processing of words as well as similarity calculation between citing sentences. Actually, with the full-text data we do not have to employ the content-level information; instead, several pieces of useful information at metadata level that were ignored previously in full text can be considered to improve the performance of ACA in mapping knowledge domains.

The number of mentioned times of references, for instance, is a typical piece of information. As pointed out by Ding, Liu, Guo, and Cronin (2013) as well as Zhao, Cappello, and Johnston (2017), the number of mentioned times of a reference represents the importance of the reference to the citing paper. However, the traditional ACA regards as identical two co-cited authors with a distinct number of mentioned times in a citing paper, which is problematic. For example, Zhao and Strotmann (2014) cited (a) Zhao and Strotmann (2008a), (b) White and McCain (1998), and (c) Hirsch (2005), but (a) was mentioned 15 times, (b) twice, while (c) only once in the citing paper. The co-citation strength of pair (a)-(b) and (b)-(c) should be different when we consider their topical relatedness. In this paper, we start to consider the number of mentioned times of references as a supplement into ACA in order to provide more accurate information for mapping knowledge domains (Bu et al., 2017a).

Besides, the citing sentences containing references could have different numbers of words. From our intuitive thinking, a reference contained in a longer citing sentence should have more topical relatedness to the citing paper than that contained in a shorter one, because longer sentences are more likely to include more details or interpretations to the reference, which is more "useful" to the citing paper—otherwise it is not likely to be cited with many interpretative words. However, the traditional ACA ignores such difference in the length of citing sentences; as a result, we start to consider the difference in the length of citing sentence and combine it into ACA in this paper.

The current paper considers the number of mentioned times and the number of context words into ACA. The results of the empirical studies show that our newly proposed approach not only shows better clustering performance but also provides more details in knowledge domain mappings. As talked in the aforementioned paragraph, the current approach adopts full-text data without using content- or semantic information. This article is outlined as follows. At first, the work related to our study and the data with the methods for our analysis are detailed. The findings as well as the comparisons between the traditional ACA and our proposed method then are presented. Finally, the conclusions and the future research are pointed out.

RELATED STUDIES

Author co-citation analysis (ACA) was proposed by White and Griffith (1981). In 1990, McCain gave a completed overview and set up a standard framework for ACA, in which four steps of ACA implementations were mentioned: (1) Data collection and processing; (2) Construction of raw co-citation matrix; (3) Transformation to correlation matrix; and (4) Data analyses (e.g., factor analysis, clustering analysis, multi-dimensional scaling (MDS) analysis, and network analysis) and result interpretations.

More than thirty years, this method has been improved a lot by revising rules to construct raw co-citation matrix (Step (2) above) and transform to correlation matrix (Step (3) above). As pointed out by Persson (2001), the elements in co-citation matrix can be defined as first-author or all-author co-citation frequency; the latter might, to some extent, provide more detailed knowledge domain maps (Eom, 2008a; Rousseau & Zuccala, 2004; Zhao, 2006). As the availability of the all authors' information, it is more common to use all authors' information instead of first authors' to run ACA. Meanwhile, the rules of defining main diagonal values was also discussed and at least six distinct ways of processing main diagonal values in raw co-citation matrix have been proposed and/or experimented (Eom, 2008b).

Additionally, several metadata of references, such as published time and venue of references and their keywords, were considered in ACA implementations (Bu *et al.*, 2016), and they were found to play positive roles in improving the performance of ACA maps. Note that the metadata they employed had been obtained from reference lists instead of full text.

In the Step (3) above, we have also observed that many researchers have worked on the strategies of transforming correlation matrix (Ahlgren, Jarneving, & Rousseau, 2003; White, 2004). The usage of Pearson's *r* and other correlation measurements has been debated in both theoretical and mathematical ways (Mêgnigbêto, 2013). Although there is not any common-believed final conclusions about which measurement should be implemented in ACA, we here follow Ahlgren *et al.* (2003)'s arguments to use cosine similarity to transform the matrices in this work.

Due to the availability of full-text scientific data, Jeong *et al.* (2014) first explored content-based ACA by comparing the similarity between citing sentences. Their empirical studies show that content-based ACA is able to mine more details in scientific intellectual depicting compared with the traditional ACA. Nevertheless, the computational complexity is high in full text processing and semantic identification, which impedes the applications of their proposed method to various domains widely.

When full-text data are used, however, it is not required to make content- or semantic-level analyses. Several non-content-level metadata are easily accessible in full text, such as the number of mentioned times and the number of context words in a citing paper. In addition, these pieces of information reflect the importance of a certain reference to the citing paper. For example, if mentioned many times than others, a reference is likely to have more topical relatedness to the citing paper; if the citing sentence containing certain reference is longer than that containing another reference, we will be more confident to expect it to interpret more details and thus has higher possibilities to relate to the citing paper (Bu et al., 2017a). As a result, this paper combines the number of mentioned times and the number of context words of references into ACA and proposes a new method so as to improve the performance of ACA in knowledge domain mappings.

METHODOLOGY

The whole process of our algorithm is shown in Figure 1. All of the dataset are derived from full-text in *Journal of the American Society for Information Science and Technology* (JASIST, currently named as *Journal of the Association for Information Science and Technology*). After data processing (see details in the "Data" section), we extract the number of mentioned times and the number of context words, and combine them into ACA (dotted area in Figure 1, see details in the "Methods" section). After the new co-citation matrix is constructed, cosine similarity is utilized to transform it to correlation matrix. Factor analysis, network analysis, and MDS-measurement are used to analyze the data, in which Gephi (Bastian, Heymann, & Jacomy, 2009) is applied to display the results of the combined author co-citation network for discussions and analyses. Note that the dotted area in Figure 1 is the major difference among the proposed and traditional ACA methods.

