Scientometrics (2016) 108:143–166
DOI 10.1007/s11192-016-1959-5

MACA: a modified author co-citation analysis method combined with general descriptive metadata of citations

Yi Bu[1] · Tian-yi Liu[1] · Win-bin Huang[1]

Received: 5 November 2015 / Published online: 3 May 2016

Abstract Author co-citation analysis (ACA) is a well-known and frequently-used method to exhibit the academic researchers and the professional field sketch according to co-citation relationships between authors in an article set. However, visualizing subtle examination is limited because only author co-citation information is required in ACA. The proposed method, called modified author co-citation analysis (MACA), exploits author co-citation relationship, citations published time, citations published carriers, and citations keywords, to construct MACA-based co-citation matrices. According to the results of our experiments: (1) MACA shows a good clustering result with more delicacy and more clearness; (2) more information involved in co-citation analysis performs good visual acuity; (3) in visualization of co-citation network produced by MACA, the points in different categories have far more distance, and the points indicating authors in the same category are closer together. As a result, the proposed MACA is found that more detailed and subtle information of a knowledge domain analyzed can be obtained, compared to ACA.

Keywords Author co-citation analysis · Co-citation analysis · Citation analysis · Bibliometrics

Mathematics Subject Classification 68T30

JEL Classification D830

✉ Win-bin Huang
huangwb@pku.edu.cn

[1] Department of Information Management, Peking University, Beijing, China

Introduction

Co-citation analysis (CA) is a significant branch of citation analysis in bibliometrics. It can be divided into at least three types according to the object of study: author co-citation analysis (ACA), document co-citation analysis (DCA), and journal co-citation analysis (JCA). H. D. White and B. C. Griffith brought ACA into Library and Information Science (LIS) in 1980s (White and Griffith 1981) in order to depict the intelligent domain of certain field(s). The main purpose of ACA is to map scientific domains from the perspective of co-cited authors by pointing out the co-citation relationships in which the object of study (i.e. the unit of analysis) is author rather than document or journal (Jeong et al. 2014). The basic assumptions of ACA can be summarized as: all cited articles play equal roles in co-citation analysis; the more two authors are co-cited, the stronger their relevance is. Moreover, four normal steps of ACA are listed as followings (McCain 1990; Eom 2008a): (1) selection of author set and retrieval of co-cited author counts; (2) forming the raw co-citation matrix; (3) transformation from the raw co-citation matrix to the correlation matrix; (4) multivariate analyses (e.g., cluster analysis, multi-dimensional scaling (MDS), factor analysis, etc.). The concepts and methods of ACA were applied frequently in other majors to exhibit scientific domains and academic researchers (Eom 1999; Tsay 2011). Recently, ACA has been further combined with content-based analysis (Jeong et al. 2014) and artificial intelligence technologies (An et al. 2011).

However, it is assumed that each citation in an article has equal contribution according to White and Griffith (1981). It could not reveal significance and relevance because the purpose of these citations could be different in citers' perspective. For example, the article, named "PageRank for ranking authors in co-citation networks" (Ding et al. 2009), has two references, coauthorship-related one (Liu et al. 2007) and PageRank-related one (Bianchini et al. 2005) with the corresponding authors, Liu and Binanchini. In fact, the authors have different interest fields, Library and Information Science (LIS) and Computer Science (CS), though their studies are co-cited. The author, Dr. Binanchini, could appear in citation networks (graph) obtained in multivariate analyses while LIS is considered. This might cause an oversight to explore the potential authors in LIS if lots of such situations occurred. In other words, its performance has been accepted and tolerated despite the fact that ACA uses author co-citation relationships as its unique information to construct a knowledge domain. And the major purpose of this paper is to reduce the oversight by involving more general information in citations based on ACA. The information can be general descriptive metadata of a citation, such as published time, the publication itself, and keywords of a citation. Specifically, in time perspective, for example, small difference between two citations' published time implies that the authors tend to focus on similar issues in the same period of time. The representation of authors' relationship might be distinctive in knowledge graph because of various concepts, methods, or even diversified demands in different periods of time. Similar journals where two authors' papers are published or similar keywords of citations they use, on the other hand, implies that they tend to research on similar issues.

As a result, the proposed method, called Modified Author Co-Citation Analysis (MACA), exploits four general descriptive metadata in citations, authors of a citation, the time when a citation is published, the carrier (i.e. journals, conferences, monographs, and even electronic sources, etc.) where a citation is published, and the keywords of a citation, to construct a citation network. Similar to ACA, the information of authors in citations (i.e. author co-citation count) is used to establish the co-citation relationships among authors.

The import of published time information in citations to every co-cited author is produced to form the co-citation matrix from time perspective, called time-based parameter. The carrier information of citations is abstracted first and their professional fields belonged are developed according to the focused issues. The relationship of co-cited authors, called carrier-based parameter, is calculated depending on the similarity of professional fields of their articles. Similarly, the professional fields to which keywords of citations belong are obtained initially based on the meaning of keywords. Fields calculate the co-cited authors' relation in keyword perspective, called keyword-based parameter by fields.

Related works are described in "Related works" section. The calculations and explanations of the proposed MACA are detailed in "Modified author co-citation analysis (MACA)" section. The dataset and pre-processing of our studies are expressed and the performance and analysis of the proposed MACA are demonstrated in "Experimental results and discussion" section. Finally, the conclusions are provided in "Conclusion" section.

Related works

ACA has been a hotspot in informetrics and scientometrics, which aims to instruct scientific research by looking for co-citation relationships between authors in academic articles set and mapping knowledge domains (McCain 1990). Much empirical research indicated that ACA is very effective and applicable in evaluating discipline development situations and identifying micro-structures of certain field and its sub-fields since it can reveal dynamic changes and future developments.

The major steps of ACA are shown in Fig. 1. An academic dataset is selected by using certain methods (e.g. selection of specific journal(s), snowballing, etc.) and the author's name should be disambiguated in the first two steps. Author name disambiguation mainly bases on the authors' affiliation, collaboration records, and research areas. Then the co-cited authors within a dataset are abstracted to construct a raw co-citation symmetric matrix based on their co-citation count regardless of whether the first-author or all-author information is counted. The raw co-citation matrix is transformed into a correlative co-citation matrix for normalization in the next step. Many correlation measurements (e.g. Pearson's r, Jaccard, cosine, Euclidean distance, etc.) should be judged and selected in this step. The final series of data analysis methods (e.g., factor analysis, cluster analysis, network analysis, and multidimensional scaling) are used to produce a more accurate interpretation of the results. For example, when trying to cluster given authors, a hierarchical agglomerative or iterative partitioning method is adopted to analyze the correlating authors. Then professionals provide some explanations based on the results before peer reviewing.

Over 30 years, four major concerns of traditional ACA can be summarized as followings: (1) Data collection methods (White and McCain 1998; Cothill et al. 1989) and database selection (Zhao and Strotmann 2008); (2) Raw matrix formation and definition or

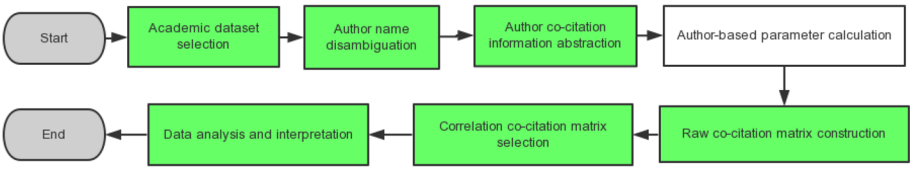

Fig. 1 The framework of ACA

modification of ACA; (3) Correlation matrix transformation and similarity measurement in ACA (Ahlgren et al. 2003; White 2003a; Bensman 2004; Leydesdorff and Vaughan 2006; Egghe 2009; Mêgnigbêto 2013); (4) Further analysis methods (e.g. factor analysis, multi-dimensional scaling, cluster analysis, network analysis, etc.) and visualization (White 2003b; An et al. 2011; Chen 1999; Moya-Anegón et al. 2007).

