Journal of Informetrics 11 (2017) 810–822

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

Combining multiple scholarly relationships with author cocitation analysis: A preliminary exploration on improving knowledge domain mappings

Yi Bu [a], Shaokang Ni [b], Win-bin Huang [b,*]

[a] School of Informatics and Computing, Indiana University, Bloomington, IN, USA
[b] Department of Information Management, Peking University, Beijing, China

ARTICLE INFO

Article history:
Received 14 December 2016
Received in revised form 17 June 2017
Accepted 17 June 2017
Available online 6 July 2017

Keywords:
Author cocitation analysis
Coauthorship analysis
Author bibliographic coupling analysis
Scholarly network
Scientific intellectual structure
Knowledge domain mapping
Exploratory factor analysis (EFA)
Bibliometrics

ABSTRACT

Author cocitation analysis (ACA) is a branch of bibliometrics and knowledge representation that aims to map knowledge domains. However, ACA has been criticized because count-based measurement is too simple, and resulting maps are insufficiently informative. Since different scholarly relationships, e.g., coauthorship and author bibliographic coupling relationships, can extract out different relationships among authors in various perspectives, combining them with ACA for constructing knowledge domain mappings is our major purpose. The proposed method constructs the hybrid matrix from all relationships in four steps: relationship normalization, calculating the similarity between scholarly relationships, calculating adjustment parameters, and constructing hybrid relationships. The important parameters for integrating these matrices are calculated according to the distance in the hyperspace transformed from the similarity among the scholarly relationships by exploratory factor analysis. Compared with ACA, the results of the proposed method show: (1) More sub-fields in the given discipline can be identified when combining other scholarly relationships; (2) The more scholarly relationships added into ACA, the more details in terms of research area the method will find; (3) Good visualization in clustering is depicted when we combine other scholarly relationships. As a result, the proposed method offers a good choice to understand researchers and to map knowledge domains in a study field for integrating more scholarly relationships at the same time.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

As an important method in bibliometrics proposed by White and Griffith (1981), author cocitation analysis (ACA) aims to utilize cocitation relationships between author pairs to map the intellectual structures of knowledge and research domains (Jeong, Song, & Ding, 2014; Kim, Jeong, & Song, 2016). McCain (1990) summarized four main steps of ACA: 1) selecting and retrieving the author set; 2) constructing the raw cocitation matrix; 3) transforming the cocitation matrix into a correlation matrix; and 4) analyzing the data and interpreting the results. ACA works well independently of the specific application domains as evidenced by its popularity for mapping various scientific disciplines (e.g., Chen & Lien, 2011; Chu, Liu, & Tsai, 2012). However, past research has noted defects of ACA, such as that its inputs only includes author cocitation count

* Corresponding author at: Fanglibangqin Building 521, Department of Information Management, Peking University, 5 Yiheyuan Road, Haidian District, Beijing 100871, China.
E-mail address: huangwb@pku.edu.cn (W.-b. Huang).

http://dx.doi.org/10.1016/j.joi.2017.06.004
1751-1577/© 2017 Elsevier Ltd. All rights reserved.

(frequency), which has been criticized for simply using count-based information (Bu, Liu, & Huang, 2016). Despite some improvements by integrating more general descriptive metadata (Bu et al., 2016; Zhao, 2006), ACA remains limited by the cocitation perspective, thus constraining its performance for mapping intellectual structures and knowledge domains.

According to Zhao (2012), there are four main research paradigms in knowledge domain mappings: the traditional paradigm (mainly using manual literature studies and review), the theory paradigm (based on sociology of science founded by Merton (1973)), the bibliometrics paradigm, and the social network analysis paradigm (primarily based on complex network/system theories and technologies). Regarding the "bibliometrics paradigm", researchers have used several scholarly relationship/network analysis methods, e.g., bibliographic couplings analysis (Boyack & Klavans, 2010; Kessler, 1963; Zhao & Strotmann, 2008a), citation analysis (Garfield, Sher, & Torpie, 1964), cocitation analysis (McCain, 1991; Small & Griffith, 1974; White & Griffith, 1981), coauthorship analysis (Beaver & Rosen, 1979), and co-word analysis (Callon, Courtial, & Turner, 1983), each of which provide different perspectives concerning scientific intellectual structures (Ma & Ni, 2012; Zhao & Strotmann, 2008a). The cocitation relationship is not the only way to show scientific intellectual structures, and combining different scholarly networks can provide broader visual thresholds. The method proposed in this article, which combines coauthorship and author bibliographic coupling analyses into ACA, thus aims to improve the performance of knowledge domain mapping. Note that the two terminologies, scholarly networks and scholarly relationships, are represented as matrices with the same meaning in this article.

The outline of this article is as follows. Related work is provided in Section 2. The point of scholarly relationship combinations is described in Section 3. The dataset used in this paper and the methods combining different scholarly relationships are proposed in Section 4. The results of the empirical studies and our observation are illustrated in Section 5. Finally, the conclusion is remarked in Section 6.

2. Related work

The basic assumptions of ACA are that each citation plays an equal role in cocitation analysis and that cocitation counts of the author pair are proportional to their relevance (White & Griffith, 1981). Thus, in ACA, two authors are connected if and only if they were cocited at least once, and the more they are cocited, the stronger cocitation relationship they will have. However, traditional ACA, founded by White and Griffith (1981) and standardized by McCain (1990), takes as input the cocitation counts of first authors, resulting in a small amount of useful information and thus negatively impacting the performance of visualization—two branches of study have emerged to explore this problem. The first branch lay in all-author cocitation analysis (AACA), pioneered by Persson (2001). Other scholars followed, classifying several kinds of ACA according to their methods of cocitation counting, such as first-author cocitation analysis (FACA), inclusive AACA, and exclusive AACA (Zhao, 2006). Zhao and Strotmann (2008b) as well as Eom (2008) found that AACA works better to capture all influential researchers in a field and can identify more sub-specialties than FACA. The other branch of research explored general descriptive metadata into ACA. For example, Bu et al. (2016) combined citation published time, citation published venue (e.g., journal, proceeding, etc.), and citation keywords to reveal more details and nuance in mapping knowledge domains.

A debate of how to transform raw cocitation matrix to correlation matrix in ACA is worth noting (Mêgnigbêto, 2013). Although researchers have adopted Pearson's r in ACA since its birth, Ahlgren, Jarneving, and Rousseau (2004) provide a theoretical perspective and rigorous mathematical proof to argue that Pearson correlation coefficient's use in ACA has several drawbacks. Nevertheless, White (2004), a representative of the "Drexel team", found that although Pearson's r might fluctuate in different experiments, clusters based on it "are much the same for the combined groups as for the separate groups" (p. 1250), but he emphasized that he had never disagreed to use other similarity measurements beyond Pearson's r for ACA, such as cosine and Jaccard similarities. Such debates have been ongoing for more than ten years without resolve (Egghe & Leydesdorff, 2009; Van Eck & Waltman, 2008). Nevertheless, during this debate, bibliometricians have had a consensus that bibliometric analyses should be carried out at least using the theoretically most appropriate methods. In this paper, therefore, we are going to employ the cosine similarity to transform our raw matrix to correlation matrix instead of Pearson's r, because the method we propose contains author cocitation, author bibliographic coupling, and coauthorship frequencies and we do not aim to obtain a uni-scholarly-network probability like traditional ACA.

