MFTACA: An Author Co-citation Analysis Method Combined with Metadata in Full Text

Yi Bu\textsuperscript{1} Binglu Wang\textsuperscript{2} Win-bin Huang\textsuperscript{3,}\textsuperscript{*} Shangkun Che\textsuperscript{4}

\textsuperscript{1}buyi@iu.edu
Indiana University Bloomington (America)

\textsuperscript{2}maplewang@pku.edu.cn
Peking University (China)

\textsuperscript{3}huangwb@pku.edu.cn
Peking University (China, corresponding author)

\textsuperscript{4}csk@pku.edu.cn
Peking University (China)

Abstract

As a frequently used method of depicting scientific intellectual structures, author co-citation analysis (ACA) has been applied to many domains. However, only count-based information is involved as the input of ACA, which is not sufficiently informative for knowledge representations. This article catches several metadata in full text of citing papers but not aims at content-level information, which could increase the amount of information input to ACA but not increase computational complexity a lot. By involving information including the number of mentioned times in a citing paper, the number of context words in a citing sentence, and the published year of a reference, we propose a new method called MFTACA (metadata-in-full-text-based ACA). We combine these pieces of information into traditional ACA and compare the results between ACA and MFTACA by using factor analysis, network analysis, and MDS-measurement. The result of our empirical study indicates that compared with traditional ACA, our proposed method shows a better clustering performance in visualizations and reveals more details in displaying intellectual structures.

Conference Topic

Citation and co-citation analysis; Mapping and visualization; Science of science

Introduction

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

Traditional ACA regards two authors with higher co-citation frequency as higher topical relatedness. Such assumption hints that every author pair with the same co-citation frequency as identical, which simply considers count-based instead of content-based information. As the availability of full-text data nowadays, Jeong \textit{et al.} (2014) firstly proposed content-based ACA method and compared the similarity between citing sentences. However, the computational cost of their content-based method could be high because of the processing of words as well as similarity calculation between citing sentences. Actually, with the full-text data, we do not have to employ the content-level information. Nevertheless, instead, several pieces of useful information at metadata level that were ignored previously in full text could be considered to improve the performance of ACA in mapping knowledge domains.
The number of mentioned times of references is a typical piece of information. As pointed out by Ding, Liu, Guo, and Cronin (2013) as well as Zhao, Cappello, and Johnston (2017), the number of mentioned times of a reference represents the importance of the reference to the citing paper. However, traditional ACA regards two co-cited authors with a distinct number of mentioned times in a citing paper as identical, which could be problematic. For example, Zhao and Strotmann (2014) cited (a) Zhao and Strotmann (2008a), (b) White and McCain (1998), and (c) Hirsch (2005), but (a) was mentioned 15 times, (b) twice, while (c) only once in the citing paper. The co-citation strength of pair (a)-(b) and (b)-(c) should be different when we consider their topical relatedness. In this paper, we start to consider the number of mentioned times of references as a supplement into ACA in order to provide more accurate information for mapping knowledge domains.

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

Moreover, the published time of references could also reveal the difference between co-cited authors. For instance, as pointed out by Bu, Liu, and Huang (2016), small difference between the published time of two references implies that the authors tend to focus on similar issues in the same period of time; therefore, the two authors’ relationship could be distinct in knowledge domain maps because they tend to use “various concepts, methods, or even diversified demands” (p. 144) in dissimilar periods of time. The information of references; published time is considered in this paper and it is combined with traditional ACA to enhance the performance of knowledge domain mappings.

This article is outlined as follows. At first, the work related to our study and the data with the methods for our analysis are detailed. The findings as well as the comparisons between traditional ACA and our proposed MFTACA then are presented. Finally, the conclusions and the future research are pointed out.