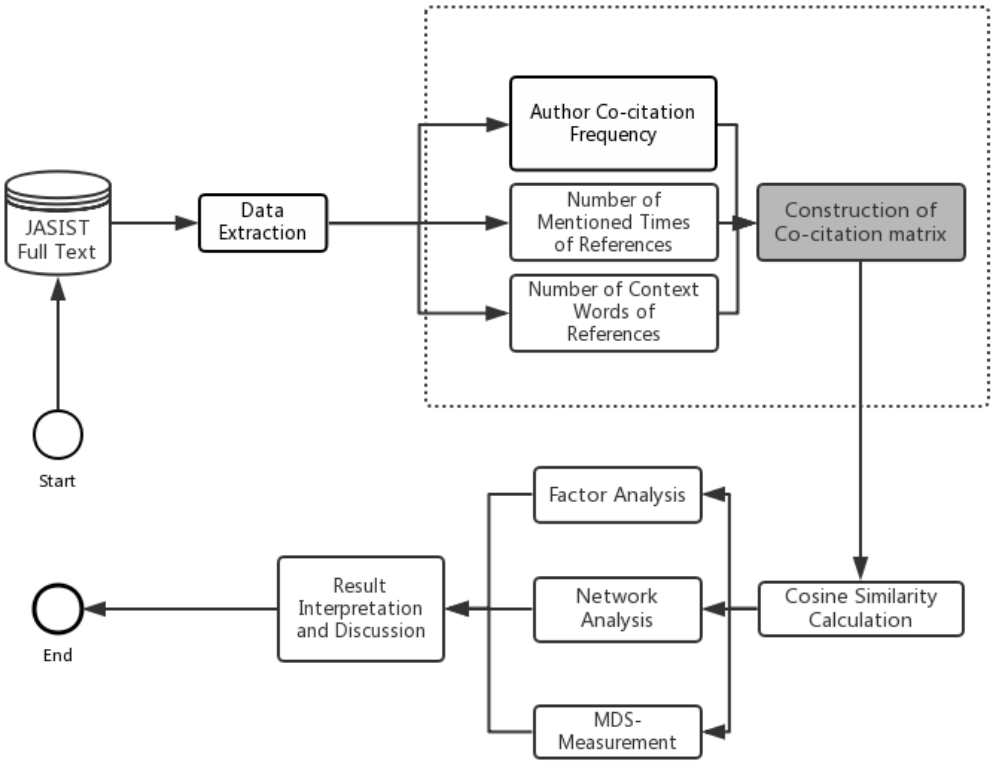

Figure 1. Flow diagram of the proposed algorithm.

Data

The dataset used in this research is the same as that in Jeong *et al.* (2014), in which 1,420 full-text articles with citation links published in JASIST between January 2003 and June 2012 are selected. These 1,420 articles containing 60,068 references are completed by 32,095 authors. In order to make the co-citation matrix denser, we extract the most popular 500 authors who have received the most number of citations; bibliometrically, the most "popular" scholars are often regarded as the representative of the research going back according to Zhao and Strotmann (2008a, 2014).

Methods

Calculation of mentioned time parameters

A paper citing other papers refers that these cited papers (references) are related and useful to the citing paper (CP). Traditionally, these co-cited papers are regarded as equal in co-citation analysis (Ding *et al.*, 2013). However, the importance of cited papers might be distinct (Cano, 1989; Case & Higgins, 2000). Specifically, some references are crucial to the CP because they might be the foundation of CP. In the full-text, the important references can be revealed as multi-mentioned cited papers. For one CP, from an intuitive thinking, the more number of mentioned times a certain reference has, the more importance it is to the CP (Ding *et al.*, 2013; Zhao *et al.*, 2017; Zhu *et al.*, 2015). For instance, Zhao and Strotmann (2014) cited (a) Zhao and Strotmann (2008a), (b) White and McCain (1998), and (c) Hirsch (2005), but (a) was mentioned 15 times, (b) twice, while (c) only once. In this case, (a) should be the most correlative reference to the CP, compared with (b) and (c). Indeed, Zhao and Strotmann (2014) tried to map the knowledge domains of information science (IS) between 2006-2010 while (a) did the same thing but between 1996-2005 with similar methods, ACA and author bibliographic coupling analysis (ABCA), which shows that (a) is closely-related to the CP. However, (b) only used ACA to map the knowledge

domain of IS instead of ABCA and the result of (b) is also very different from (a) and the CP. The reason (c) is cited is simply because the indicator "h-index" was used and mentioned. Hence, we can see that the number of mentioned times of a reference is indeed positively related to its relatedness with CP.

Similarly, in ACA, if two co-cited authors are both mentioned many times in a CP, their topical relatedness tends to be high because both of them are closely related to the CP. In the above example, the topical relatedness of (a)-(b) should be higher than that of (b)-(c). Indeed, (a) and (b) both focused on mapping the field of IS by employing ACA, but (c) simply introduced an indicator to evaluate scholars. Thus, in our proposed algorithm, we assume that two co-cited authors with more number of mentioned times should be assigned more weights in co-citation analysis because they have higher possibilities to be related with each other topically.

Mathematically, suppose that the authors A_i and A_j are co-cited for x_{ij} times. Specifically, the CP is annotated as P_{ij_1}, P_{ij_2}, …, $P_{ij\_x_{ij}}$. In P_{ij_k} $(k = 1,2, \dots, x_{ij})$, assume that the author A_i is mentioned for λ_{ik} times and the author A_j is mentioned for λ_{jk} times. If we annotate the mentioned time of the cited author with the maximum number of mentions in the paper P_{ij_k} as λ_{k_max}, the mentioned time parameter between the authors A_i and A_j in the paper P_{ij_k}, MT_{ij_k}, is calculated as:

$$MT_{ij\_k} = \frac{\lambda_{ik}\lambda_{jk}}{\lambda_{k\_max}^2} \quad \text{(Eq. 1)}$$