Bibliographic coupling analysis (BCA) reveals that the more similar the research topics of two articles, the more references they share, and was first proposed by Kessler (1963). Coupling strength (coupling frequency) in BCA is defined as the number of references two papers share (cocite). BCA is regarded as a static analysis because the coupling strength of coupled papers does not change. Zhao and Strotmann (2008a) proposed author bibliographic coupling analysis (ABCA) by using the author as the object of research and indicated that ABCA reveals the active researchers in a field and provides a perspective of the structure of the research front. Furthermore, an important reflection of Zhao and Strotman (2008a)'s study is that the coupling strength of the coupled authors in ABCA may change while the authors cocite more papers.

ABCA was used in combination with ACA from its inception. Zhao and Strotmann (2008a) argued that ABCA can be regarded as a complement of ACA and their combination will provide a more comprehensive view of the intellectual structures than either of them can provide on its own; specifically, ACA is considered as a look back in time (historical analysis) and ABCA as a view of the presence (current research front). They also pointed out that the extrapolation of major differences between ACA and ABCA could provide a forecasting method—a glimpse of the likely future development of the field under analysis. More recently, Zhao and Strotmann (2014) verified that their forecast made in 2008 had been remarkably good, and proceeded to provide another forecast. Ma and Ni (2012) used China Social Science Citation Index (CSSCI) to perform empir-

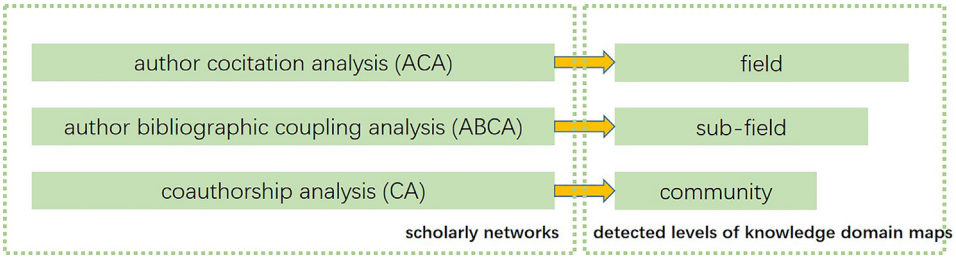

Fig. 1. Relationships and performances on different levels of knowledge domain maps detecting among three types of scholarly relationships.

ical studies and found that ABCA is a satisfactory method for depicting scientific intellectual structures. They also concluded that if ABCA can be combined with ACA, the intellectual structures could be more accurate. Based on these previous studies, this paper combines ABCA and coauthorship analysis into ACA to supplement more perspectives into mapping knowledge domains to make it more precise, informative, and scientific.

Coauthorship analysis (CA) is often used to measure scientific collaboration (Milojević, 2010; Zitt, Bassecoulard, & Okubo, 2000). Two authors have coauthorship relationships when they have coauthored at least one article. The more articles they have coauthored, the stronger the relationship. CA could be used to construct small "communities" under the level of "sub-fields". Such small "communities" form and strengthen clusters of "sub-fields", which can then allow for accurate and reliable mapping of knowledge domains. Zhang, Bu, and Ding (2016) used CA to map knowledge domains, in which some influence factors of scientific collaboration are also explored for analyzing the coauthorship network.

3. Scholarly relationship combinations

The similarities among authors' research areas are mainly considered in knowledge domain mappings via different kinds of scholarly relationships. In other words, the more similar two authors' research interests (topical relatedness) are, the nearer they should be located in the knowledge domain maps. In spite of this, different types of scholarly networks provide distinct perspectives and levels of analyses for mapping knowledge domains and depicting scientific intellectual structures. As shown in Fig. 1, ACA forms large-scale structures at the level of *fields*, ABCA specifies small *sub-fields*, and CA mines deeper to identify clusters of *communities*.

ACA is able to cluster authors in knowledge domain mappings by calculating cocitation counts (Ding, Chowdhury, & Foo, 1999; McCain, 1990; White & Griffith, 1981). Specifically, ACA positions authors having similar research fields nearer and correspondingly keeps those in different research fields at a distance; the results of this method of mapping shows clusters of authors representing different research fields. A cluster in the map indicates that many authors in the dataset pay attention to the authors in it. Their issues are regarded as a field because it catches many authors' participation. However, this method of distance calculation could lead to a purely macro view of a knowledge domain without details.

Unlike ACA, ABCA concentrates on the relationships of the authors citing common articles. Different from ACA in which the nodes represent more highly-cited authors, the nodes revealed in ABCA are the prolific authors in a domain and have a certain number of authors having common interests in a period of time. Moreover, a cluster in the map indicates that a number of prolific authors research on similar topics. These topics can be regarded as sub-fields in a field because they indicate that a group of authors' research is related in a time period, an observation also illustrated by Zhao and Strotmann (2008a). Meanwhile, CA provides the perspective of scientific collaboration, whereby more collaborations between authors, the stronger the relationship between them and the nearer they appear on the map. These collaborating authors become a small "community" under the level of a "sub-field". Such small "communities" form and strengthen clusters of "sub-fields", which adds further nuance, accuracy, and reliability to the knowledge domain map.

Based on the above, ACA can outline the knowledge base of a field and plays a fundamental role in mapping knowledge domains. This reveals the reason why our proposed method is "ACA-centered" (See details in Section 4.2). The other two scholarly relationships can be integrated into ACA in order to present more information in the same knowledge domain maps. In this paper, the combination of these three types of scholarly relationships is regarded as a combination with different weights that contribute to the creation of a knowledge domain map.

4. Methodology

The steps of our algorithm is shown in Fig. 2. All data are sourced from AMiner (Tang et al., 2008), an academic social network analysis and mining system. After the data are extracted from AMiner, the author cocitation, author bibliographic coupling, and coauthorship relationships are calculated in the traditional way. Specifically, the constructions of author cocitation and author bibliographic coupling relationships are based on "pure-first author-cocitation" and "pure-first author-bibliographic-coupling" methods, as detailed by Rousseau and Zuccala (2004). That is to say, when two papers are cocited, the bibliographic coupling is only formed between the first authors or sole-authors of each paper. A raw matrix is constructed by combining the three relationships according to the proposed method which is called ACA-centered combination and will

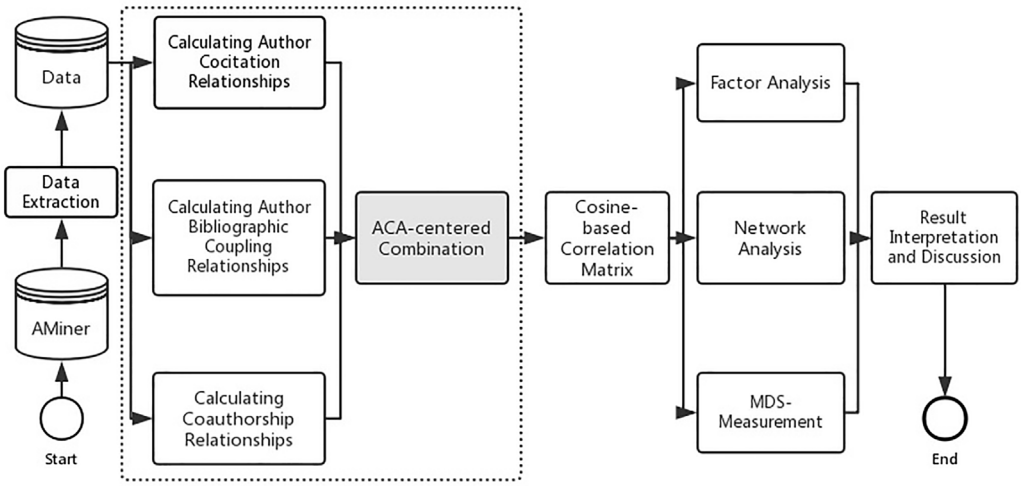

Fig. 2. Flow diagram of the proposed algorithm. Note that the area outlined by the dotted line is the major difference between the proposed method and others.