Related Studies

Author co-citation analysis (ACA) was proposed by White and Griffith (1981). In 1990, McCain gave a completed overview and set up a standard framework for ACA, in which four steps of ACA implementations were mentioned: (1) Data collection and processing; (2) Construction of raw co-citation matrix; (3) Transformation to correlation matrix; and (4) Data analyses (e.g., factor analysis, clustering analysis, multi-dimensional scaling (MDS) analysis, and network analysis) and result interpretations. More than thirty years, this method has been improved a lot by revising rules to construct raw co-citation matrix and transform to correlation matrix. As pointed out by Persson (2001), the elements in co-citation matrix could be defined as first-author co-citation frequency or all-author co-citation frequency; the latter could to some extent provide more detailed knowledge domain maps (Rousseau & Zuccala, 2004; Zhao, 2006). Meanwhile, the rules of defining main diagonal values (Mégnigbéto, 2013) and transforming correlation matrix (Ahlgren, Jarneving, & Rousseau, 2003; White, 2004) were both discussed and improved in detail to make the data processing more accurate. Additionally, several
metadata of references, such as published time and venue of references and their keywords, were considered in ACA implementations (Bu et al., 2016), and they were found to play positive roles in improving the performance of ACA maps. Note that the metadata they employed had been obtained from reference lists instead of full text.

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

However, when full-text data are used, it is not required to make content- or semantic-level analyses. Several non-content-level metadata are easily accessible in full text, such as the number of mentioned times and the number of context words in the citing paper. In addition, these pieces of information reflect the importance of reference to citing paper. For example, if mentioned many times than others, a reference could have more topical relatedness to the citing paper; if the citing sentence containing certain reference is longer than that containing other references, it is more likely to be interpreted more detailedly and thus has higher possibilities to be related to the citing paper. As a result, this paper combines the number of mentioned times and the number of context words of references into ACA and proposes a new method called MFTACA so as to improve the performance of ACA in knowledge domain mappings.

**Methodology**

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

**Data**

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

Methods

Calculation of mentioned time parameters
A paper citing other papers refers that these cited papers (references) are related and useful to the citing paper (CP). Traditionally, these co-cited papers are regarded as equal in traditional co-citation analysis (Ding et al., 2013). However, the importance of cited papers could be distinct (Cano, 1989; Case & Higgins, 2000). Specifically, some references are crucial to the CP because they might be the foundation of CP. In the full-text, the important references could be revealed as multi-mentioned cited papers. For one CP, from an intuitive thinking, the more number of mentioned times a reference has, the more importance it is to the CP (Ding et al., 2013; Zhao et al., 2017). For instance, Zhao and Strotmann (2014) cited (a) Zhao and Strotmann (2008a), (b) White and McCain (1998), and (c) Hirsch (2005), but (a) was mentioned 15 times, (b) twice, while (c) only once. In this case, (a) should be the most important reference to the CP, compared with (b) and (c). Indeed, Zhao and Strotmann (2014) tried to map the knowledge domains of information science (IS) between 2006-2010 while (a) did the same thing but between 1996-2005 with similar methods, ACA and author bibliographic coupling analysis (ABCA), which shows that (a) is closely-related and important to the CP. However, (b) only used ACA to map the knowledge domain of IS instead of ABCA and the result of (b) is also very different from (a) and the CP. The reason (c) is cited is nothing but simply because the indicator “h-index” was used. Hence, we can see that the number of mentioned times of a reference is indeed positively related to its relatedness with CP.

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

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

$$MT_{ij,k} = \frac{\lambda_{ik}\lambda_{jk}}{\lambda_{k,max}}$$

(Eq. 1)

If we consider all of the $MT_{ij,k}$ in citing papers $P_{ij,1}, P_{ij,2}, \ldots, P_{ij,x_{ij}}$, the mentioned time parameter between author $A_i$ and $A_j$ among dataset, $MT_{ij}$, could be defined as:

$$MT_{ij} = \sum_{k=1}^{x_{ij}} MT_{ij,k}$$

(Eq. 2)

**Calculation of context word parameters**

When citing references, CPs tend to use one or more sentences to set up an argument, which is called citing sentences (or cite (Jeong et al., 2014)). However, the length of citing sentences could be distinct. For example, Zhao and Strotmann (2014) cited: (d) Finlay, Sugimoto, Li, and Russell (2012), (e) Milojević, Sugimoto, Yan, and Ding (2011), as well as (f) Sugimoto, Li, Russell, Finlay, and Ding (2011) in the same sentence with 27 words. They shared these 27 words and each of them has been assigned nine words averagely. Meanwhile, Zhao and Strotmann (2014) also cited (g) Zhao and Strotmann (2008b) in a sentence with 31 words. Although all of these references are cited once in the CP, their numbers of context words assigned could be diverse, 9, 9, 9, and 31, respectively.