In the method of raw matrix formation and definition or modification of ACA, researchers focus on diagonal values in the raw co-citation matrix (White and McCain 1998; McCain 1991) and first- or all-author co-citation analysis (Persson 2001; Zhao and Logan 2002; Zhao 2006; Rousseau and Zuccala 2004; Zhao and Strotmann 2008; Schneider and Larsen 2009; Eom 2008b). The latter research has made traditional ACA more informative since more authors' co-citation relationships were imported. However, these studies only focused on author-related information instead of other available metadata in citations. Moreover, some researchers studied on content-based ACA. Jeong et al. (2014), for example, tried to use the similarity of citance (i.e. citing sentences) to modify traditional ACA, the essence of which is to improve the step of the raw co-citation "count" calculation. The results showed that content-based ACA performed better than the previous methods. Nevertheless, content-based ACA requires full-text data in TXT or HTML format and more calculative complexity. Concerning these disadvantages, in this paper, we hope to modify the construction of raw co-citation matrix combined with other citation descriptive metadata (i.e., citations' published time, citations' published carrier, and citations' keywords) in order to integrate more types of information and to improve the performance of ACA. This paper tries to modify traditional ACA by adding an "author-based parameter calculation" step (white block in Fig. 1).

Modified author co-citation analysis (MACA)

The framework of the proposed MACA, which analyzes the relationship of two authors by using general descriptive metadata of citations including the published time, keywords, and carrier, is shown in Fig. 2. Obviously, the major difference between ACA and MACA is the stage of constructing raw co-citation matrix. The authors' names, published time, carriers and keywords of each citation should be abstracted in the first stage. The co-citation matrix of MACA is then constructed by four matrices, called author-based parameter, time-based parameter, carrier-based parameter, and keyword-based parameter,

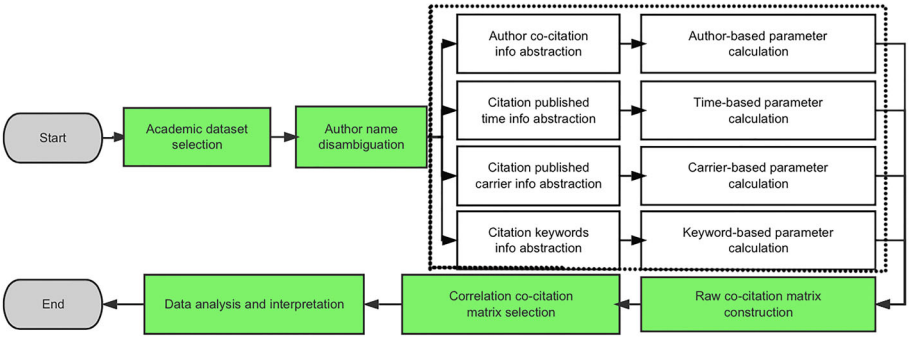

Fig. 2 The framework of the proposed MACA

based on the four kinds of corresponding descriptive metadata, respectively. Note that in Fig. 2, the white blocks refer to new steps we introduce, while the green blocks mean traditional steps. The calculations of three different parameters and the co-citation matrix are detailed in the following.

Calculation of the time-based parameter between two authors

An academic article usually exposes the research interest, professional field, and specific contribution of an author. The published time of an article may also implicitly show the authors' research period on this work. According to the observation of general academic research procedure, the researchers usually read literatures first and formulate their problem inside the studies, then looked for the current solutions or algorithms related to their problems. The researchers, especially in engineering field, cite recent studies for exploiting, modifying, or comparing. It simply implies that two authors' works could be related, cooperated, or continued while the published time of their articles, especially co-cited by an article, is near.

Nevertheless, the purpose of the citations in an article, more often than not, could be different, and they might belong to different professional fields (Bu et al. 2015; Brooks 1985). For example, a mathematic theory proposed in a citation is cited for conducting an algorithm, and the method of another citation belonged to bibliometrics is cited for evaluating its results. The analytical result of author co-citation combined with the calculation of their published time could not be influenced while the analysis in a specific field is mainly considered. However, the authors belonging to different professional fields would actually be shown obviously in the knowledge graph. In other words, the relationship between authors of two citations within similar published time should be reflected on the knowledge graph if their studies are in the similar research field.

Three academic researchers within their professional fields are indexed and shown in Table 1. The histogram of the number of pairwise authors' are co-cited according to their time difference as demonstrated in Fig. 3 as well. The distributions of the pairwise authors, 1 and 2, 2 and 3, 1 and 3, are drawn as a solid line, placing a circular, triangular, and square markers at the data points, respectively. Obviously, a total of 36 articles are co-cited, and 72 % have less than 3-year difference. These articles, closed at the published time, have similar or related issues in network science after examining them artificially. The similarity is also revealed in the observation of other pairwise authors. Moreover, there are not many co-citations with more than a 5-year difference, and one of them could be a literature review or a classic study in a professional field. It implies that the interest field of the authors might be related in the same period while their articles having only a small difference in published time are co-cited. In other words, the authors having a number of co-citations with small differences in published time can have closer positions on the knowledge graph.

Table 1 Three authors and their area of interests

Author	P. Ahlgren	A. Barabási	J. Bar-llan
Index	1	2	3
Area of interests	Text mining/NLP/data analysis/ sentiment analysis	Network science/statistic physics/biophysics	Internet research/network science/informetrics

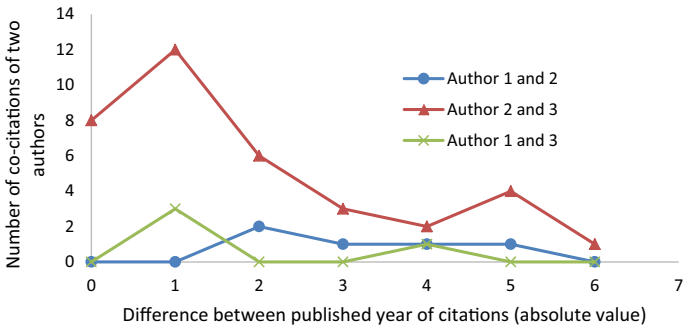

Fig. 3 The distribution of the number of pairwise authors' co-citations to their time difference

As a result, two basic assumptions on calculation of time-based parameter between two authors show the following: (1) A small difference between two citations' published time implies that authors tend to study similar issues in the same time period. (2) An obvious difference between them refers that though authors may research in similar issues in different periods of time, the representation of the authors' relationship should be distinct in the knowledge graph because of various concepts, methods, or even diversified demands in different periods. Therefore, the time-based parameter of MACA indicates the quantity of relationship in time dimension between two authors whose works are co-cited in one or more articles. Figure 4 shows the block diagram to calculate the time-based parameter. At first, all referred papers made by author A_i are collected respectively. Then the published time of all papers cited are extracted and inputted to time-based relation calculator. After that, the time-based parameter of MACA will be produced.