If we consider all of the MT_{ij_k} in citing papers P_{ij_1}, P_{ij_2}, …, and $P_{ij\_x_{ij}}$, the mentioned time parameter between the authors A_i and A_j among dataset, MT_{ij}, could be defined as:

$$MT_{ij} = \sum_{k=1}^{x_{ij}} MT_{ij\_k} \quad \text{(Eq. 2)}$$

Calculation of context word parameters

When citing references, CPs tend to use one or more sentences to set up an argument, which is called citing sentences (or "citance" proposed by Nakov, Schwartz, and Hearst, 2004) (Jeong *et al.*, 2014). However, the lengths of citing sentences are probably different. For example, Zhao and Strotmann (2014) cited: (a) Finlay, Sugimoto, Li, and Russell (2012), (b) Milojević, Sugimoto, Yan, and Ding (2011), as well as (c) Sugimoto, Li, Russell, Finlay, and Ding (2011) in the same sentence with 27 words. They shared these 27 words and each of them has been assigned nine words averagely. Meanwhile, Zhao and Strotmann (2014) also cited (d) Zhao and Strotmann (2008b) in a sentence with 31 words. Although all of these references are cited once in the CP, their numbers of context words assigned are diverse, 9, 9, 9, and 31, respectively.

Basically the number of context words assigned in a CP reflects the importance of the reference. Specifically, more numbers of context words assigned in a CP reveal that more details and interpretations of the reference is likely to be stated, which hints that it has higher topical relatedness to the CP. For example, CP uses ACA and ABCA to map the knowledge domain of IS field, while (d) explores AACA at a methodology level; both of them are closely related to the ACA research. Nevertheless, (a) analyzed Library Science (LS) using titles and keywords, (b) employed article title words to depict the scientific structure of LIS, and (c) focused on North American LIS dissertation using Latent Dirichlet Allocation (LDA) model, all of which are not as close as (d) in terms of the topical relatedness with the CP from an intuitive perspective. Indeed, the number of context words assigned to (d) is much more than that to (a), (b), and (c). These show that the number of context words assigned in a CP is positively related to its topical relatedness to the CP. Similarly in ACA, if two co-cited authors are both assigned many words than others, their topical relatedness should be higher because both of them are closely related to CP.

To test this assumptions, we randomly select 150 citing sentences from all citing sentences in our

corpus. Then two bibliometricians manually labeled the importance of the references containing in the sentences to the raw papers; specifically, they labeled the importance as "important", "neutral", and "unimportant". In total, their labels on 138 sentences out of 150 (92%) are consistent; we therefore target on these 138 citing sentences. We then implement a mean-based T test and found a significant difference in regard to the length of the citing sentences (i.e., the number of context words in the sentences) among the three groups.

In P_{ij_k}, assume that during its μth mention ($\mu = 1,2,\dots,\lambda_{ik}$), the citing sentence containing the author A_i (A_j) includes $w_{ik\mu}$ ($w_{jk\mu}$) words and mentions $a_{ik\mu}$ ($a_{jk\mu}$) distinct authors ($w_{ik\mu}, w_{jk\mu}, a_{ik\mu},, a_{jk\mu} > 0$). The number of context words of the author A_i in the paper P_{ij_k}, cw_{i_k}, can be calculated as (similar to A_j):

$$cw_{i\_k} = \sum_{\mu=1}^{\lambda_{ik}} \frac{w_{ik\mu}}{a_{ik\mu}} \quad \text{(Eq. 3)}$$

If we annotate the largest number of context words of cited author in the paper P_{ij_k} as cw_{k_max}, the context word parameter between the authors A_i and A_j in the paper P_{ij_k}, CW_{ij_k}, is calculated as:

$$CW_{ij\_k} = \frac{cw_{i\_k} cw_{j\_k}}{cw_{k\_max}{}^2} \quad \text{(Eq. 4)}$$

If we consider all of the CW_{ij_k} in citing papers P_{ij_1}, P_{ij_2}, …, and $P_{ij\_x_{ij}}$, the context word parameter between the authors A_i and A_j among dataset, CW_{ij}, could be defined as:

$$CW_{ij} = \frac{1}{x_{ij}} \sum_{k=1}^{x_{ij}} CW_{ij\_k} \quad \text{(Eq. 5)}$$

Construction of the co-citation matrix based on the two parameters

The co-citation matrix in our proposed algorithm is based on the above parameters to be normalized into [0,1]. We here annotate the largest co-citation frequency among the dataset regardless of which author pairs as x_{max}, and the weight values for co-citation, mentioned time, and context word parameters as w_c, w_{MT}, and w_{CW}, respectively. To better compare different parameters, we run four different models based on these parameters. In Model 0, we simply employ the traditional ACA without importing any other factors. The co-citation matrix in Model 0, $M_0 = (m_{1,i,j})$, as:

$$m_{0,i,j} = \frac{x_{ij}}{x_{max}} \quad \text{(Eq. 6)}$$

In Model 1, we combine the raw co-citation matrix with the mentioned time parameter; we construct the new co-citation matrix in Model 1, $M_1 = (m_{1,i,j})$, as:

$$m_{1,i,j} = w_c \cdot \frac{x_{ij}}{x_{max}} + w_{MT} \cdot MT_{ij} \quad \text{(Eq. 7)}$$

where $w_c + w_{MT} = 1.0$. In Model 2, we combine the raw co-citation matrix with the context word parameter; thus the new matrix in Model 2, $M_2 = (m_{2,i,j})$, as:

$$m_{2,i,j} = w_c \cdot \frac{x_{ij}}{x_{max}} + w_{CW} \cdot CW_{ij} \quad \text{(Eq. 8)}$$

where $w_c + w_{CW} = 1.0$. In Model 3, we combine the raw co-citation matrix with both of the two parameters; thus the new matrix in Model 3, $M_3 = (m_{3,i,j})$, as:

$$m_{3,i,j} = w_c \cdot \frac{x_{ij}}{x_{max}} + w_{MT} \cdot MT_{ij} + w_{CW} \cdot CW_{ij} \quad \text{(Eq. 9)}$$

where $w_c + w_{MT} + w_{CW} = 1.0$. Essentially the Model 0 acts as a baseline for comparing with other models.