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

In $P_{ij,k}$, assume that during its $\mu$th mention ($\mu = 1, 2, \ldots, \lambda_{ik}$), the citing sentence containing author $A_i$ ($A_j$) include $w_{ik\mu}$ ($w_{jk\mu}$) words and mentions $a_{ik\mu}$ ($a_{jk\mu}$) distinct authors ($w_{ik\mu}, w_{jk\mu}, a_{ik\mu}, a_{jk\mu} > 0$). The number of context words of author $A_i$ in paper $P_{ij,k}$, $cw_{i,k}$, could be calculated as (similar to $A_j$):

$$cw_{i,k} = \sum_{\mu=1}^{\lambda_{ik}} \frac{w_{ik\mu}}{a_{ik\mu}}$$

(Eq. 3)
If we annotate the largest number of context words of cited author in paper $P_{ij,k}$ as $cw_{k,max}$, the context word parameter between authors $A_i$ and $A_j$ in paper $P_{ij,k}$, $CW_{ij,k}$, is calculated as:

$$CW_{ij,k} = \frac{cw_{k,cw_{ij,k}}}{cw_{k,max}^2}$$  \hspace{1cm} (Eq. 4)

If we consider all of the $CW_{ij,k}$ in citing papers $P_{ij,1}, P_{ij,2}, \ldots$, and $P_{ij,x_{ij}}$, the context word parameter between authors $A_i$ and $A_j$ among dataset, $CW_{ij}$, could be defined as:

$$CW_{ij} = \frac{1}{x_{ij}} \sum_{k=1}^{x_{ij}} CW_{ij,k}$$  \hspace{1cm} (Eq. 5)

**Calculation of published year parameter**

Besides the mentioned time and context word parameters, we also employ published year parameter proposed by Bu et al. (2016), in which it is called time-based parameter. According to Bu et al. (2016), a small difference between two references’ published time refers that they tend to focus on similar research topics in the same time period; a large difference between two references’ published time refers that the representation of the co-cited authors’ relationships should be distinct because they are likely to use different methods, concepts, and tools although they might have similar research issues.

Suppose that $P_{ij,k}$ cites $r_{ik}$ papers published by author $A_i$ and $r_{jk}$ papers published by author $A_j$. The published year of these $r_{ik}$ papers published by author $A_i$ and $r_{jk}$ papers published by author $A_j$ is respectively $t_{ik1}, t_{ik2}, \ldots, t_{ikr_{ik}}, t_{jkl}, t_{jkl2}, \ldots, t_{jkr_{jk}}$. Then the average of published year of author $A_i$ in the article $P_{ij,k}$ is calculates as:

$$TB_{i,k} = \frac{1}{r_{ik}} \sum_{v=1}^{r_{ik}} t_{ikv}$$  \hspace{1cm} (Eq. 6)

The published year parameter between authors $A_i$ and $A_j$ in paper $P_{ij,k}$ is defined as the sum of the average of published year of authors $A_i$ and $A_j$:

$$TB_{ij,k} = \frac{1}{1+ln(1+|TB_{i,k}-TB_{j,k}|)}$$  \hspace{1cm} (Eq. 7)

As a result, the average of published year of cited papers written by author $A_i$ among dataset is calculated as:

$$TB_{ij} = \sum_{k=1}^{x_{ij}} TB_{ij,k}$$  \hspace{1cm} (Eq. 8)

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

The co-citation matrix in our proposed algorithm is based on three above parameters. All of the parameters are normalized into $[0,1]$. If we annotate the largest co-citation frequency among the dataset regardless of which author pairs as $x_{max}$, we can construct the new co-citation matrix, $M = (m_{i,j})$, as:

$$m_{i,j} = w_c \cdot \frac{x_{ij}}{x_{max}} + w_{MT} \cdot MT_{ij} + w_{CW} \cdot CW_{ij} + w_{TB} \cdot TB_{ij}$$  \hspace{1cm} (Eq. 9)

where the four positive weights, $w_c, w_{MT}, w_{CW},$ and $w_{TB}$, are relatively the weight of co-citation, mentioned time, context word, and published year parameters. Note that $w_c + w_{MT} + w_{CW} + w_{TB} = 1.0$.