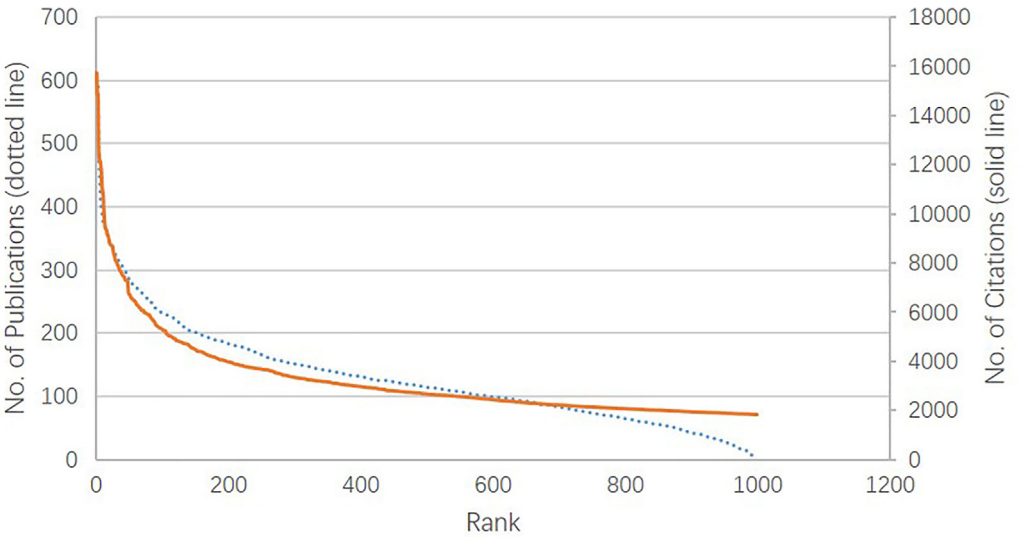

Fig. 3. Basic descriptive statistics of the author set. The blue dotted line and the solid red line represent the number of publications and that of citations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

be detailed in Section 4.2. The cosine similarity is then employed to transform the raw matrix into the correlation matrix. Gephi (Bastian, Heymann, & Jacomy, 2009), a visualization tool, is used to display the results of the hybrid network for discussion and analysis. We use the R programming language to implement our methods, factor analysis, network analysis, and MDS-measurement.

4.1. Data

We examine the AMiner dataset, a platform which allows for the search and analysis of academic social networks, and covering most important conference and journals in Computer Science (Tang et al., 2008). It includes 2,092,356 papers published from 1936 to 2014, 1,207,061 unique authors, and 8,024,869 citation relationships. The 1000 most-cited authors who have received at least 132 citations until the year of 2014 are selected in our experiment. The distributions of number of publications and that of citations among the authors are shown in Fig. 3, and obviously they are approximately in accordance with the long tail law. We use these selected authors to calculate the author cocitation-, coauthorship-, and author bibliographic coupling-relationships. For example, the value of author cocitation relationship between two authors is equal to the number of these two authors cocited; the other two relationships are calculated in the similar way. Then author cocitation, coauthorship, and author bibliographic coupling networks, are constructed according to these relationships, respectively.

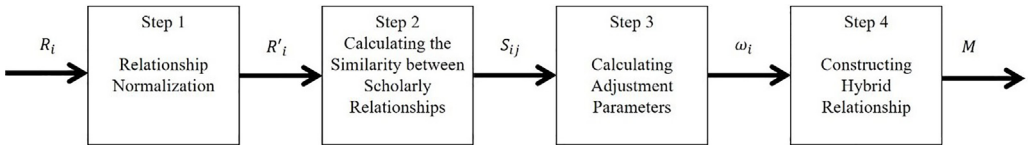

Fig. 4. Steps of ACA-centered combinations.

4.2. ACA-centered combination

ACA-centered combination indicates that ACA is regarded as a baseline while combining different scholarly relationships because ACA plays a fundamental role in mapping knowledge domains. The key step of the proposed method is to firstly determine the similarity among these relationships after normalization. The final hybrid relationship is then the summation of each relationship multiplied by its corresponding parameter (weight value) adjusted according to their similarities. As a result, the four steps illustrated in Fig. 4 are included in the proposed ACA-centered combination and detailed in the following.

Step 1: Relationship Normalization

Assume that $(\psi - 1)$ distinct scholarly relationships are combined into the author cocitation relationship and δ is the number of authors in the dataset. R_i, of which the element $r_{i,p,q}$ is the i-type scholarly relationship between author p and q, is a $\delta \times \delta$ relationship matrix, where $i = 1, 2, \ldots, \psi$ indicates the corresponding scholarly relationships. Specifically, R_1 refers to the author cocitation relationship, and $R_x\,(x = 2, 3, \ldots, \psi)$ indicates the other combined scholarly relationships. R_i is symmetric and the elements on the main diagonal of R_i is set by zero. Then the normalization can be defined as:

$$R_i' = \frac{1}{max_i} \cdot R_i \tag{1}$$

where max_i is the maximum value in R_i, and R_i' is a normalized matrix of the i-type scholarly relationship.

Step 2: Calculating the Similarity between Scholarly Relationships

After normalization, symmetric Kullback-Leibler divergence (Wang et al., 2011) is exploited to calculate the similarities between scholarly relationships. Symmetric Kullback-Leibler divergence (sKL divergence) is calculated as the sum of Kullback-Leibler divergence between two matrices. The reasons why we employ sKL divergence to measure the similarity of two matrices lie in the fact that these matrices essentially contain possibility values normalized between zero and one. The similarity, s_{ij}, between R_i' and R_j', can be formulated as:

$$s_{ij} = \frac{1}{2}\left[tr\left(R_i'^{T} \cdot R_j'\right) + tr\left(R_j'^{T} \cdot R_i'\right) - \ln\frac{|R_j'|}{|R_i'|} - \ln\frac{|R_i'|}{|R_j'|}\right], \quad \forall i, j = 1, 2, \ldots, \psi \tag{2}$$

where $tr(R_i')$ refers to the trace of R_i', i.e., the sum of the elements on the main diagonal of R_i', $R_i'^{T}$ is the transpose of R_i', and $|R_i'|$ is the value of the determinant corresponding to R_i'. After the calculation, the results can be formed as a $\psi \times \psi$ similarity matrix S, composed by $s_{ij}\,(\forall i, j = 1, 2, \ldots, \psi)$. Essentially S contains the similarities between each pair of scholarly relationships measured by Frobenius Norm.