RESULTS AND DISCUSSION

Factor Analysis

In factor analysis, we extract the factors whose Eigen Factor is 1.0 or more as the result of factor analysis, regardless in Models 0-3. We have known that the model becomes more complex and involves more information from Model 0 to Model 3. As shown in Table 1, the number of factors extracted from Model 3 is 16, which is five more than that in the traditional ACA (i.e., Model 0). Models 1 and 2 are found to extract 13 factors in the experiments. In terms of the total variance explained, the factor analysis of Model 0 is able to explain about 82.8% of total variance while that of Model 3 explains 85.6%.

Table 1. Factor analysis results overview.

Methods	*Number of factors extracted*	*Total variance explained*
Model 0	11	0.828
Model 1	13	0.844
Model 2	13	0.839
Model 3	16	0.856

Note: In Models 1 and 2 $w_c = 0.6$ and Model 3 $w_c = 0.6, w_{MT} = 0.2, w_{CW} = 0.2$, which are finally determined after examining lots of possible experiments. The same below.

Based on some previous research (Janssens, Leta, Glänzel, & Moor, 2006; Yang, Han, Wolfram, & Zhao, 2016), several core sub-fields of information science are dug out and more details are supposed to be refined. Table 2 shows the factor analysis results of all four models. Core sub-fields of Library and Information Science are extracted and identified by all methods (models), including: (1) information retrieval, (2) information seeking behavior, (3) language model, query, and clustering, (4) text mining, machine learning, (5) user interface, (6) evaluation indicator, index, (7) webometrics, social network analysis, (8) scholarly communication, (9) journal citation analysis, interdisciplinarity, evaluation of algorithms, (10) network analysis, and (11) bioinformatics. Although bioinformatics is not the main scope of JASIST, there is still one author, Don Swanson, appearing in that factor, which confirms Jeong *et al.* (2014)'s result.

Table 2. Factor analysis results.

ID	*Factor*	*Model 0*	*Model 1*	*Model 2*	*Model 3*
1	**Information retrieval**	√	√	√	√
2	Information behavior, digital library, information usage	√	√		√
3	**Language model, query, clustering**	√	√	√	√
4	Classification, information organizations			√	√
5	**Text mining, machine learning**	√	√	√	√
6	**User interface**	√	√	√	√
7	User acceptance of information technology		√		√
8	Information systems				√
9	Data Mining, data Analysis		√	√	√
10	**Evaluation indicator, index**	√	√	√	√
11	**Webometrics**, social network analysis	√	√	√	√
12	Visualization, mapping		√	√	√
13	**Scholarly communication**	√	√	√	√
14	**Journal citation analysis, interdisciplinarity,** evaluation	√		√	√

	of algorithms				
15	**Network analysis**	√	√	√	√
16	**Bioinformatics**	√	√	√	√

Although using the same dataset as Jeong *et al.* (2014), we input approximately 500 authors into factor analysis while Jeong *et al.* (2014) did 100. We try to compare our results with theirs, which shows in **bold** in Table 2, where we can find that the results are similar and confirm each other's. The factors extracted by Models 1 and 2 are more in detail than Model 0, i.e., the baseline. On the other hand, with respect to the factor analysis result of Model 3, we can find many detailed sub-fields of information science, such as visualization and data mining. These newly-detected domains are able to showcase the nuance and the emerging topics of information science recently. Therefore, we believe that our proposed methods, when inputting more metadata in full text into ACA, can reveal more details and nuance in depicting scientific intellectual structures.

Network Analysis

Figures 2-5 show the scientific intellectual structures by using the four models, respectively, where each node represents an author and the size of the node is proportional to the degree of the node in the given network. The distance between nodes are determined by ForceAtlas2 (Jacomy, Venturini, Heymann, & Bastian, 2014), a frequently used layout algorithm in Gephi. If two nodes lie near in the map, for instance, their relationship could be strong; and *vice versa*. For visualization, we employ Modularity algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008), in which nodes classified in the same cluster have significantly more links than those in different clusters, compared with a null model. Based on the results of Modularity algorithm, we assign nodes with different colors. The nodes (authors) within the same color indicate that their research interests are similar, while those in different colors show that their research interests should be distinct. The labels of the clusters are manually given by our reading literatures of the authors as well as browsing their personal websites. From Figure 2 we can see that four clusters are detected, bibliometrics, information retrieval, information behavior, and library science/qualitative research. The results are similar to Jeong *et al.* (2014)'s result, where they also found four clusters, bibliometrics, information retrieval (I), information retrieval (II), and library science. Moreover, Figures 3 and 4 detect five clusters; besides four clusters having been detected in Figure 2, it also finds "test mining/data mining". In Figure 5, another cluster, namely "network-based information science" is detected. These two new clusters reveal the nuance of the development of IS and are micro-level sub-fields in IS. These indicate that our proposed models provide more details in knowledge domain maps and help better understand the domain, and that the more full text-based metadata involved, the more details we can obtain.