In time-based relation calculator, assume that an article, P_l and $l \in [1, n]$, has the references, D_r and $r \in [1, m]$, with their authors, A_i and $i \in [1, I]$, and their published year, t_r and $r \in [1, m]$. Then the average of published time of an author A_i in the article P_l is

$$\text{Pub\_ave\_}t_{A_i,P_l} = \frac{1}{m}\sum_{r=1}^{m} t_r \tag{1}$$

The time-based parameter of two authors is calculated by all average of published time of two authors, A_i and A_j, in all n articles and shown as

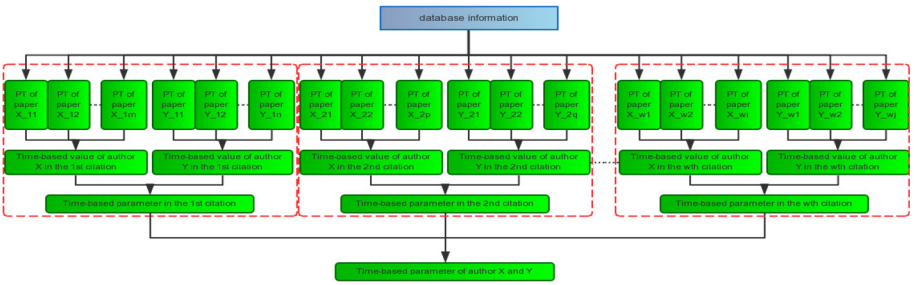

Fig. 4 The procedure of calculating time-based parameter in MACA (*PT* published time)

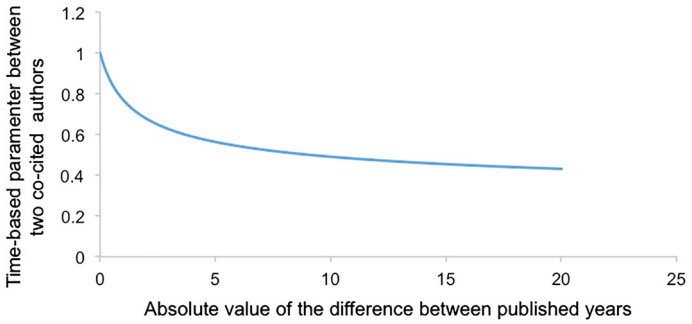

Fig. 5 The value of time-based parameter

$$\text{Ave\_FTR}_{A_i,A_j} = \frac{1}{n}\sum_{l=1}^{n}\left(1+\ln\left(1+\left|\text{Pub\_ave\_}t_{A_i,P_l}-\text{Pub\_ave\_}t_{A_j,P_l}\right|\right)\right)^{-1} \quad (2)$$

where $\text{Pub\_ave\_}t_{A_i,P_l}$ and $\text{Pub\_ave\_}t_{A_j,P_l}$ are average published time of two co-cited authors, A_i and A_j, in the same article, P_l. Then the range of Eq. (2) is $[0, 1]$ with its domain $[1, +\infty)$ and is shown in Fig. 5. Apparently, it reaches the maximum value 1 when $\text{Pub\_ave\_}t_{A_i,P_l}$ is equal to $\text{Pub\_ave\_}t_{A_j,P_l}$, and it is closer to 0 if the difference is larger enough. Note that this design has two advantages: (1) The function can easily reflect the citation relationship in time dimension between two authors; (2) It can simply be merged into the calculation of traditional ACA for normalizations because its range is $[0, 1]$.

For example, Table 2 shows that an article X has four references, D_1, D_2, D_3, and D_4, with their authors, A, A, B, and B, respectively. According to Eq. (2), the time relation of each author in the references can be calculated as $\left(1+\ln\left(1+\left|\frac{1990+2002}{2}-\frac{2010+2010}{2}\right|\right)\right)^{-1} \approx 0.27$. Suppose that the two authors are also co-cited in another two articles with their time correlation 1.00 and 0.59, respectively. Then their time-based parameter is $(0.27 + 1.00 + 0.59)/3 = 0.62$.

Calculation of the carrier-based parameter between two authors

A carrier here is defined as a form of a publication, such as journals, conferences, magazines, books, electronic resources, etc. Carriers often have specific concentrations in a professional field because they typically dedicate a specific group of readers. The articles in a carrier usually have similar issues and characteristics, such as special issues, special columns, or distinguishing themes, etc. Authors also would like to publish their studies in the carrier in which focused topics are matched. In other words, authors of similar or related fields are co-cited when their articles are published in the same or field-related carriers.

For example, three major topics are discussed, information retrieval and technology, Internet information and information searching behavior, citation analysis and term co-occurrence research, after analyzing all articles in 1999–2008 on *Journal of the American Society for Information Science and Technology*[1] (JASIST) (Li and Gong 2010). A. Spink,

[1] Its name was *Journal of the American Society for Information Science* before 2001. Nevertheless, this journal changed its name to *Journal of the Association for Information Science and Technology* in 2014.

Table 2 Examples of four papers and their published time

(Paper, author)	Published year	(Paper, author)	Published year
(D_1, A)	1990	(D_3, B)	2010
(D_2, A)	2002	(D_4, B)	2010

a famous scientist in information retrieval, published 22 articles, 1.205 % of the whole papers, on JASIST from 2001 to 2010 (Yang 2013). Obviously, the author's particular interest is included in the scope of JASIST. Meanwhile, another author in the carrier might have similar research field to Dr. Spink if their studies are co-cited. A similar example lies in Y. Ding publishing many of her articles on *Journal of Informetrics* and *Scientometrics* (Ding 2011; Ding et al. 2000, 2013). However, the disciplines of the authors could be discerned even if their articles published in different kind of carriers are co-cited because their cited purposes are distinct. Again, we take the example of "PageRank for ranking authors in co-citation networks" (Ding et al. 2009) in which two papers, namely "Inside PageRank" (Bianchini et al. 2005) and "Co-authorship networks in the digital library research community" (Liu et al. 2007), were co-cited. The journal of the former paper is *ACM Transactions on Internet Technology*, obvious a journal in CS, while that of the other paper is *Information Processing and Management*, a typical LIS journal. Meanwhile, their authors belong to corresponding fields as well.

Thus, three basic assumptions on calculation of carrier-based parameter between two authors lie on the followings: (1) The information carriers have their specific knowledge range, even though cross-disciplinary sources have strong pertinence. As a result, the knowledge range of information carriers can be cataloged and indexed according to their research areas. (2) The papers published in particular carriers are relative to some extent because the carriers usually focus on particular issues or have given features. (3) The authors would normally submit and publish their articles in information carriers whose concerns are matched with the directions of their studies. Therefore, carrier-based parameter of MACA indicates the quantity of relationship in information carrier perspective between two authors whose works are co-cited in one or more articles. Figure 6 shows the block diagram to calculate the carrier-based parameter. At first, all referred papers made by the author A_i are collected respectively. All information carriers of the papers cited are extracted and are given indexes in field indexer according to their focus areas. Then the field indexes of all papers cited are computed and inputted to carrier-based relation calculator. After that, the carrier-based parameter of MACA will be produced.

In carrier-based relation calculator, suppose that there are K distinct information carriers in dataset, which are divided into ξ different professional fields. An article, P_l and $l \in [1, n]$, has the references, D_r and $r \in [1, m]$, with their authors, A_i and $i \in [1, I]$, and their information carrier, c_q and $q \in [1, K]$. A field distribution matrix, showing the field relation of a reference D_r and its author A_i in article P_l, is formulated as $F = \left(f_{l,i,j,r}\right)$

$$f_{l,i,j,r} = \begin{cases} 1, & c_q \text{ of } D_r \text{ with } A_i \text{ is related to } j\text{th field} \\ 0, & c_q \text{ of } D_r \text{ with } A_i \text{ is not related to } j\text{th field} \end{cases} \tag{3}$$

where $j \in [1, \xi]$ is the field index, and the field relation, FR of an author A_i in article P_l on field j is further defined as

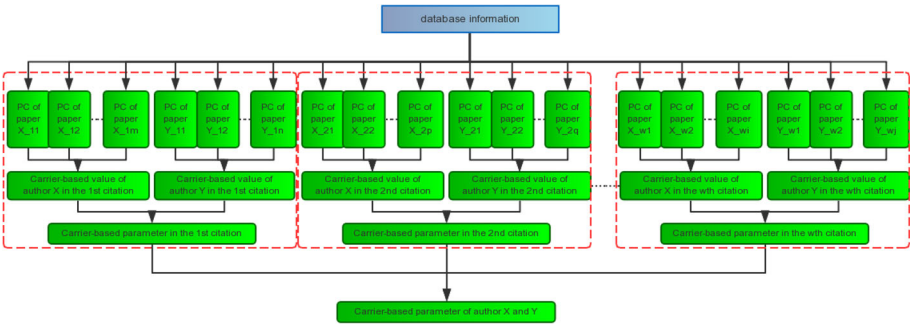

Fig. 6 The procedure of calculating carrier-based parameter in MACA (*PC* published carriers)