Step 3: Calculating Adjustment Parameters

The parameter corresponding to each scholarly relationship is required to adjust their proportion when different scholarly relationships are combined. Taking the matrix S containing similarity measurements of the scholarly relationships as the input, Exploratory Factor Analysis (EFA) (Cattell, 1965) is then exploited to transform different scholarly relationships into a hyperspace. As pointed out by Bryant and Yarnold (1995), EFA is a method to uncover a small number of latent factors that represent the observed variables; this method is distinct from Principal Component Analysis (PCA) that instead detects components to represent the observed variables. EFA *de facto* considers the variables as the linear combination of common factors while PCA uses principal components as the linear combination of the variables. Therefore, EFA provides a better capability to interpret factors and their internal relationships without prior information. The output of EFA is a hyperspace containing different scholarly relationships, as shown in Fig. 5. Each node in the hyperspace is represented as a scholarly relationship and their distances by using Euclidean metric are exploited to formulate the proportion among these scholarly relationships. The adjustment parameter (weight value) for each scholarly relationship is calculated based on the distance between each other in the hyperspace we have generated. Specifically, those who are similar with other scholar relationships should be weighed as a smaller adjustment parameter while those dissimilar with other scholar relationships should be weighed as a larger adjustment parameter. As different scholarly relationships can deliver various information, the purpose of this rule is to display more perspectives of information in the knowledge domain mappings because a scholarly relationship similar to other relationships should yield some of the total proportion of adjustment parameter (weight value) to other dissimilar scholarly relationships; this rule shows more contributions and more diverse information to the scientific intellectual structures.

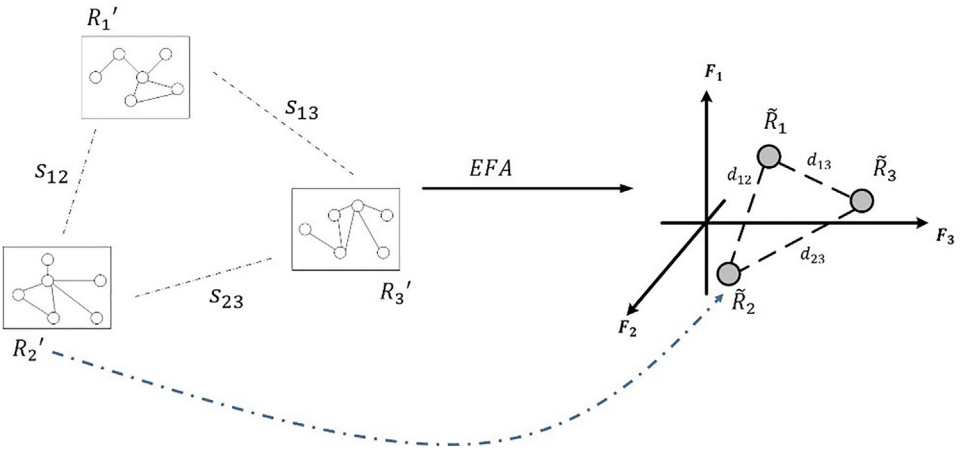

Fig. 5. Details for Step 3 in the ACA-centered combination method.

After similarity calculation, the ith row of S, represented as $S_i = \left(s_{i1}, s_{i2}, \ldots, s_{i\psi}\right)$, indicates the similarities between R_i' and other scholarly relationships. Assume that S_i can be represented as:

$$S_i = \xi_{i1}F_1 + \xi_{i2}F_2 + \ldots + \xi_{in}F_n + \xi_i U_i + e_i \left(n \le \max\left\{i, j\right\}\right) \tag{3}$$

where $F_\sim$ refers to common factors, $\xi_\sim$ factor loadings, U_i unique factor, and e_i random errors (Cattell, 1965). Essentially S_i can be regarded as a vector in an n-dimension space constructed by F_i. Then ξ_i should be the projection of vector S_i on the F_i axis, and it is determined by the position of S_i in the given n-dimension space. However, n could be unlimited to decompose S completely and easily. Assume that p impacted factors of $F_\sim$ are sufficient to formulate S, and Z is defined as a function to select the factors:

$$Z(\{F_1, F_2, \ldots, F_n\}, t) = \left\{F_{\mu_1}, F_{\mu_2}, \ldots, F_{\mu_p}\right\}, \quad (p \le n) \tag{4}$$

where t refers to the threshold value of the Eigenfactor of $F_\sim$, and the index $\mu_\sim \subseteq \left\{1, 2, \ldots, n\right\}$ and $|\mu_\sim| = p$. According to Kaiser's law (1958), we select the factors of $F_\sim$ with an Eigenfactor greater than or equal to 1.0 ($t \ge 1.0$). As a result, the i-type scholarly relationship can be presented as a node, $\tilde{R}_i$, in a p-dimension hyperspace with the coordinate $\xi_i = \left(\xi_{i1}, \xi_{i2}, \ldots, \xi_{ip}\right)$. Then the distance between R_i' and R_j', calculated by using Euclidean metric, defined as:

$$d_{ij} = \sqrt{\left(\xi_{i1} - \xi_{j1}\right)^2 + \left(\xi_{i2} - \xi_{j2}\right)^2 + \ldots + \left(\xi_{ip} - \xi_{jp}\right)^2} \tag{5}$$

According to the distances in the hyperspace, the parameter for the i-type scholarly relationship is then calculated as follows:

$$\omega_i = (1 - \omega_1)\frac{\sum_{i \ne l} d_{il}}{\sum_{k>1}\sum_{k \ne l} d_{kl}}, \forall l, k = 1, 2, 3, \ldots, \psi, i > 1 \tag{6}$$

where ω_1, defined as the adjusted parameter of the ACA relationship, is manually set since the proposed method is majorly based on ACA-centered. And the summation of the parameter of all scholarly relationships should be one, i.e. $\sum_{i=1}^{\psi}\omega_i = 1$. Essentially formula (6) reveals that the scholarly relationships similar to others should be assigned as a smaller weight value and those dissimilar to others are featured as a larger weight value.

Step 4: Constructing Hybrid Relationship

The hybrid relationship is then constructed with ψ types of scholarly relationships with their corresponding parameters. Assume that $M_{\delta\times\delta}$ is the hybrid scholarly relationship based on the author cocitation relationship and other $(\psi - 1)$ scholarly relationships. The element m_{kl} of M can be calculated as:

$$m_{kl} = \sum_{i=1}^{\psi} r_{i,k,l} \cdot \omega_i \tag{7}$$

where $i \in \left(1, \psi\right)$, $l, k \in \left(1, \delta\right)$, and $r_{i,k,l}$ refers to the value of the k-th row and the l-th column in the matrix R_i', and ω_i represents the parameter (weight value) of matrix R_i'. Finally, the hybrid scholarly relationship matrix M is used to map the knowledge domains and to do further analyses.

Table 1
"Distance" (d_{ij}) between the scholarly networks.

	R_1	R_2	R_3
R_1 (Author cocitation)	–	0.85	0.21
R_2 (Coauthorship)	0.85	–	0.70
R_3 (Author bibliographic coupling)	0.21	0.70	–

Table 2
Factor analysis results of each combination. A ticked mark (√) in table indicates that the factor can be extracted in the corresponding combination. Bolded factors are those identified by $\mathbf{R}_1$.