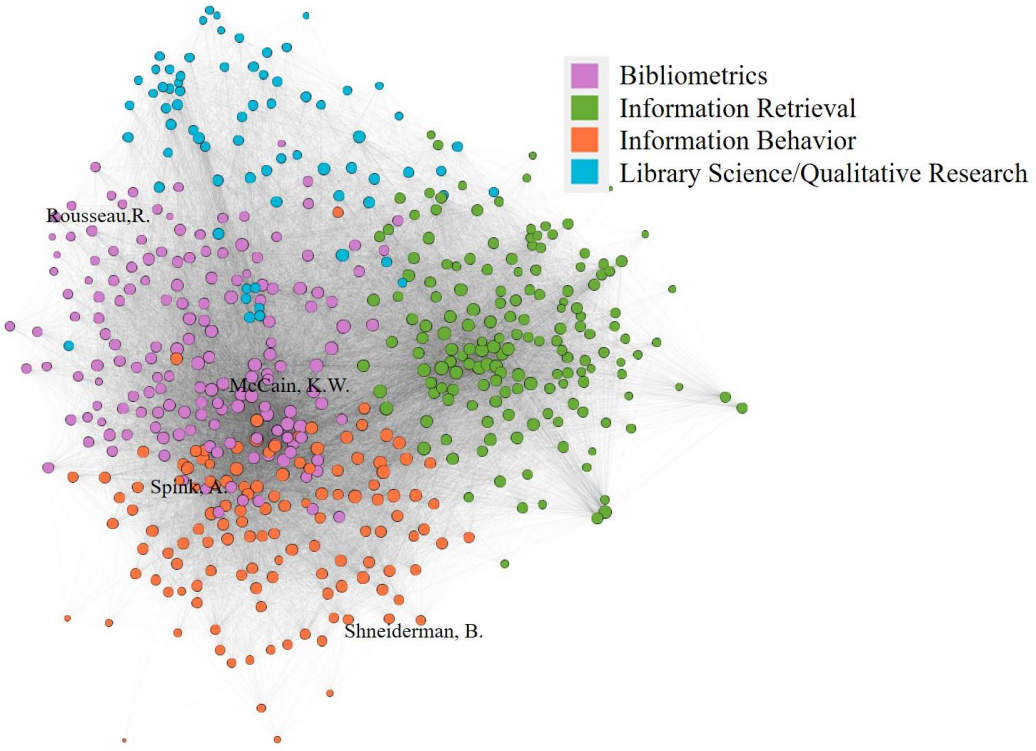

Figure 2. Knowledge domain map: Model 0 (traditional ACA).

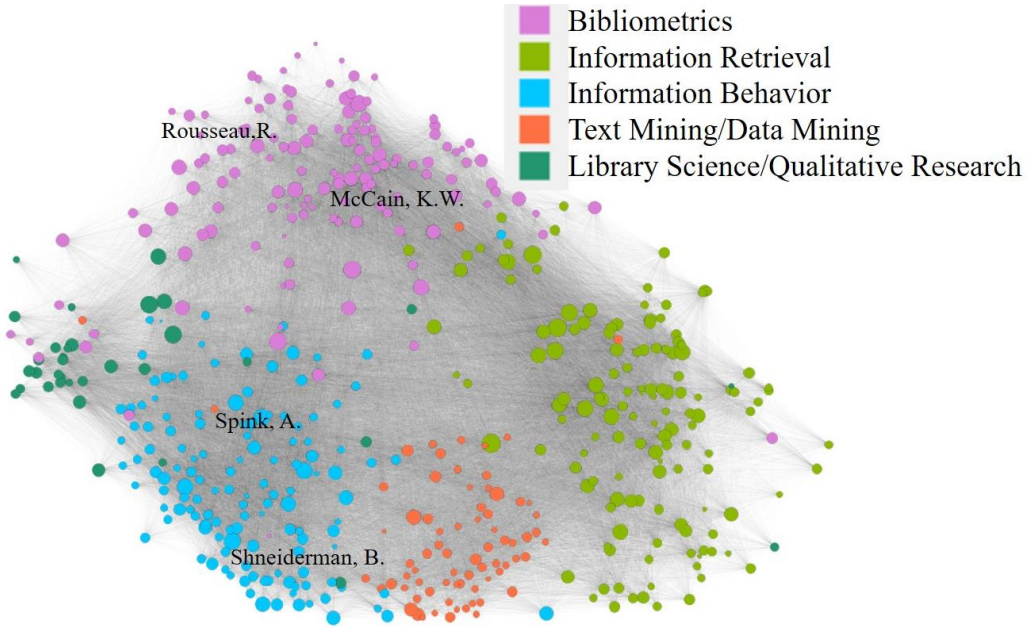

Figure 3. Knowledge domain map: Model 1.

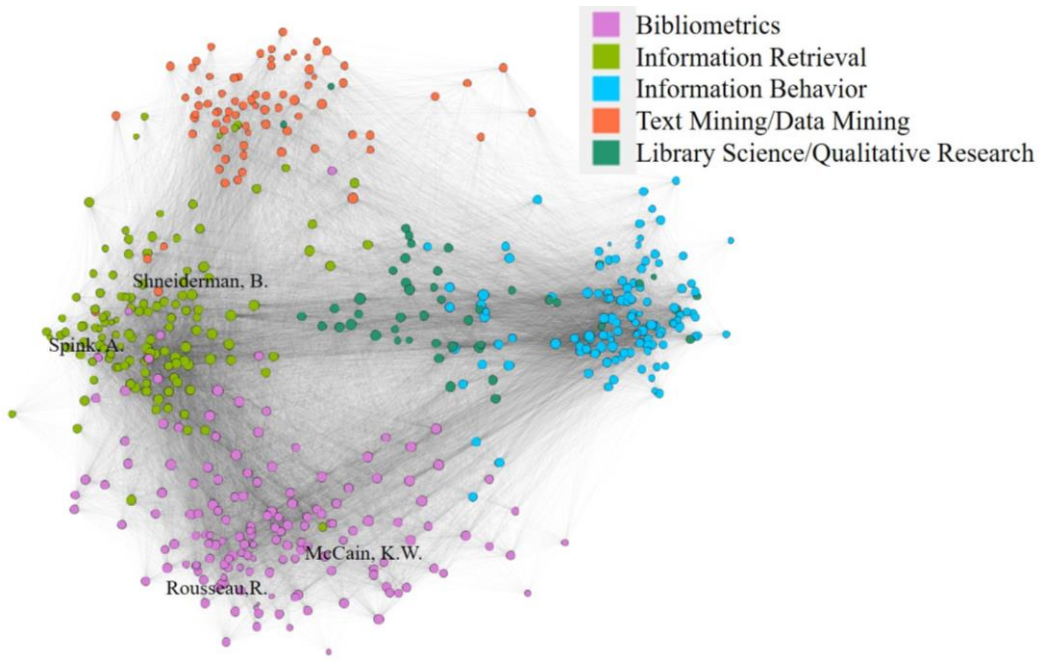

Figure 4. Knowledge domain map: Model 2.