**Results and Discussion**

**Factor Analysis**

In factor analysis, we extract the factors whose Eigen factor is 1.0 or more as the result of factor analysis, regardless in ACA or MFTACA. As shown in Table 1, the number of factors extracted from MFTACA is 18, which is seven more than that in ACA. In terms of the total variance
explained, the factor analysis of ACA could explain 82.8% of total variance while that of MFTACA explains 86.1%.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Number of factors extracted</th>
<th>Total variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACA</td>
<td>11</td>
<td>0.828</td>
</tr>
<tr>
<td>MFTACA</td>
<td>18</td>
<td>0.861</td>
</tr>
</tbody>
</table>

**Note:** Here, $w_c = 0.7, w_{MT} = 0.1, w_{CW} = 0.1, w_{TB} = 0.1$, which are finally determined after examining lots of possible experiments. The same below.

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

<table>
<thead>
<tr>
<th>ID</th>
<th>Factor</th>
<th>ACA</th>
<th>MFTACA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Information retrieval</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>Information seeking behavior</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3</td>
<td>Information usage, digital library</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>Language model, query, clustering</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>5</td>
<td>Classification algorithms, information organizations</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>6</td>
<td>Text mining, machine learning</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>7</td>
<td>User interface</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>8</td>
<td>User acceptance of information technology</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>9</td>
<td>Information systems</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>10</td>
<td>Data Mining, Data Analysis</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>11</td>
<td>Evaluation indicator, index</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>12</td>
<td>Webometrics, social network analysis</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>13</td>
<td>Visualization, mapping</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>14</td>
<td>Modeling</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>15</td>
<td>Scholarly communication</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>16</td>
<td>Journal citation analysis, interdisciplinarity, evaluation of algorithms</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>17</td>
<td>Network analysis</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>18</td>
<td>Bioinformatics</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Although using the same dataset as Jeong et al. (2014), we input approximately 500 authors into factor analysis while Jeong et al. (2014) did 100. We try to compare our results with theirs, which shows in **bold** in Table 2, where we can find that the results are similar and confirm each other’s. On the other hand, with respect to the factor analysis result of MFTACA, we can find
many detailed sub-fields of information science, such as visualization and data mining. These newly-detected domains could showcase the nuance and emerging topics of information science recently. Therefore, we believe that MFTACA could reveal more details and nuance in depicting scientific intellectual structures.

Network Analysis
Figures 2 and 3 show the scientific intellectual structures by using the two methods, ACA and MFTACA, where each node represents an author and the size of the node is proportional to the degree of the node in the given network. The distance between nodes are determined by ForceAtlas2 (Jacomy, Venturini, Heymann, & Bastian, 2014), a frequently used layout algorithm in Gephi. If two nodes lie near in the map, for instance, their relationship could be strong; and *vice versa*. For visualization, we employ Modularity algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008), in which the nodes (authors) within the same colour indicate that their research interests are similar, while those in different colors show that their research interests could be distinct. The labels of the clusters are manually given by our reading literatures of the authors as well as browsing their personal websites. From Figure 2 we can see that four clusters are detected, bibliometrics, information retrieval, information behaviour, and library science/qualitative research. The results are similar to Jeong et al. (2014)’s result, where they also found four clusters, bibliometrics, information retrieval (I), information retrieval (II), and library science. On the other hand, Figure 3 detects six clusters; besides four clusters having been detected in Figure 2, it also finds two additional clusters, i.e. text mining/data mining and network-based information science. These two clusters reveal the nuance of the development of IS and are detailed sub-fields in IS. These indicate that our proposed MFTACA method could provide more details in knowledge domain maps and help better understand the domain.
Intuitively, the nodes within the same cluster lie nearer, and the nodes in different clusters lie farther in Figure 3 than in Figure 2. These indicate a better clustering performance in MFTACA method. Take Drs. K. W. McCain and R. Rousseau as examples. Both of them focus on bibliometrics during their main scientific careers. Specifically, they are both interested in ACA, where Dr. McCain (1990) gave a comprehensive overview of ACA and used ACA to map knowledge domain of IS field between 1972 and 1995 (White & McCain, 1998). Dr. Rousseau, as a representative in typical European bibliometricians, proposed several types of ACA by classifying them according to distinct requirements (Rousseau & Zuccala, 2004) and discussed whether Pearson’s $r$ should be used in ACA (Ahlgren, Jarneving, & Rousseau, 2004). Their positions in Figure 3 are nearer than in Figure 2, which shows that MFTACA could play a role of closing authors with similar research interests. Another examples come from Drs. A. Spink and B. Shneiderman, in which the former researcher is an expert in information seeking behavior (Spink, Ozmutlu, & Ozmutlu, 2002; Spink & Cole, 2005) while the latter has been concentrating on user behavior analysis (Shneiderman, 1978). Although the nodes representing these two authors are not closely with each other in Figure 2, they move nearer in the visualization of MFTACA method. However, the distance between Drs. Spink and McCain becomes farther in Figure 3 than Figure 2, indicating that our proposed MFTACA method separates authors sharing different research interests in knowledge domain maps. All of these facilitate the quality of maps in terms of the clustering performance.