Table 3 Examples of four papers and their field distributions

(Paper, author)	F_1	F_2	F_3	F_4	F_5
(D_1, A)	0	0	0	1	0
(D_2, A)	0	0	1	0	0
(D_3, B)	0	0	1	0	0
(D_4, B)	0	0	0	0	1

$$\mathrm{FR}_{A_i,P_l j} = \begin{cases} 1 & \text{if } \sum_{r=1}^{m} f_{l,i,j,r} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Then the field correlation between two co-cited author, A_i and A_z, in article P_l is calculated by

$$\mathrm{FCR}_{A_i,A_z,P_l} = \frac{1}{\xi}\sum_{j=1}^{\xi} \mathrm{FR}_{A_i,P_l j} \cdot \mathrm{FR}_{A_z,P_l j} \tag{5}$$

Therefore, the carrier-based parameter of the two authors within the range $[0, 1]$, shown in Eq. (6), is the average of their field correlation in the dataset.

$$\mathrm{Ave\_FCR}_{A_i,A_z} = \frac{1}{n}\sum_{l=1}^{n} \mathrm{FCR}_{A_i,A_z,P_l} \tag{6}$$

For example, Table 3 shows that an article X has four references, D_1, D_2, D_3, and D_4, with their authors, A, A, B, and B, respectively. And these references are belonged to F_4, F_3, F_3, and F_5, individually in total five different professional fields. According to Eq. (4), Table 4 indicates the research field relation of each author in the references. After calculation, the two authors' field correlation is 1. Suppose that the two authors are also co-cited in other two articles with their field correlation 2 and 1, respectively. Then their carrier-based parameter is $[(1 + 2 + 1)/5]/3 \approx 0.27$.

Table 4 The field relation of two authors in article *X*

Author	F_1	F_2	F_3	F_4	F_5
A	0	0	1	1	0
B	0	0	1	0	1

Calculation of the keyword-based parameter between two authors

Generally, keywords in an article are usually important access points relevant to readers' interests and authors' studies. There are several ways of choosing keywords in writing academic papers, and fitting into the categories that have already been prescribed by the journal's "instruction to authors" would be a possible method of choosing keywords. Keywords are sometimes generated automatically by the library information systems at proof stage (Hartley 2008). According to Hartley, these keywords are selected from the following series of suggested categories: the discipline (e.g., economic, computer science, mathematics), methods (e.g., experiment, case study, questionnaire, algorithm), data source (e.g., primary, secondary, library), location (e.g., city, institution), or topic (e.g., information security, image processing, nature language processing). Due to the limitation of the number of keywords, some researchers have to judge and weigh between keywords. In most cases, keywords often orientate the main professional field of an article and they might expose the authors' interests in an academic field. As a result, the interested fields of two authors, whose articles are co-cited, might be correlated while the keywords of the articles belong to closer professional fields.

For example, keywords of all articles published by Y. Ding, a productive researcher in LIS, on JASIST before 2015 and the number of them used in her articles are shown in Table 5. Note that keywords of her partial articles are not found in the PDF version and those added by *Web of Science* system are automatically selected in our observation. Some generalized keywords, like "science", "library", "time", etc., are deleted here. These keywords can be roughly divided into eight parts: citation analysis, bibliometrics/scientometrics, social networks, knowledge management, topic modeling, semantic web, scientific collaboration, and scientific evaluation. The partitions obviously indicate Dr. Ding's interests and professional fields. After examining the statements on her website,[2] we found that her interest fields includes semantic web, healthcare, social network, citation analysis, knowledge engineering, and information retrieval. These partitions of keywords are matched with her interest fields except for "healthcare" because it is not involved in her articles published on JASIST. Apparently the more keywords collected could reflect the interest fields of an author. Moreover, L. Bornmann, a well-known sociologist of science, was co-cited with Y. Ding many times, for instance, "Generalized preferential attachment considering aging" (Wu et al. 2014). In that paper, (Ding et al. 2013) and (Bornmann and Daniel 2008) were co-cited. The keywords of the former paper are "content-based citation analysis", "citation", "mentioning", and "citation analysis", while those of the other paper are "reference services", and "bibliometrics systems". In this case, the articles of both authors, Y. Ding and L. Bornmann, may be related to bibliometrics/scientometrics, and presumably they should have similar interests, i.e. bibliometrics/scientometrics. After examining his personal website,[3] this assumption is correct—the area of L. Bornmann

[2] This is the URL of Y. Ding's personal website: http://info.ils.indiana.edu/~dingying.

[3] This is the URL of L. Bornmann's personal website: http://www.lutz-bornmann.de.

Table 5 All keywords of Y. Ding's articles published on JASIST (*NK* the number of keywords used)

Keywords	NK	Keywords	NK	Keywords	NK
Algorithm	2	Documents	1	Network analysis	1
Author citation	1	Eigenfactor	2	Networks	1
Author co-citation analysis	1	Evaluation	1	Neural-network research	1
Authors	1	Folksonomies	1	North-American library	1
Betweenness centrality	1	Graph	1	Pagerank algorithm	5
Bibliographic citations	1	*h*-index	1	Patents	1
Bibliographic coupling	1	Impact	2	Patterns	1
Bibliometrics	1	Impact factor	1	Performance	1
Centrality	1	Index	1	Power	1
Citation	1	Information science	1	Primary-care	1
Citation analysis	3	International collaboration	1	Research productivity	1
Citation networks	2	JASIS	1	Retrieval/search	2
Co-authorship network(s)	3	Joint authorship	1	Scholarly communication	2
Co-citation analysis	1	Journal articles	2	Scientific and technical information	1
Combined co-citation	2	Journal impact factor	2	Scientific publications	1
Communication	2	Journal self-citation	1	Scientometrics	2
Community	1	LIS	2	Social network(s)	2
Complex networks	1	Macro	1	Social tagging	1
Content analysis	1	Mathematical-theory	1	Text mining/processing	3
Co-words	1	Mechanism	1	United States	1
Discipline	1	Multiple authorship	1	Word analysis	1
Doctoral programs	1	Natural language processing	1		

includes research evaluation, peer review, bibliometrics, and altmetrics, very similar to that of Y. Ding. As a result, keywords the authors used could reflect their area of interests.

Therefore, three basic assumptions on method of calculation of keyword-based parameter between two authors lie on the followings: (1) Similar to the information carrier, keywords can also be divided into specific research areas and types. (2) The same or similar meaning of keywords used in different articles indicates that there is certain connection or relation on these studies and authors' interests. (3) The more the number of similar keywords appeared in two articles are, the stronger the relation between them. Therefore, the keyword-based parameter of MACA indicates the quantity of relationship in keyword perspective between two authors whose works are co-cited in one or more articles. Figure 7 shows the block diagram to calculate the keyword-based parameter. After collecting all referred papers made by the author A_i individually, all keywords of the papers cited are extracted and given indexes in field indexer according to their focusing areas. Then the field indexes of all papers cited are computed and inputted to

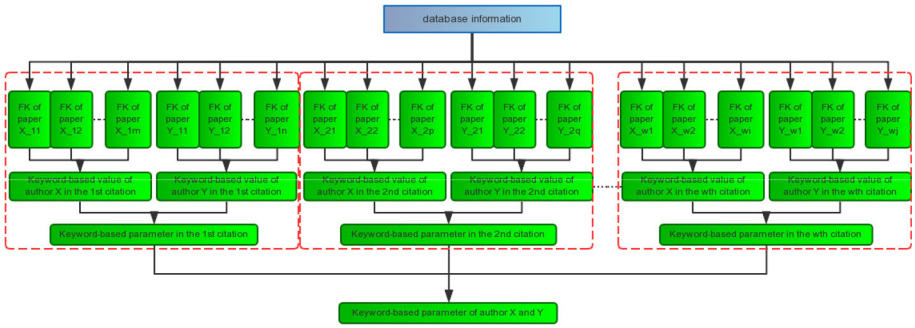

Fig. 7 The procedure of calculating keyword-based parameter in MACA (*FK* fields of keywords. Note that *FK* is a vector instead of one value)

keyword-based relation calculator. After that, the keyword-based parameter of MACA will be produced.