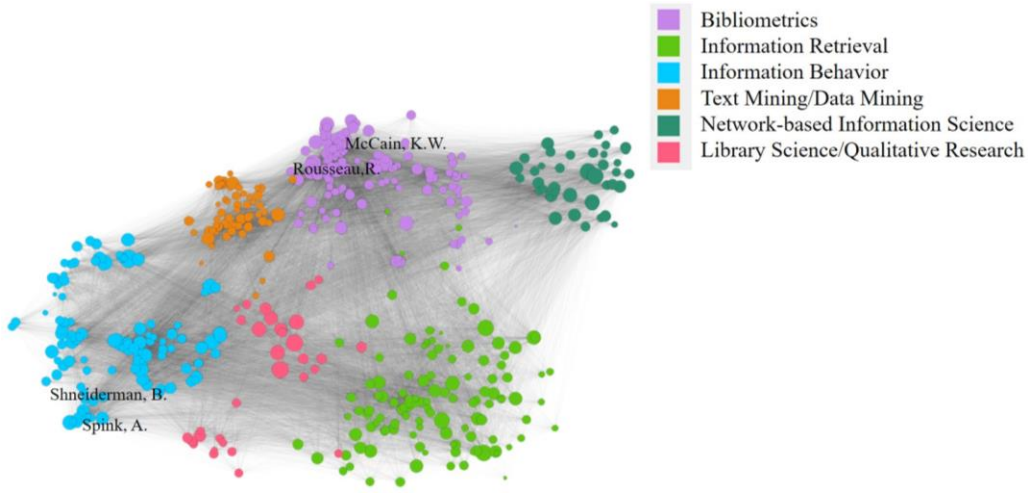

Figure 5. Knowledge domain map: Model 3.

Intuitively, the nodes within the same cluster lie nearer, and the nodes in different clusters lie farther in Figures 3-5 than in Figure 2; and those distances are larger in Figure 5 than in Figures 3-4. These indicate a better clustering performance in Models 1-3 than Model 0 and better in Model 3 than Models 1-2. Take K. W. McCain and R. Rousseau as examples. Both of them focus on bibliometrics during their main scientific careers. Specifically, they are both interested in ACA, where McCain (1990) gave a comprehensive overview of ACA and used ACA to map knowledge domain of IS field between 1972 and 1995 (White & McCain, 1998). Rousseau proposed several types of ACA by classifying them according to distinct requirements (Rousseau & Zuccala, 2004) and discussed whether Pearson's *r* should be used in ACA (Ahlgren, Jarneving, & Rousseau, 2004). Their positions in Figures 3-5 are nearer than those in Figure 2, which shows that the proposed method plays a role of closing authors with similar research interests. Another examples come from A. Spink and B. Shneiderman, in which the former researcher is an expert in information seeking behavior (Spink, Ozmutlu, & Ozmutlu, 2002; Spink & Cole, 2005) while the latter has been concentrating on user behavior analysis (Shneiderman, 1978). We know that both of them are behavior scientists. Although the nodes representing these two authors are

not closely with each other in Figure 2, they move nearer in the visualization of new proposed models. However, the distance between Spink and McCain becomes farther in Figures 3-5 than Figure 2, indicating that our proposed method separates authors sharing different research interests in knowledge domain maps. All of these facilitate the quality of maps in terms of the clustering performance.

To understand the difference among methods more clearly, we compare two properties of the networks generated by our four methods, network density and average clustering coefficient (ACC). We know that all networks are weighted instead of binary, i.e., the edges in the networks range from zero to a certain positive number but not purely zero and one; therefore, to calculate the properties, we employ Barrat, Barthelemy, Pastor-Satorras, & Vespignani (2004)'s algorithm. Based on network science theories, a denser network indicate more interactions among nodes—in the case of scientific intellectual structures, more pieces of information are therefore presented—; a network with greater ACC shows a better clustering performance. Table 3 shows the density and ACC for networks generated by four methods, in which we can find that Model 0 (traditional ACA) has a low density, but when we add the factor of mentioned time or context word, the density of the network increases. If both factors are involved, the density doubles compared with that in Model 0. In terms of the ACC, we find that when more metadata in full text are employed, the values of ACC raise. All of these descriptive statistics indicate that our proposed ACA methods combining metadata in full text enhance the clustering performance and provide more information in scientific intellectual structure depicting, which echoes our findings in the aforementioned sections.

Table 3. Descriptive statistics of the network properties in all implemented methods.

Model	*Density*	*ACC**	*Model*	*Density*	*ACC*
Model 0	0.06	0.03	Model 2	0.09	0.06
Model 1	0.09	0.07	Model 3	0.12	0.11

Note: ACC = Average cluster coefficient.

MDS-measurement

In order to evaluate the performance of our proposed method quantitatively, we employ multi-dimensional scaling measurement (MDS-measurement) (Bu *et al.*, 2016) to supplement our qualitative arguments in the "Network Analysis" section. Different from MDS that is a typical way to show nodes in a two- or three-dimensional map (essentially a visualization algorithm), MDS-measurement, a bibliometric indicator to evaluate the clustering performance provided by MDS, aims to calculate the MDS-measurement value (σ), which is equal to the ratio between the sum of the distance between the nodes within the same cluster (c), and the sum of the distance between the nodes in different clusters (S), i.e., $\sigma = c/S$. Intuitively, a smaller σ indicates better clustering performance in knowledge domain maps in which nodes within the same cluster lie nearer while those in different clusters lie farther. Table 4 shows the MDS-measurement result, where we can see that the MDS-measurement value (σ) of Models 1-3 is smaller than that of Model 0 (traditional ACA), indicating a better clustering result in knowledge domain map. Also, σ in Model 3 is the smallest among Models 1-3, showing that the cluster performance of Model 3 is the best among our proposed methods. This confirms our observation in the "Network Analysis" section.