	Factor	Combination			
		R_1	R_1+R_2	R_1+R_3	$R_1+R_2+R_3$
1	**Software Engineering**	√	√	√	√
2	Information Management and Systems				√
3	Embedded Systems				√
4	**Databases**	√	√	√	√
5	Information Retrieval				√
6	**Computation, Algorithm, and Data Structure Studies**	√	√	√	√
7	Information Security			√	√
8	Programming Languages		√	√	√
9	**Multimedia Technologies**	√	√	√	√
10	Computer Graphics and 3D Studies				√
11	**Natural Language Processing**	√	√	√	√
12	**Computer Systems and Computer Architecture**	√	√	√	√
13	Distributed Computing				√
14	Computer Hardware				√
15	Electronic Engineering		√	√	√
16	**Communication and Computer Networks**	√	√	√	√
17	**Data Analysis**	√	√	√	√
18	Artificial Intelligence, Machine Learning, and Robotics		√	√	√
19	Applied Computer Science		√	√	√
Number of Factors		**8**	**12**	**13**	**19**

5. Results and discussion

Using on the records from our dataset related to the selected high-impact authors, we construct several networks, i.e. author cocitation networks (R_1), coauthorship networks (R_2), and author bibliographic coupling networks (R_3). Specifically, if two authors' publications were cocited in a certain article, the weight between their ties in R_1 is added; if two authors coworked in certain publication, the weight between their ties in R_2 is added, and; if two authors' works have once cocited others' publication, the weight between their ties in R_3 is added.

Table 1 shows the distance in hyperspace among the scholarly networks, in which the smaller the distance between two networks, the more similar they are. In this example, author cocitation network is similar to author bibliographic coupling networks, and the author bibliographic coupling network is more dissimilar to coauthorship networks compared to author cocitation networks; these confirm Yan and Ding (2012)'s conclusion when they compared the similarities among several scholarly networks. Moreover, the author cocitation network's parameter (w_1) is initially set as 1.0, 0.6, 0.6, and 0.5 in R_1, R_1+R_2, R_1+R_3, and $R_1+R_2+R_3$, respectively, and the weight value for each scholarly network is set based on the result shown in Table 1 and the method provided above.

5.1. Factor analysis

Factor analysis is often used to mine latent relationships between authors in knowledge domain mappings (McCain, 1990). In most of the ACA studies, the identified factors are often regarded as different research specialties (White & McCain, 1998). Zhao and Strotmann (2008b), for example, detected 11 sub-fields (factors) of Library and Information Science (LIS) by factor analysis with 120 most-cited LIS authors in 2001–2005. Jeong et al. (2014) identified 10 and 24 sub-fields (factors) respectively when comparing traditional ACA and their proposed content-based ACA. Table 2 shows four different combinations of factor analysis results applied to our data. The combination of more scholarly relationships can extract more factors, which refer to more sub-fields identified. In other words, combining scholarly networks can detect more second-level disciplines and more details compared with traditional ACA. As shown in Table 2, the factors extracted in R_1 are software engineering; databases; computation, algorithm, and data structure studies; multimedia technologies; natural language processing (NLP); computer systems and computer architecture; communication and computer networks; and data analysis (**bolded** in Table 2). All of the above are core areas of computer science (Brookshear, 2005; Dale, 2015). In $R_1+R_2+R_3$, not only these disciplines are extracted but also newly-formed disciplines are included, such as multi-media

Table 3
The proportion of explained variance in each combination.

Combination	The proportion of explained variance
R_1	80.3%
$R_1 + R_2$	80.9%
$R_1 + R_3$	81.4%
$R_1 + R_2 + R_3$	85.1%

Table 4
Descriptive statistics of the scholarly network combinations.

Network	Density	ACC[a]	Network	Density	ACC
R_1	0.13	0.07	$R_1 + R_3$	0.24	0.16
$R_1 + R_2$	0.26	0.14	$R_1 + R_2 + R_3$	0.43	0.20

[a] Note: ACC = Average clustering coefficient.

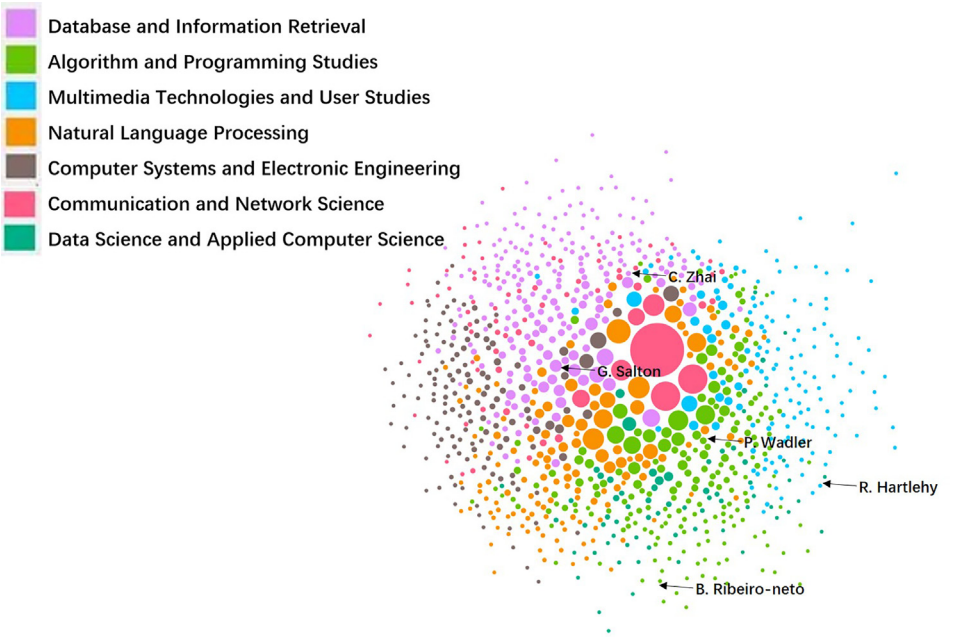

Fig. 6. Network visualization for author cocitation network (R_1). The nodes represent authors. Node sizes are proportional to their average degrees of citations. Nodes within the same color indicate that the authors represented by these nodes could have similar research area and belong to the same cluster. The same below.

studies, computer graphics and 3D studies, as well as distributed computing, in which the last discipline had not emerged until the arrival of the so-called big data era.

The proportion of explained variance of a scholarly relationship usually indicates the performance of factor analysis. According to Table 3, all proportions are above 80%, indicating acceptable performances.

5.2. Network analysis

Table 4 shows the descriptive statistics of different scholarly network combinations, including network density and average clustering coefficient. Note that Barrat, Barthelemy, Pastor-Satorras, and Vespignani (2004)'s algorithm is used to calculate and normalize because all of the networks are weighted instead of binary (zero or one). As shown in Table 4, the density in R_1 is small but the network becomes denser when more scholarly relationships are combined. Meanwhile, the average clustering coefficient is larger when more relationships are considered, hinting the clustering performance could be better in $R_1 + R_2 + R_3$ than that in other combinations. The results can also be reflected in the visualization graphs and the corresponding evaluations (see details in Section 5.3).