In keyword-based relation calculator, assume that there are total K distinct citations' keywords in dataset which are divided into ξ different professional fields. Similar to the definition of carrier-based parameter, an article, P_l and $l \in [1, n]$, has the references, D_r and $r \in [1, m]$, with their authors, A_i and $i \in [1, I]$, and their keywords, k_q and $q \in [1, K]$. A big matrix, showing the field distribution of the keyword k_q in a reference D_r and its author A_i in article P_l, is formulated as $F = \left(f_{l,i,j,r}\right)$ and

$$f_{l,i,j,r,q} = \begin{cases} 1, & k_q \text{ of } D_r \text{ with } A_i \text{ is related to the } j\text{th field} \\ 0, & k_q \text{ of } D_r \text{ with } A_i \text{ is not related to the } j\text{th field} \end{cases} \tag{7}$$

where $j \in [1, \xi]$ is the field index. Define the field relation, DFR, of a reference D_r of an author A_i in article P_l on field j as

$$\mathrm{DFR}_{A_i,P_l,D_r,j} = \frac{\sum_{q=1}^{K} f_{l,i,j,r,q}}{\sum_{j=1}^{\xi}\sum_{q=1}^{K} f_{l,i,j,r,q}} \cdot \varepsilon \tag{8}$$

where $\varepsilon \in N^*$, normally $3 \le k \le 7$, is adaptive variable for normalization of keyword-based parameter, and the field relation, FR, of author A_i in article P_l on field j can be calculated by

$$\mathrm{FR}_{A_i,P_l,j} = \frac{\sum_{r=1}^{m} \mathrm{DFR}_{A_i,P_l,D_r,j}}{\mathrm{NZ}\left[\mathrm{DFR}_{A_i,P_l,D_r,j}\right]} \tag{9}$$

Here, NZ[·] is a function that assigns the number of nonzero entries to an input matrix. Then the field correlation between two co-cited authors, A_i and A_k, in article P_l is calculated by

$$\mathrm{FKR}_{A_i,A_z,P_l} = \frac{1}{\xi}\sum_{j=1}^{\xi}\left[1 + \left(\mathrm{FR}_{A_i,P_l,j} - \mathrm{FR}_{A_z,P_l,j}\right)^2\right]^{-1} \tag{10}$$

Therefore, the keyword-based parameter of the two authors within the range $[0, 1]$, shown in Eq. (11), is the average of their field correlation in the dataset.

$$\text{Ave\_FKR}_{A_i,A_z} = \frac{1}{n}\sum_{l=1}^{n} \text{FKR}_{A_i,A_z,P_l} \tag{11}$$

For example, an article X having four references, D_1, D_2, D_3, and D_4, with their corresponding authors, A, A, B, and B, is shown in Table 6. And Table 7 indicates that total ten keywords in citations with their field distribution. Obviously, the field distribution of the referred papers in keyword perspective can be calculated and their results are shown in Table 8.

The field relation of each reference with its corresponding author on these fields is then computed according to Eq. (8). Here, ε is set to 4 in this case and their calculated results are shown in Table 9. Each author's field distribution is estimated one by one and exhibited in Table 10. Hence, the field correlation between two co-cited authors A and B in paper X can be calculated as:

$$\text{FCR}_{A,B,X} = \frac{1}{5} \times \left[\frac{1}{1+(2.2-1.4)^2} + \frac{1}{1+(0.5-0.0)^2} + \frac{1}{1+\left(0.5-\frac{1}{3}\right)^2} + \frac{1}{1+(0.4-0.0)^2} + \frac{1}{1+\left(0.4-\frac{34}{15}\right)^2} \right] \approx 0.69$$

If A and B are also co-cited in other two papers, Y, and Z, with their field correlation $FCR_{A,B,Y} = FCR_{A,B,Z} = 1$, the keyword-based parameter of the two authors is $(0.69 + 1 + 1)/3 \approx 0.897$.

Table 6 Examples of four papers and their field distributions

Paper	Author	Keywords
D_1	A	k_1, k_2, k_3, k_6, k_7
D_2	A	k_1, k_2, k_4, k_5
D_3	B	k_1, k_7, k_8, k_9, k_{10}
D_4	B	k_1, k_2, k_3, k_5, k_8, k_{10}

Table 7 Examples of the field distribution of overall keywords in citations

Paper	F_1	F_2	F_3	F_4	F_5
k_1	1	0	0	0	0
k_2	1	0	0	0	0
k_3	1	0	0	0	0
k_4	0	1	0	0	0
k_5	0	0	1	0	0
k_6	0	0	0	1	0
k_7	0	0	0	0	1
k_8	0	0	0	0	1
k_9	0	0	0	0	1
k_{10}	0	0	0	0	1

Table 8 Examples of field distribution of the referred papers in keyword perspective

(Paper, author)	F_1	F_2	F_3	F_4	F_5
(D_1, A)	3	0	0	1	1
(D_2, A)	2	1	1	0	0
(D_3, B)	1	0	0	0	4
(D_4, B)	3	0	1	0	2

Table 9 Field relation of all references in keyword perspective ($\varepsilon = 4$)

(Paper, author)	F_1	F_2	F_3	F_4	F_5
(D_1, A)	$3/5 \times 4 = 2.4$	$0/5 = 0.0$	$0/5 \times 4 = 0.0$	$1/5 \times 4 = 0.8$	$1/5 \times 4 = 0.8$
(D_2, A)	$2/4 \times 4 = 2.0$	1.0	1.0	0	0
(D_3, B)	$1/5 \times 4 = 0.8$	0.0	0.0	0.0	3.2
(D_4, B)	$3/6 \times 4 = 2.0$	0.0	2/3	0.0	4/3

Table 10 The field distribution of the authors in keyword perspective

Author	F_1	F_2	F_3	F_4	F_5
A	$(2.4 + 2.0)/2 = 2.2$	$(0.0 + 1.0)/2 = 0.5$	$(0.0 + 1.0)/2 = 0.5$	$(0.8 + 0.0)/2 = 0.4$	$(0.8 + 0.0)/2 = 0.4$
B	1.4	0.0	1/3	0.0	34/15

Construction of the co-citation matrix based on three above parameters

The proposed MACA mainly construct raw co-citation matrix synthesized author-based co-citation matrix of ACA with time-based parameter, carrier-based parameter, and keyword-based parameter. Furthermore, each entry of these matrices should be normalized to the same space for calculation and its range is set to $[0, 1]$ in this paper. Only the normalization of author-based co-citation matrix in ACA is required because the range of other three parameters mentioned is satisfied. Here, the original author-based co-citation matrix in ACA after normalization is defined as

$$\text{Nor\_RCM}_{A_iA_z} = \frac{\text{RCM}_{A_iA_z}}{\text{Max}\left(\text{RCM}_{A_iA_z}\right)} \tag{12}$$

where $\text{RCM}_{A_iA_j}$ is author-based co-citation matrix in ACA between the two authors, A_i and A_z, and the function $\text{Max}(\cdot)$ output the maximal entry of a matrix. Then the co-citation matrix, notated as $M = \left(m_{i,z}\right)$, in MACA is formulated as

$$m_{i,z} = w_t \cdot \text{Ave\_FTR}_{A_i,A_j} + w_c \cdot \text{Ave\_FKR}_{A_i,A_z} + w_k \cdot \text{Ave\_FCR}_{A_i,A_z} + w_A \cdot \text{Nor\_RCM}_{A_iA_z} \tag{13}$$

where w_t, w_c, w_k, and w_A indicate the weight of time-, carrier-, and keyword-based parameter, as well as author-based co-citation matrix value. Besides, each weight is larger than 0 and the summation of these weights is 1.