Table 4. MDS-measurement results.

Method	c	S	$\sigma(= c/S)$
Model 0	546.61	4303.53	12.70%
Model 1	526.11	4308.42	12.21%

Method	c	S	$\sigma(=c/S)$
Model 2	529.54	4319.73	12.26%
Model 3	518.83	4396.14	11.80%

CONCLUSION

This paper proposes a novel method combining the numbers of mentioned times and context words into traditional author co-citation analysis (ACA). The results show that compared with the traditional method, our newly proposed approach not only shows better clustering performance but also provides more details in knowledge domain mappings. Considering that this method does not need a large volume of calculation such as content-based ACA (Jeong *et al.*, 2014; Kim, Jeong, & Song, 2016; Hsiao & Chen, 2017), we believe that our proposed method are easily applied to various disciplines so as to depict scientific intellectual structures by involving more information and improving the traditional ACA.

Besides the method itself and its advantages compared with the traditional ACA, this study provides several implications to the future researchers. Firstly, we use full-text data but not intend to analyze content- or semantic-level information, which breaks the conventional thinking to use complex natural language processing technologies aiming at mining content- or semantic-level data so as to map knowledge domains. Secondly, our approach inspires future researchers to duplicate this method on other scholarly network analyses, such as author bibliographic coupling analysis (Zhao & Strotmann, 2008a) and coauthorship analysis (Bu et al., 2017a; Ding, 2011b; Zhang et al., 2018). Specifically, the metadata in full text can be involved into an author bibliographic coupling network by normalizing and assigning certain weight values. Furthermore, this research supplements the framework of bibliometric elements proposed by Morris and Vander Veer Martens (2008), in which papers, paper authors, paper journals, references, reference authors, reference journals, and index terms are included. Our study provides "citing sentences" as a bridge between "papers" and "references", and shows the potential detailed affiliations upon "citing sentences" such as the number of mentioned times and the number of context words of references. These have offered significant foundations for future supplements of the bibliometric element framework when more full-text data are involved.

Nevertheless, simply duplicating this method is *not always* wise. Although we find that our method can mine more details that traditional ACA cannot do, a combination between traditional method and our proposed approaches should often get much better performance. Combining two methods can provide more distinct perspectives to make sense on the bibliometrical relationship among authors from a retrospective view. Practically, if many full text-based metadata have been extracted from the raw dataset, a traditional ACA will also be doable.

One of the key steps of our proposed method is aggregating paper- into author-level information. For instance, when calculating the mentioned time parameter, we first do the parameter in a given *paper* and then aggregate it to a given *author* by calculating the mean values. Following this idea, we can then aggregate the parameters into other levels, including topics, journals, and even disciplines.

However, there are several limitations in this research. For example, we only used first authors' information instead of all authors'. The accuracy might thus be negatively affected. Moreover, there are still many other types of metadata that are not used to involve in ACA in previous or the current studies, such as the sequence of co-cited authors (He, Ding, & Yan, 2012) and the number of figures or tables (Lee, West, & Howe, 2018). We would like to focus on these under the context of ACA as well as other scholarly network analyses in the future.

ACKNOWLEDGMENT

The present study is an extended version of an article presented at the 16th International Conference on Scientometrics and Informetrics, Wuhan (China), 16-20 October 2017. The authors would like to thank Bikun Chen and Ming-Yueh Tsay for their insightful suggestions on the first draft of this article. We are also grateful to two anonymous reviewers and the guest editors of this special issue for their improving the quality of this paper.

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 and Technology, 54*(6), 550-560.

Ahlgren, P., Jarneving, B., & Rousseau, R. (2004). Author co-citation analysis and Pearson's *r*. *Journal of the American Society for Information Science and Technology, 55*(9), 843-844.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Proceeding of the third International Conference on Web and Social Media* (pp. 361-362), May 17-19, 2009, San Jose, California, U.S.A.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of United State of America, 101*(11), 3747-3752.

Blondel, D.V., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 10*, 1000.

Bruer, J.T. (2010). Can we talk? How the cognitive neuroscience of attention emerged from Neurobiology and Psychology, 1980-2005. *Scientometrics, 83*(3), 751-764.

Bu, Y., Liu, T., & Huang, W.-B. (2016). MACA: A modified author co-citation analysis combined with general metadata of citations. *Scientometrics, 108*(1), 143-166.

Bu, Y., Ni, S., & Huang, W.-B. (2017a). Combining multiple scholarly relationships with author cocitation analysis: A preliminary exploration on improving knowledge domain mappings. *Journal of Informetrics, 11*(3), 810-822.

Bu, Y., Wang, B., Huang, W.-B.*, & Che, S. (2017b). MFTACA: An author cocitation analysis method combined with metadata in full text. In *Proceedings of the 16th International Conference on Scientometrics and Informetrics (ISSI 2017)* (pp. 916-927), October 16-20, 2017, Wuhan, Hubei, China.

Cano, V. (1989). Citation behaviour: Classification, utility, and location. *Journal of the American Society for Information Science, 40*(4), 284-290.

Case, D.O., & Higgins, G.M. (2000). How can we investigate citation behaviour? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science, 51*(7), 635-645.

Chen, L.C., & Lien, Y.H. (2011). Using author co-citation analysis to examine the intellectual structure of E-leaning: A MIS perspective. *Scientometrics, 89*(3), 867-886.

Chu, K.C., Liu, W.I., & Tsai, M.Y. (2012). The study of cocitation analysis and knowledge structure on healthcare domain. In *Proceedings of the sixth Global Conference on Power Control and Optimization* (pp. 247-253), August 6-8, 2012, Las Vegas, Nevada, U.S.A.