The networks constructed by using different scholarly relationships are shown in Figs. 6, 7, 8 and 9, respectively. In these networks, nodes represent authors and node size their sizes are proportional to their average degrees of citations. Additionally, the colors of the nodes are determined by Modularity algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). Nodes within the same color indicate that the authors represented by these nodes could have similar research area

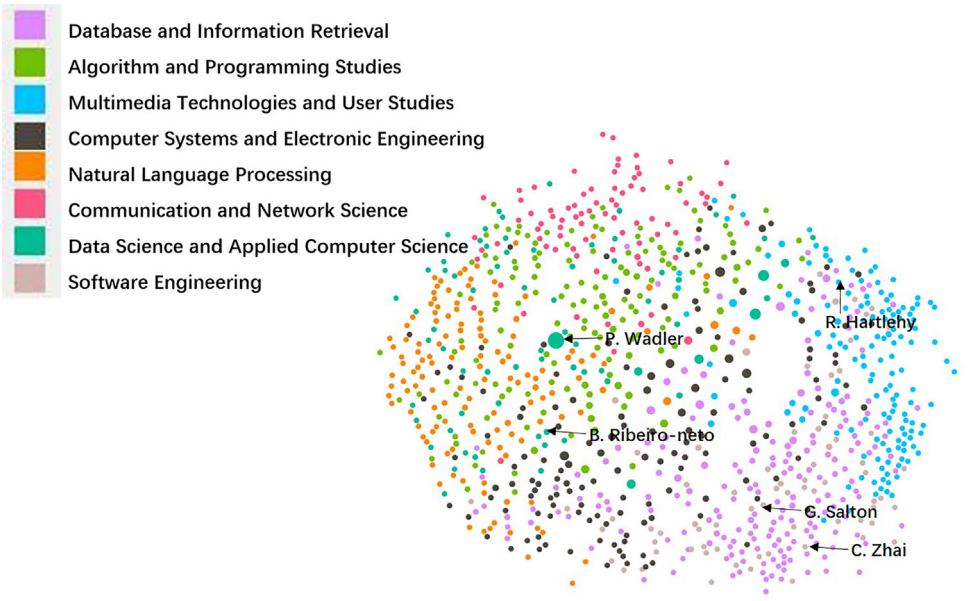

Fig. 7. Network visualization for the combination of author cocitation network and coauthorship network ($\mathbf{R}_1 + \mathbf{R}_2$).

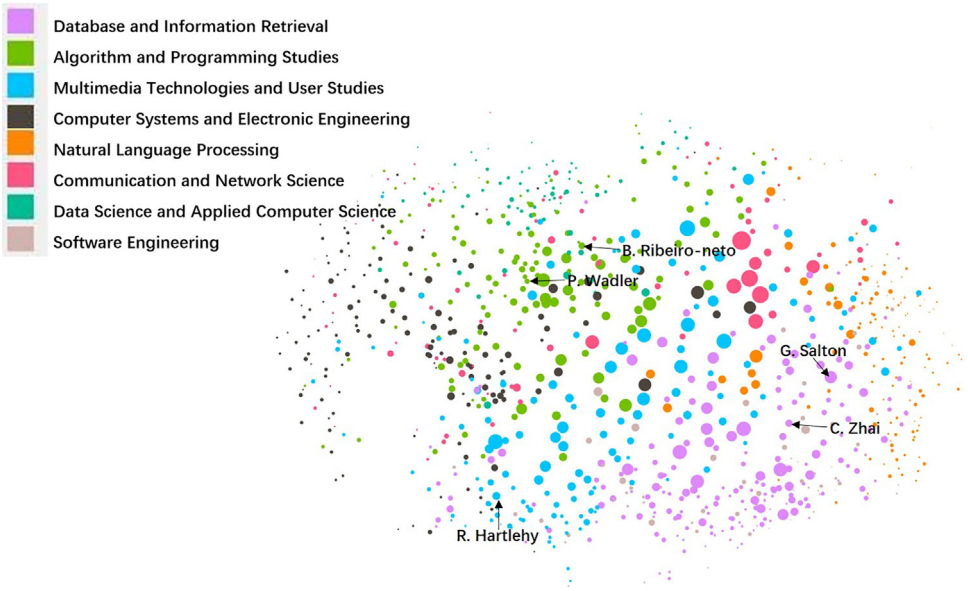

Fig. 8. Network visualization for the combination of author cocitation network and author bibliographic coupling network ($\mathbf{R}_1 + \mathbf{R}_3$).

and belong to the same cluster. The edges are the authors' relationships integrated matrix value (in different network combinations) of the author pair. Then we manually label the clusters by searching the authors' website, Google Scholar, and ResearchGate, as well as reading their academic publications. Comparing these figures, the structures of these networks are similar but there are subtle differences between them. In total, we detect 7, 8, 8, and 11 clusters in Figs. 6, 7, 8 and 9, respectively. The ACA's network in Fig. 6 demonstrates a fundamental role in mapping knowledge domains although not all nodes in it are perfectly clustered and thus can be regarded as the knowledge base of the field. When other kinds of scholarly networks are integrated into ACA we see more clusters emerge, a finding that benefits future endeavors in knowledge domain mapping and the study of the formation of scientific communities. This observation is the same as the results in our factor analysis and also confirms our argument in the Section 3, the relationships of the three networks and their reflected entities, fields, sub-fields, and communities.

Comparing different visualization maps, more sub-fields can be revealed when R_3 are combined with R_1, which illustrates that author bibliographic coupling relationship takes effects in clustering sub-fields based on the clustered bases/outlines in a field by author cocitation relationship, as shown in Figs. 6 and 8, respectively. Some of these sub-fields representing

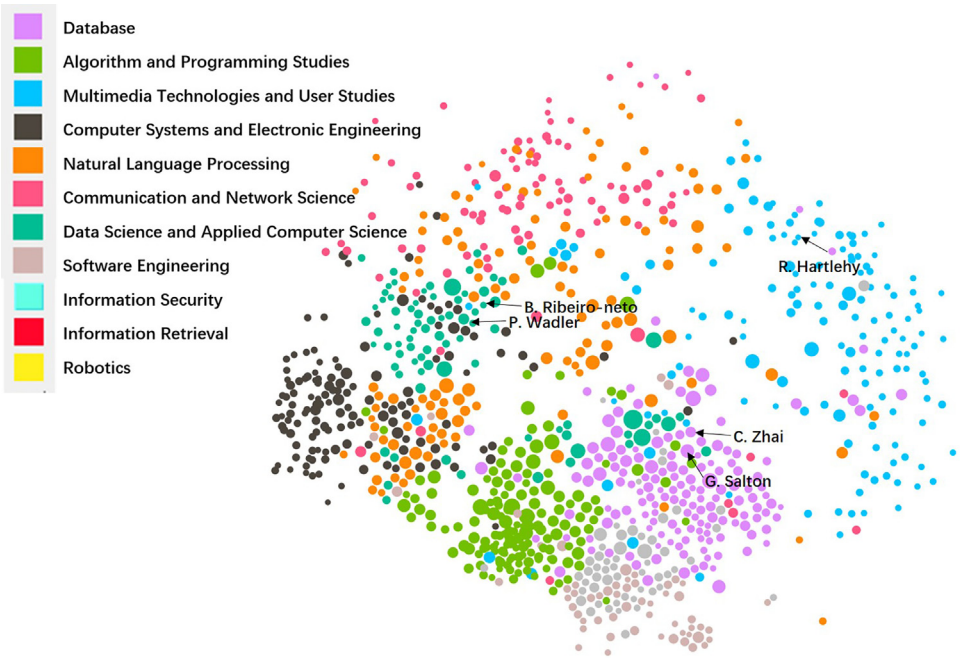

Fig. 9. Network visualization for the combination of three different networks ($\mathbf{R}_1 + \mathbf{R}_2 + \mathbf{R}_3$).

research frontiers (Zhao & Strotmann, 2008a) are clustered more clearly. The phenomena of "sub-fields" could also be identified from the perspective of more larger-size nodes, representing their more centralized statuses in the combined scholarly relationship in $R_1 + R_3$ than in R_1. Actually some authors might not be very impactful or centralized in the larger field, but they could be an active and important member of a sub-field (Zhao and Strotmann, 2008a). Thus, more larger-size nodes could hint more specialized sub-fields under clusters outlined by author cocitation relationship.