Experimental results and discussion

Dataset and preprocessing

The primary dataset used in the paper is the articles in JASIST from January 2003 to December 2012. All general descriptive metadata of the articles and their citations, including title, authors, published time, published carriers, volume and issues, keywords, and the number of pages, are exploited. Totally 2038 articles and 68,606 citations are used after preliminary refinement. For citations, only the first-author information of an article is adopted and their names are processed for disambiguation and artificial filtration. The authors appeared <10 times are ignored for keeping experimental quality, and then 958 authors and 30,512 citations were left. At last, 100 most popular authors, i.e. they received most citations, are selected for reducing computation complexity. The diagonal entries of the citation matrices are set to 0 in our experiment. Multi-type analyses including MDS and factor analysis are executed for showing the performance of the proposed MACA, and all results are demonstrated in a two-dimensional graph by using ALSCAL in SPSS 20.0. The factor analysis abstracts all principal components whose values are more than 1 and the "rotation solution" is outputted by using the "maximum variance analysis".

Indicating the affiliated professional field of keywords and information carriers

The affiliated professional field of keywords and information carriers should be provided before calculating carrier- and keyword-based parameters in the proposed MACA. Thesaurus utilization and manual classification are major ways to index their professional fields in the paper. The procedure for classifying the belonging file of keywords is described as follows:

(1) Extracting all keywords of citations in dataset.
(2) Removing duplication and filtering simply (e.g. "method", "methods" and "methodology" are regarded as the same keyword).
(3) Classifying keywords according to subject headings in thesauruses. (Note that some keywords may belong to more than one field)
(4) Classifying keywords manually that are not available in thesauruses. (A few academic professionals in LIS area would examine the classified results)

In the dataset, total 2053 keywords are extracted and 6 major categories are defined after the procedure. Table 11 shows these fields of keywords with their examples, basic statistics, and the indexes assigned. Note that the categories can be more specific, but six fields would be enough in our experiments for demonstrating the performance of MACA. Some keywords can be classified into more than one field, and "information retrieval" would be both in category 5 and 6, for example.

Similarly, the procedure for classifying the belonging field of information carriers is described as follows:

(1) Extracting all carriers of citations in dataset.
(2) Searching the catalogs of each information carrier on Essential Science Indicator (ESI).
(3) Classifying carrier manually which are not available on ESI.

Table 11 Catalogs of citations keywords and their indexes

Fields of keywords	Examples	Number/rate of keywords (%)	Index
LIS research (quantitative)	Author co-citation analysis	464/22.6	6
LIS research (qualitative)	Human information behavior	578/28.2	5
Computer science/ engineering	Software	285/13.9	4
Medical science/biology	Cardiopathy	348/17.0	3
MIS/business research	Knowledge management	127/6.2	2
General keywords/other fields	Comparative research; tenth century	247/12.1	1

Table 12 Catalogs of information carriers and their indexes

Fields of carriers	Index
Informetrics/data science/LIS quantitative research	5
IR/behavior studies/HCI/Other LIS research	4
Computer science/engineering	3
MIS/business research	2
Other fields	1

(a) Downloaded its contents and reading more than 50 the articles of each carrier in the experimental period.
(b) Classified them according to their keywords, the characteristic of contents, and judgments.
(c) The classified results would be examined by a few academic professionals in LIS area.

In the dataset, all primary articles from JASIST have more focus on LIS. The citation sources are majorly divided into five categories in our experiment after the procedure. Table 12 shows these fields of information carriers with its index assigned. Obviously, some information carriers also have more than one belonging fields. For instance, the carrier "*iConference*" might affiliate both category 4 and 5.

Multi-dimensional scaling (MDS)

Multi-dimensional scaling (MDS), usually for visualizing the level of similarity of individual cases in a dataset, is employed in showing the performance of the proposed MACA. Three parameters majorly in MACA are proposed to combine with raw co-citation matrix in ACA. For showing their performance separately, the notations of the ACA combined with the parameters are defined as follows:

(1) MDS-A: MDS result of the traditional ACA (ACA).
(2) MDS-AT: MDS result of ACA combined with time-based parameter (ACA + T). The weights used for author- and time-based parameters are 0.6 and 0.4, respectively.

(3) MDS-AC: MDS result of ACA combined with carrier-based parameter (ACA + C). The weights used for author- and carrier-based parameters are 0.6 and 0.4, respectively.
(4) MDS-AK: MDS result of ACA combined with keyword-based parameter (ACA + K). The weights used for author- and keyword-based parameters are 0.6 and 0.4, respectively.
(5) MDS-ATC: MDS result of ACA combined with T and C (ACA + TC). The weights used for author-, time- and carrier-based parameters are 0.5, 0.25, and 0.25, respectively.
(6) MDS-ATK: MDS result of ACA combined with T and K (ACA + TK). The weights used for author-, time- and keyword-based parameters are 0.5, 0.25, and 0.25, respectively.
(7) MDS-ACK: MDS result of ACA combined with C and K (ACA + CK). The weights used for author-, carrier-, and keyword-based parameters are 0.5, 0.25, and 0.25, respectively.
(8) MDS-M: The MDS result of ACA combined with all three parameters (MACA). The weights used for author-, time-, carrier-, and keyword-based parameters are 0.5, 0.2, 0.1, and 0.2, respectively.

All the weights used in these algorithms combined with other parameters are finally decided after examining all possible experiments. All of these have 0.5 or more for author-based parameter because the author co-citation relationship should be a basic element to construct the network. Figure 8 demonstrates MDS results of ACA combined with the parameters separately. MDS-A, MDS-AT, MDS-AC, MDS-AK, MDS-ATC, MDS-ATK, MDS-ACK, and MDS-M are shown in from the 1st row to the 4th row left and right, respectively.

The area of each aggregation in MDS-M is smaller than that in MDS-A due to the data set from JASIST focusing on particular fields. MDS-M also has remoter results between points in different aggregations because three categories split in the dataset are different in a sense. The professional fields of three categories are shown in Table 13. Some authors having more than one study within these fields would be located in the junction of the aggregations. In all MDS results, the points in the category, i.e., information retrieval/information behavior/user studies, are more separated because most of the authors in the aggregation have studies combined with other fields. In MDS-M, the authors in semantic-related aggregation are more gathered due to their commonly focusing on the algorithms. And their studies are closer to retrieval-related/behavior-related aggregation. For example, some text mining methods are exploited to construct their solutions and explain experimental results in the area of user studies (e.g., Davis 2004; Park and Park 2014). The points in informetrics-related aggregation are also concentrated because the issues in the field are more specific. And information retrieval-related authors also have studies cited with the articles in informetrics-based research (Swanson et al. 2001).