**MDS-measurement**

In order to evaluate the performance of our proposed method quantitatively, we employ multi-dimensional scaling measurement (MDS-measurement) (Bu et al., 2016) to supplement our argument in “Network Science” section that in a quantitative way. Note that MDS-measurement is NOT the same as MDS. The basic principle of MDS-measurement is to calculate the MDS-measurement value ($\sigma$), which is equal to the ratio between the sum of the distance between the nodes within the same cluster ($c$), and the sum of the distance between the nodes in different clusters ($S$). Intuitively, a smaller $\sigma$ indicates a better clustering performance in knowledge domain maps in which nodes within the same cluster lie nearer while those in different clusters...
lie farther. Table 3 shows the MDS-measurement result, where we can see that the MDS-measurement value ($\sigma$) of MFTACA is smaller than that of ACA, indicating a better clustering result in knowledge domain map. This confirms our observation in “Network Analysis” section.

Table 3. MDS-measurement results.

<table>
<thead>
<tr>
<th>Method</th>
<th>$c$</th>
<th>$S$</th>
<th>$\sigma$ ( = $c/S$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACA</td>
<td>546.61</td>
<td>4303.53</td>
<td>12.70%</td>
</tr>
<tr>
<td>MFTACA</td>
<td>513.17</td>
<td>4401.68</td>
<td>11.66%</td>
</tr>
</tbody>
</table>

Conclusions

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

Besides the method itself and its advantages compared with traditional ACA, this study provides several implications to the future researchers. First, we use full-text data but not intend to analyse content-level information, which breaks the conventional thinking to use complex natural language processing technologies to mine content- or semantic-level (Bu, Huang, Ding, & Ai, 2017) data so as to map knowledge domains. Second, such idea enables to be applied into not only ACA but also other scholarly network analyses, such as author bibliographic coupling analysis (Zhao & Strotmann, 2008a) and coauthorship analysis (Ding, 2011b). Last but not least, this research supplements the framework of bibliometric elements proposed by Morris and Vander Veer Martens (2008), in which papers, paper authors, paper journals, references, reference authors, reference journals, and index terms are included. Our study provides “citing sentences” as a bridge between “papers” and “references”, and shows the potential detailed affiliations about “citing sentences” such as the number of mentioned times and the number of context words of references. These have offered significant foundations for future supplements of the bibliometric element framework when more full-text data are involved.

However, there are several limitations in this research. For example, we only used first authors’ information instead of all authors’. The accuracy could thus be negatively affected. Moreover, there are still many other types of metadata that are not used to involve in ACA in previous studies and this paper, such as the sequence of co-cited authors (He, Ding, & Yan, 2012) and the number of figures or tables (Lee, West, & Howe, 2016). We would like to focus on these in the future.

References