Compared with Figs. 8 and 9, we can see that when combining R_2 into $R_1 + R_3$, several small communities based on the sub-fields occur, and more nodes with the same color are clustered nearly in the figures. This phenomenon also confirms our argument in Section 3, in which coauthorship is regarded as a good relationship to mine the communities under sub-fields explored by author bibliographic coupling relationship (Newman, 2001, 2004, 2006; Zhang et al., 2016).

When comparing Fig. 6 with Fig. 8, we can see that the range of the nodes with the same color is getting smaller and the distance between them is shorter. The size of some nodes has become smaller as well. As mentioned in Section 3, the information provided by R_3 indicates that those authors citing common articles have common interest in a period of time. It results in their corresponding nodes having shorter distance in knowledge domain mappings and their relationships influencing the importance of some authors cocited by others (the degree of a node). In other words, some nodes having stronger connections are grouped and a few big nodes in R_1 are reduced. Furthermore, small communities among authors in the same category occur when we combine R_2 into ACA because the structure in R_1 is affected by the solid relationships among the coauthors, as shown in Fig. 7. The results of R_1 and $R_2\langle errorimage\rangle$'s impacts on the network structure when we combine R_1, R_2, and R_3 together are that the nodes with different colors are more separated and those with the same color are closer and forming a few small groups, as shown in Fig. 9. It means that more sub-fields are found and their interdisciplinary relationships can be revealed when we combine more scholarly relationships.

As an example for observing nodes whose corresponding authors have the same research category, we take two researchers, Gerald Salton and Chengxiang Zhai; Dr. Salton is an important founder in Information Retrieval (IR) and Dr. Zhai has many famous studies in IR-based natural language processing (NLP). Their corresponding nodes in these figures are identified in the category of Database and IR and the distance between them shortens when more scholarly relationships are combined. A similar example is that of Drs. Philip Wadler and Berthier Ribeiro-neto, whose common research interests in the semantic web are more strongly revealed when we add more scholarly networks, bringing their nodes closer together. These examples illustrate how combinations of multiple scholarly relationships can group scholars having related research areas closer together in the resulting knowledge domain maps. On the contrary, for nodes whose categories are different, their distance becomes larger when more scholarly networks are combined. For instance, Dr. Berthier Ribeiro-neto focuses on semantic web and IR and Dr. Richard. I. Hartlehy is interested in multi-media technologies and communication engineering. They have fewer intersections, and their corresponding nodes are more distant as more networks are added, as we can see in Fig. 9.

Furthermore, in Fig. 9 we see authors belonging to the NLP category have a spread through a large area of the map, overlapping and mixing with other categories. It means that this category can be regarded as an important field of computer

Table 5
MDS-measurement results of different combinations.

Combination	c	S	σ (%)
R_1	550.86	4010.07	13.74
$R_1 + R_2$	500.43	4209.92	11.89
$R_1 + R_3$	526.04	4622.16	11.38
$R_1 + R_2 + R_3$	436.83	4980.44	8.77

Table 6
Further analyses of MDS-measurement results.

Combinations	Scholarly Relationship	Δc	ΔS	$\Delta\sigma$ (%)
R_1 & $R_1 + R_2$	Coauthorship	50.43	199.85	1.85
$R_1 + R_2 + R_3$ & $R_1 + R_3$		89.21	358.28	2.61
Average Change of "R_2"		**69.82** = (50.43 + 89.21)/2	**279.07**	**2.23**
R_1 & $R_1 + R_3$	Author bibliographic	24.82	612.09	2.36
$R_1 + R_2 + R_3$ & $R_1 + R_2$	coupling	63.60	770.52	3.12
Average Change of "R_3"		**44.21** = (24.82 + 63.60)/2	**693.31**	**2.74**

science and is separated to several sub-fields diffused with others under computer science, such as database systems, applied computer science (e.g., computational linguistics, emotion analysis), information management and information systems, and computer systems. This confirms the conclusion from Nadkarni, Ohno-Machado, and Chapman (2011), in which they argued that the development of NLP diffused much knowledge and research paradigms into diverse fields; on the contrary, these areas also affected and provided many algorithms to NLP. Additionally, although combining multiple scholarly networks inputs more information of perspectives other than cocitation relationship, it could lower the impacts of a few classical authors due to the reduction of the size of a node in the knowledge domain mappings.

5.3. MDS-measurement analysis

Multi-dimensional scaling measurement (MDS-measurement) is exploited to evaluate different graphs quantitatively (Bu et al., 2016), and it can also be regarded as a supplement of network analysis. A good visualization in MDS-measurement often has a higher cohesion in the same category and a high separation in different categories. According to the definition of MDS-measurement, two parameters, c and S, are employed to measure the cohesion and separation, respectively. c is defined as the summation of the distances between nodes within one cluster, while S is the summation of the distances between nodes in different clusters. Higher cohesion within a category (smaller c) and high separation between categories (larger S) indicate a good clustering result. Hence, smaller $\sigma\left(=\frac{c}{S}\right)$ refers to better performance in MDS-measurement. The results of MDS-measurement among different combinations are shown in Table 5. We see that c and σ decrease and S increases when we combine other scholarly relationships into ACA. This indicates the visualized tendency of these networks – the "distance" between authors in the same category becomes smaller and smaller, while that between authors in distinct categories becomes larger and larger.

The differences of c, S, and σ in different combinations, respectively noted as Δc, ΔS, and $\Delta\sigma$, between combinations are calculated when only differing from one scholarly network, as shown in Table 6. The average of Δc when adding coauthorship relationship is larger than that adding author bibliographic coupling relationship. Besides, the average of ΔS is larger when adding author bibliographic coupling relationship. This shows that coauthorship plays a role of bringing similar authors closer in knowledge mappings – corresponding to its role that form more *communities*; on the contrary, author bibliographic coupling relationship makes dissimilar authors farther in knowledge domain mappings – corresponding to its role in forming more *sub-fields* based on ACA's fundamental knowledge structures. This observation is the same as the results in Section 5.2, of which the distances of the nodes with the same color in $R_1 + R_2$ is smaller than that of $R_1 + R_3$ and this result is contrary when the nodes belong to different categories. It also confirms our prior arguments that clarify the differences between ABCA and CA in mapping knowledge domains in Section 3.

6. Conclusions

A method combining distinct types of scholarly relationships into ACA is proposed to integrate more diverse information in mapping knowledge domains and explore more relationships in different perspectives. The major difference between the method and traditional ACA is the raw matrix constructed by using symmetric Kullback-Leibler (sKL) divergence and Exploratory Factory Analysis (EFA) when combining these scholarly relationships. According to our empirical studies, the results of the proposed method provide more detailed information in knowledge domain mappings and performs better than traditional ACA in factor analysis and network analysis. Specifically, the nodes within the same category lie nearer while nodes in different categories lie farther, and more sub-fields can also be discovered when mining nuance of the given field. As a result, the proposed method can be regarded as another option to understand researchers and to map knowledge domains in a scientific field.