In order to explain the nuances among these algorithms, six authors in the dataset are selected and their interest areas with mark given are shown in Table 14. The location of these authors are identified and colored in every graph of MDS results in Fig. 8. In the case of the authors 1–3, their locations on MDS-A are more separate than those on MDS-M. Meaningfully, MACA indicate the authors' studies are relatively similar, and several studies of them in semantic and network-based research are surely covered after examining their studies. Besides, that their locations on MDS-AT, MDS-AC, and MDS-AK, have different relative distance indicates each parameter have various impacts on their

Fig. 8 MDS results produced by ACA and MACA-series (ACA + T, ACA + C, ACA + K, ACA + TC, ACA + TK, ACA + CK, MACA)

Table 13 Authors' interests of each aggregation in MDS results

Mark	Area of interests
I	Informetrics; scientometrics; data analysis; information analysis and decision
II	Information retrieval; information behavior; user studies
III	Semantic research; network-based research; NLP; text mining; engineering

correlations. In the case of the author 5 and 6, they have closer distance between their locations on MDS-A. Due to the correlation of their studies, the locations between them on MDS-M and others indicate the actuality. Moreover, the two author groups, 1–3 and 5–6, have resembling interests, such as the studies of authors 1 and 5. Thus these five authors should be similar and their locations on MDS-M are also typical closer in visualization. In the case of the author 4, her field classified is unlike others' and her locations on MDS-A and MDS-M are also far from them. In fact, examining the areas of interests of author 4 and others also reveal these dissimilarities shown in the graphs. Furthermore, as for author 7, his position in MDS-A is far from that of 1, 2, and 3; yet in MDS-M, their distance decreases, which implies that his field is in some degree related to 1, 2, and 3's. Indeed, author 7 has some relative studies, such as semantic-based methods to analyze the researchers' citing behavior (Case and Miller 2011). That explains why his position is closer to category III as more general information of the citations is involved.

MDS-measurement

In order to exhibit MDS results more quantitatively, MDS-measurement, named, σ, is deduced by two variables, c and S, indicating cohesion and separation. These two variables are majorly exploited to evaluate the effect of a clustering result (Kaufman and Rousseeuw 1990). Assume that all ϕ authors are divided into ξ categories by their field belongs in MDS graph G. In p, $p \in [1, \xi]$, category there are n_p authors with their coordinate $\left(x_q^p, y_q^p\right)$, $q \in \left[1, n_p\right]$. And the coordinate of central point is $\left(x_c^p, y_c^p\right)$. The Euclidean metric, ρ_i^p, of a certain point in a category, can be defined as:

$$\rho_i^p = \sqrt{\left(x_i^p - x_c^p\right)^2 + \left(y_i^p - y_c^p\right)^2}, \quad \forall i \in \left[1, n_p\right] \tag{14}$$

The sum of Euclidean metric between all points and central points within their categories is

$$c = \sum_{p \in \{1,2,\ldots,\xi\}} \frac{1}{2n_p} \sum_{i \in \left\{1,2,\ldots,n_p\right\}} \rho_i^p \tag{15}$$

Besides, the sum of Euclidean metric between every two points in different categories is

$$S = \sum_{s \in p \cap v \notin p} c_{sv} \tag{16}$$

Then MDS-measurement is defined as

$$\sigma = c/S \tag{17}$$

Here c represents the degree of cohesion in clustering result and S is the degree of separation. Higher cohesion (bigger c) in the same category and higher separation (smaller

Table 14 Seven authors and their area of interests

Author	Mark	Area of interests
P. Ahlgren	1	Text mining/NLP/data analysis/sentiment analysis
A. Barabási	2	Network science/statistic physics/biophysics
J. Bar-llan	3	Internet research/network science/informetrics
S. Y. Rieh	4	Information behavior/information seeking behavior
M. Thelwall	5	Webometrics/quantitative methods/social networks
M. Newman	6	Network science/computer simulation/complex physic systems
D. O. Case	7	Information behavior/social and educational effects of ITs

Table 15 MDS-measurement results of different models

Models	c	S	σ (%)
ACA	53.29	437.60	12.18
ACA + T	44.95	451.35	9.96
ACA + C	49.87	476.51	10.47
ACA + K	45.39	457.85	9.91
ACA + TC	45.16	488.01	9.25
ACA + TK	40.72	461.01	8.83
ACA + CK	44.36	488.45	9.08
ACA + TCK (MACA)	39.71	488.98	8.12

S) in different category would be regarded as a good result. MDS-measurements of ACA and MACA are shown in Table 15 and $\sigma(\text{MACA}) < \sigma(\text{ACA} + \text{TK}) < \sigma(\text{ACA} + \text{K}) < \sigma(\text{ACA})$ reveals that σ becomes smaller as more factors are involved in the experiments. This implies that points in the same category are closer and those in different categories are more separate while more elements are involved. In addition, in cases of the same number of factors, the parameter K does impact more than T and C in our experiment. Observationally, the parameters, T, K, and C, have different impact on ACA results. In the difference of MDS-measurement between ACA + T and ACA + TC, smaller c with larger S refers that the carrier where citations are published has more impact on the points with different categories. The parameter servers the authors whose research is in different fields. This also can be observed in ACA + K and ACA + CK, or ACA + TK and MACA. In MDS-measurement among ACA + T, ACA + C, and ACA + K, furthermore, larger c with small S refers that the published time and keywords of citations has more impact on the points within one category. The parameter gathers the authors whose research is in the same field.

Factor analysis

Factor analysis, a statistical method, is usually utilized to describe variability among observed variables and factors produced are correlated variables concerning a potentially lower number of unobserved variables. Table 16 shows the results of factor analysis based on different models with different authors (load factor >0.3). Total five factors, notating 1–5, are obtained and their indexes represent information retrieval and seeking, traditional

Table 16 Factor analysis to all algorithms on three example authors (*NK* number of keywords, and *ACT* accmulative contribution value. Note that the authors' load factor is larger than 0.3 in this table)

Model	ACV (%)		S. Brin	D. R. Swanson	T. D. Wilson	Model	ACV (%)		S. Brin	D. R. Swanson	T. D. Wilson
ACA	1	36.9		0.447	0.759	ACA + TC	1	52.3		0.571	0.986
	2	57.1	0.918				2	65.2	0.546		
	3	76.2					3	77.8	0.417	0.323	
	4	90.9		0.784			4	89.0		0.559	
	5	97.8					5	95.5		0.584	
ACA + T	1	52.0		0.479	0.909	ACA + TK	1	55.1		0.505	0.895
	2	68.6	0.860				2	68.9	0.642		
	3	80.5					3	80.1		0.424	
	4	91.0		0.791			4	90.5		0.671	0.300
	5	98.8		0.350			5	96.0		0.491	
ACA + C	1	38.4		0.421	0.842	ACA + CK	1	53.0		0.516	0.898
	2	59.4	0.892				2	66.3	0.511		
	3	74.6					3	78.0	0.301	0.398	
	4	94.2		0.648			4	89.4		0.484	
	5	97.8		0.596	0.914		5	97.5		0.464	
ACA + K	1	52.2		0.596	0.914	MACA	1	56.9		0.428	0.921
	2	68.4	0.672				2	68.4	0.523		
	3	81.8					3	79.5	0.396	0.302	
	4	92.3		0.525			4	89.9		0.313	0.309
	5	98.2		0.388			5	96.1		0.465	

LIS and information analyses, informetrics and data-science related research, information (seeking) behavior and user studies, and semantic- or network-based analysis, respectively. More than a factor existed in an author indicate that the author probably has different study fields. The accumulative contribution value of each factor in different algorithms is also shown. For example, the accumulative contribution values of the 1–5 factors in ACA are 36.9, 57.1, 76.2, 90.9, 97.8 %, respectively. The prominent 1st factor reveals that the authors, whose study field belongs to information retrieval and seeking, are popular and authoritative.

The five factors in different algorithms also demonstrate varying degree of author's interested areas. For example, Dr. Don R. Swanson, a famous researcher in LIS, has many important studies in different professional areas. According to the investigation, his main area of interest is information retrieval (Swanson 1979), user psychology, and behavior analysis (Swandon 1977). The factors, 1 and 4, in ACA can obviously establish its factualness. The 5th factor identified in ACA + T and ACA + K provides a clue that he has several studies related to the area (Swanson 1960). It can't be observed in ACA + C because the carriers of these articles probably might not have strong attributes of the area. Moreover, the 3rd factor which emerged in four other algorithms indicates that his research areas are perhaps related to informetrics and data science. The observation produced by these algorithms is correct after examining his publication (Swanson et al. 2001). Observations in the other two authors likewise reveals that professional field of their partial research can be explored in MACA. Compared with ACA, as a result, MACA could obtain more details and nuances from the dataset while more information is imported.