Ding, Y. (2011a). Topic-based PageRank on author co-citation networks. *Journal of the American Society for Information Science and Technology, 62*(3), 449-466.

Ding, Y. (2011b). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics, 5*(1), 187-203.

Ding, Y., Chowdhury, G., & Foo, S. (1999). Mapping intellectual structure of Information Retrieval: An author cocitation analysis, 1987-1997. *Journal of Information Science, 25*(1), 67-78.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics, 7*(3), 583–592.

Eom, S. (1999). Decision Support Systems Research: Current State and Trends. *Industrial Management and Data Systems, 99*(5), 213-221.

Eom, S. (2008a). All author co-citation analysis and first author co-citation analysis: A comparative empirical investigation. *Journal of Informetrics, 2*(1), 53-64.

Eom, S. (2008b). *Author co-citation analysis: Quantitative methods for mapping the intellectual structure of an academic discipline*. Hershey, NY: Information Science Reference.

Finlay, C.S., Sugimoto, C.R., Li, D., & Russell, T.G. (2012). LIS dissertation titles and abstracts (1930-2009): Where have all the librar* gone? *Library Quarterly, 82*(1), 29-46.

He, B., Ding, Y., & Yan, E. (2012). Mining patterns of author orders in scientific publications. *Journal of Informetrics, 6*(3), 359-367.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569-16572.

Hsiao, T.-M., & Chen, K.-H. (2017). Yet another method for author co-citation analysis: A new approach based on paragraph similarity. In *Proceedings of the 80th Annual Meeting of the Association for Information Science and Technology* (pp. 170-178), October 27-November 1, 2017, Washington D.C., U.S.A.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2: A continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE, 9*(6), e98679.

Janssens, F., Leta, J., Glänzel, W., & Moor, B.D. (2006). Towards mapping library and information science. *Information Processing and Management, 42*(6), 1614-1642.

Jeong, Y.-K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics, 8*(1), 197-211.

Kim, H.-J., Jeong, Y.-K., & Song, M. (2016). Content- and proximity-based author co-citation analysis using citation sentences. *Journal of Informetrics, 10*(4), 954-966.

Lee, P., West, J.D., & Howe, B. (2018). Viziometrics: Analyzing visual information in the scientific literature. *IEEE Transactions on Big Data, 4*(1), 117-129.

McCain, K.W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science, 41*(6), 433-443.

Mêgnigbêto, E. (2013). Controversies arising from which similarity measures can be used in co-citation analysis. *Malaysian Journal of Library and Information Science, 18*(2), 25-31.

Milojević, S., Sugimoto, C.R., Yan, E., & Ding, Y. (2011). The cognitive structure of Library and Information Science: Analysis of article title words. *Journal of the American Society for Information Science and Technology, 62*(10), 1933-1953.

Morris, S.A., & Van der Veer Martens, B. (2008). Mapping research specialties. *Annual Review of Information Science and Technology, 42*(1), 213-295.

Nakov, P.I., Schwartz, A.S., & Hearst, M.A. (2004). Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the Twenty-seventh ACM SIGIR Conference Workshop on Search and Discovery in Bioinformatics*, July 25-29, 2004, Sheffield, U.K.

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), 513-529.

Shneiderman, B. (1978). Jump searching: A fast sequential search technique. *Communications of the ACM, 21*(10), 831-834.

Spink, A., & Cole, C. (2005). Human information behavior: Integrating diverse approaches and information use. *Journal of the American Society for Information Science and Technology, 57*(1), 25-35.

Spink, A., Ozmutlu, H.C., & Ozmutlu, S. (2002). Multitasking information seeking and searching processes. *Journal of the American Society for Information Science and Technology, 53*(8), 639-652.

Sugimoto, C.R., Li, D., Russell, T.G., Finlay, S.C., & Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using Latent Dirichlet Allocation. *Journal of the American Society for Information Science and Technology, 62*(1), 185-204.

White, H.D. (2004). Author cocitation analysis and Pearson's *r*. *Journal of the Association for Information Science and Technology, 55*(9), 843-844.

White, H.D., & Griffith, B.C. (1981). Author co-citation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*(3), 163-171.

White, H.D., & McCain, K.W. (1998). Visualizing a discipline: An author cocitation analysis of information science, 1972–1995. *Journal of the American Society for Information Science, 49*(4), 327–355.

Yang, S., Han, R., Wolfram, D., & Zhao, Y. (2016). Visualizing the intellectual structure of information science (2006-2015): Introducing author keyword coupling analysis. *Journal of Informetrics, 10*(1), 132-150.

Zhao, D. (2006). Towards all-author co-citation analysis. *Information Processing and Management, 42*(6), 1578-1591.

Zhao, D., Cappello, A., & Johnston, L. (2017). Functions of uni- and multi-citations: Implications for weighted citation analysis. *Journal of Data and Information Science, 2*(1), 51-69.

Zhao, D., & Strotmann, A. (2008a). Evolution of research activities and intellectual influences in Information Science 1996-2005: Introducing author bibliographic coupling analysis. *Journal of the American Society for Information Science and Technology, 59*(13), 2070-2086.

Zhao, D., & Strotmann, A. (2008b). Comparing all-author and first-author co-citation analyses of Information Science. *Journal of Informetrics, 2*(3), 229-239.

Zhao, D., & Strotmann, A. (2014). The knowledge base and research front of Information science 2006-2010: An author co-citation and bibliographic coupling analysis. *Journal of the Association for Information Science and Technology, 65*(5), 996-1006.

Zhao, H., Zhang, F., & Kwon, J. (2017). Corporate social responsibility research in international business journals: An author co-citation analysis. *International Business Review*, doi:10.1016/j.ibusrev.2017.09.006.

Zhang, C., Bu, Y., Ding, Y., & Xu, J. (2018). Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science and Technology, 69*(1), 72-86.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, *66*(2), 408-427.