Besides the method itself, this article explores more deeply the relationships between author cocitation, author bibliographic coupling, and coauthorship relationships in terms of mapping knowledge domains. As a supplement of Zhao and Strotmann (2008a, 2014) who argued that ACA is regarded as a "look back" while ABCA is considered as a present view, our article implies more nuance between these scholarly networks. Specifically, focusing more on impactful authors, ACA explores basic clusters and helps form "fields" in depicting scientific intellectual structures. ABCA is more detailed, caring more about the most prolific author and revealing "sub-fields" in a domain. CA provides a more fine-grained perspective of "community" under the level of a "sub-field". As pointed out above, such small "communities" form and strengthen the clusters of "sub-fields", which facilitates knowledge domain maps to be more accurate and reliable (Newman, 2001, 2006).

For future researchers who hope to replicate this method, we do not always recommend them to combine as many scholarly networks as possible. Since different networks play distinct roles in mapping knowledge domains, researchers are supposed to choose the network(s) that best fit their requirements. For instance, if their research goal is to draw a general map concerning rough vision, an author cocitation network or a combination between author cocitation and author bibliographic coupling networks are sufficient as their input. However, if detailed understandings and thorough examinations are necessary, coauthorship network should also be included to detect the communities under sub-fields. Another issue that needs to be considered is the data availability, as argued by Zhao and Strotmann (2008a). In particular, coauthorship network requires information for all authors, but the other two networks need only the first authors' information when we use the "pure-first author-cocitation/author-bibliographic-coupling analysis" (Rousseau & Zuccala, 2004). If all authors' information is not available or it is available but with high cost (e.g., author name disambiguation), then combining these networks with the coauthorship network might be impractical.

This article has several limitations and many related studies could therefore be implemented in the future. First of all, this article simply uses the first authors' instead of all authors' information, which might cause some inaccuracy to the results (Zhao, 2006; Zhao & Strotmann, 2008b). Nevertheless, this is acceptable because using the first authors' information still demonstrates the performance and applicability of our proposed method, and showcases its improvements over traditional ACA. Meanwhile, from the perspective of combining scholarly networks, citation, co-word, and topical networks (Yan & Ding, 2012) could also be input into traditional ACA. Specifically, references' keyword information could be used when we try to construct co-word networks; and topic modeling (Blei, Ng, & Jordan, 2003) could be used to form topical networks. Moreover; the relationship between author cocitation and coauthorship networks could be mined according to author rank (sequence) in articles so that the weights of the first authors and the corresponding authors should increase when future algorithms are designed. On the other hand; documents; affiliations; and research areas could also be used as the study objects instead of only authors.

Acknowledgements

The authors would like to thank Dr. Erjia Yan and Mr. Dakota S. Murray for their kind suggestions on this article. We are also very grateful for the constructive comments from three anonymous reviewers and the editor-in-chief of *Journal of Informetrics*, Dr. Ludo Waltman.

Author contributions

Yi Bu: Conceived and designed the analysis, Contributed data or analysis tool, Performed the analysis, Wrote the paper.

Shaokang Ni: Conceived and designed the analysis, Wrote the paper.

Win-bin Huang: Conceived and designed the analysis, Contributed data or analysis tool, Performed the analysis, Wrote the paper.

References

Ahlgren, P., Jarneving, B., & Rousseau, R. (2004). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, *55*(9), 550–560.

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.

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).

Beaver, D. D., & Rosen, R. (1979). Studies in scientific collaboration, Part III: Professionalization and the natural history of modern scientific coauthorship. *Scientometrics*, *1*(3), 231–245.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

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.

Boyack, K. W., & Klavans, R. (2010). Cocitation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404.

Brookshear, J. G. (2005). *Computer science: An overview*. Boston, Massachusetts: Addison Wesley.

Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm, & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99–136). Washington, DC: American Psychological Association.

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

Callon, M., Courtial, J. P., & Turner, W. A. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Sciences Information*, *22*(2), 191–235.
Cattell, R. B. (1965). Factor analysis: An introduction to essentials (II): The role of factor analysis in research. *Biometrics*, *21*, 405–435.
Chen, L. C., & Lien, Y. H. (2011). Using author co-citation analysis to examine the intellectual structure of e-learning: A MIS perspective. *Scientometrics*, *89*(3), 867–886.
Chu, K. C., Liu, W. I., & Tsai, M. Y. (2012). The study of co-citation analysis and knowledge structure on healthcare domain. In *Proceedings of the sixth global conference on power control and optimization* (pp. 247–253).
Dale, N. B. (2015). *Computer science illuminated* (5th ed.). Beijing: China Machine Press.
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.
Egghe, L., & Leydesdorff, L. (2009). The relation between Pearson's correlation coefficient *r* and Salton's cosine measure. *Journal of the American Society for Information Science and Technology*, *60*(5), 1027–1036.
Eom, S. (2008). All author co-citation analysis and first author co-citation analysis: A comparative empirical investigation. *Journal of Informetrics*, *2*(1), 53–64.
Garfield, E., Sher, I., & Torpie, R. J. (1964). *The use of citation data in writing the history of science*. Philadelphia, Pennsylvania: Institute for Scientific Information.
Jeong, Y.-K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics*, *8*(1), 197–211.
Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–200.
Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25.
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.
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.
Ma, R., & Ni, C. (2012). Author coupling analysis: An exploratory study on a new approach to discover intellectual structure of a discipline. *Journal of Library Science in China*, *38*(2), 4–11.
McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, *41*(6), 433–443.
McCain, K. W. (1991). Mapping economics through the journal literature: An experiment in journal co-citation analysis. *Journal of the American Society for Information Science*, *42*(4), 290–296.
Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago, Illinois: University of Chicago Press.
Milojević, S. (2010). Modes of collaboration in modern science: Beyond power laws and preferential attachment. *Journal of the American Society for Information Science and Technology*, *61*(7), 1410–1423.
Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551.
Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 404–409.
Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5200–5205.
Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United Stated of America*, *3*(23), 8577–8582.
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.
Small, H., & Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science Studies*, *4*(1), 17–40.
Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 990–998).
Van Eck, N. J., & Waltman, L. (2008). Appropriate similarity measures for author co-citation analysis. *Journal of the American Society for Information Science and Technology*, *59*(10), 1653–1661.
Wang, H., Ding, Y., Tang, J., Dong, X., He, B., Qiu, J., et al. (2011). Finding complex biological relationships in recent PubMed articles using Bio-LDA. *PLoS One*, *6*(3), e17243.
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 co-citation analysis of information science (1972–1995). *Journal of the American Society for Information Science*, *49*(4), 327–335.
White, H. D. (2004). Author cocitation analysis and Pearson's *r*: Reply. *Journal of the American Society for Information Science and Technology*, *55*(9), 1250–1259.
Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks related to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313–1326.
Zhang, C., Bu, Y., & Ding, Y. (2016). Understanding scientific collaboration from the perspective of collaborators and their network structures. In *iConference 2016 partnership with society*.
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). 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), 995–1006.
Zhao, D. (2006). Towards all-author co-citation analysis. *Information Processing and Management*, *42*(6), 1578–1591.
Zhao, D. (2012). Bibliometric approach about mapping knowledge domains. *Library and Information Service*, *56*(6), 107–110.
Zitt, M., Bassecoulard, E., & Okubo, Y. (2000). Shadows of the past in inter-national cooperation: Collaboration profiles of the top five producers of science. *Scientometrics*, *47*(3), 627–657.