Conclusion

A Modified Author Co-Citation Analysis (MACA) method is proposed in the paper for eliciting a bird's eye view of intellectual structure in a research field. Four kinds of different general descriptive metadata, authors of a citation, citations' published time, citations' published carrier, and keywords of a citation, are exploited in MACA to construct a co-citation network. The major difference of MACA from ACA is the stage at which the former constructs the raw co-citation matrix when calculating the author co-citation relationship, the difference of their published time, the relationship in professional fields based on their carriers, and keywords. In our experimental results, more professional fields of an author are explored in MACA and the distribution of each field indicates the number of research district. Compared with ACA, MACA have more detailed and sensitive demonstrations in MDS. The main contributions of the proposed MACA are as follows: (1) By adding more information to the author co-citation analysis one can provide more details and nuance analysis to the dataset; (2) MACA has a good demonstration of the analysis of knowledge domain while extra calculations required in MACA are just a little more than what is required in ACA; (3) Different additional information has different impacts on the clustering results. For example, the published carriers would obviously separate authors whose interests are in different fields and the keywords of the authors' articles effectively gather authors with different interests. As a result, MACA can be another option to understand researchers and the knowledge map in a study field with higher fineness. Furthermore, content-based ACA exploiting the content in an article can also be combined with MACA for improving the accuracy and efficiency of analysis.

However, the two parameters, carriers and keywords, in MACA are derived from the classification of professional fields. In this paper, a simple way to reconstruct the categories of the fields for analysis is proposed and we believe that there are many other methods to establish these, such as classification, machine learning, and ontology-based method, etc. The more categories one can divide will reveal nuances from the results of clustering in MACA. Thus, we would like to focus on the classifying methods and adaptive size of categories for MACA in the future.

Acknowledgments The authors would like to thank two research assistants in our group, Binglu Wang and Chengyue Gong, for their kind help on part of data collections and preliminary author-filtering. We are also very grateful for the constructive comments and helpful suggestions from the anonymous reviewers and the editor in chief of *Scientometrics*, Dr. Wolfgang Glänzel.

References

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a citation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology, 54*(6), 550–560.

An, L., Zhang, J., & Yu, C. (2011). The visual subject analysis of library and information science journals with self-organizing map. *Knowledge Organization, 38*(4), 299–320.

Bensman, S. J. (2004). Pearson's *r* and author co-citation analysis: A commentary on the controversy. *Journal of the American Society for Information Science and Technology, 55*(10), 935.

Bianchini, M., Gori, M., & Scarselli, F. (2005). Inside PageRank. *ACM Transactions on Internet Technology, 5*(1), 92–128.

Bornmann, L., & Daniel, H. D. (2008). What do citation count measure? A review of studies on citing behavior. *Journal of Documentation, 64*, 45–80.

Brooks, T. A. (1985). Private acts and public objects: An investigation of citer motivations. *Journal of the American Society for Information Science, 36*(4), 223–229.

Bu, Y., Liu, T., & Huang, B. (2015). Exploration on research of improving traditional author co-citation analysis: A novel author co-citation analysis method combining with publishing time of cited papers. *Library and Information Knowledge, 32*(6), 89–97.

Case, D. O., & Miller, J. (2011). Do bibliometricians cite differently from other scholars? *Journal of the American Society for Information Science and Technology, 62*(3), 421–432.

Chen, C. (1999). Visualizing semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management, 35*(3), 401–420.

Cothill, C. A., Rogers, E. M., & Mills, T. (1989). Co-citation analysis of the scientific literature of innovation research traditions. *Science Communication, 11*(2), 181–208.

Davis, P. (2004). Information-seeking behavior of chemists: A transaction log analysis of referral URLs. *Journal of the American Society for Information Science and Technology, 55*(4), 326–332.

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

Ding, Y., Chowdhury, G., & Foo, S. (2000). Journal as markers of intellectual space: Journal co-citation analysis of information retrieval area, 1987–1997. *Scientometrics, 47*(1), 55–73.

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

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *Journal of the Association for Information Science and Technology, 60*(11), 2229–2243.

Egghe, L. (2009). New relations between similarity measures for vectors based on vector norms. *Journal of the American Society for Information Science and Technology, 60*(2), 232–239.

Eom, S. (1999). Decision support systems research: Current state and trends. *Industrial Management and Data Systems, 99*(5), 213–221.

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

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

Hartley, J. (2008). *Academic writing and publishing: A practical guide*. New York: Routledge.

Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics, 8*(1), 197–211.
Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: John Wiley and Sons.
Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the web environment. *Journal of the American Society for Information Science and Technology, 57*(12), 1616–1628.
Li, J., & Gong, J. (2010). Frontier and trend analysis of information science research based on JASIST. *Research in Library Science, 3*, 2–6.
Liu, L., Xuan, Z., Dang, Z., Guo, Q., & Wang, Z. (2007). Weighted network properties of Chinese nature science basic research. *Physica A-Statistical Mechanics and Its Applications, 377*(1), 302–314.
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.
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.
Moya-Anegón, S. G., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Munoz-Fernández, F. J., & Herrero-Solana, V. (2007). Visualizing the marrow of science. *Journal of the American Society for Information Science and Technology, 58*(14), 2167–2179.
Park, H., & Park, M. (2014). Cancer information-seeking behaviors and information needs among Korean Americans in the online community. *Journal of Community Health, 39*(2), 213–220.
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, 55*(6), 513–529.
Schneider, J. W., & Larsen, B. (2009). A comparative study of first and all-author co-citation counting, and two different matrix generation approaches applied for author co-citation analyses. *Scientometrics, 80*(1), 103–130.
Swandon, D. (1977). Critique of psychic energy as an explanatory concept. *Journal of the American Psychoanalytic Association, 25*(3), 603–633.
Swanson, D. (1960). Searching natural language text by computer. *Science, 132*, 1099–1104.
Swanson, D. (1979). XMARC: A system for experimental indexing and searching of MARC records. *Journal of Education for Librarianship, 20*(2), 91–106.
Swanson, D., Smalheiser, N., & Bookstein, A. (2001). Information discovery from complementary literatures: Categorizing viruses as potential weapons. *Journal of the American Society for Information Science and Technology, 52*(10), 797–812.
Tsay, M. Y. (2011). The subject structure of randomized controlled trials: An author co-citation analysis. In Noyons, E., Ngulube, P., & Leta, J. (Eds.). *Proceedings of ISSI 2011—The 13th international conference on scientometrics and informetrics.* Durban, pp. 1067–1069.
White, H. D. (2003a). Author co-citation analysis and Pearson's r. *Journal of the American Society for Information Science and Technology, 54*(13), 1250–1259.
White, H. D. (2003b). Pathfinder networks and author co-citation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology, 54*(5), 423–434.
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.
Wu, Y., Fu, T., & Chiu, D. (2014). Generalized preferential attachment considering aging. *Journal of Informetrics, 8*(3), 650–658.
Yang, J. (2013). The Library and Information Science research hotpot perspective based on JASIST. *Library Work and Study, 2*, 22–25.
Zhao, D. (2006). Towards all-author co-citation analysis. *Information Processing and Management, 42*(6), 1578–1591.
Zhao, D., & Logan, E. (2002). Citation analysis using scientific publications on the web as data source: A case study in the XML research area. *Scientometrics, 54*(2), 449–472.
Zhao, D., & Strotmann, A. (2008). Comparing all-author and first-author co-citation analyses of information science. *Journal of Informetrics, 2*(3), 229–239.