IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 1, 2019, accepted April 19, 2019, date of publication May 1, 2019, date of current version May 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2913995

Clustering Scientific Document Based on an Extended Citation Model

SHUAI ZHANG, YANGBING XU, AND WENYU ZHANG
School of Information, Zhejiang University of Finance and Economics, Hangzhou 310018, China

Corresponding author: Wenyu Zhang (wyzhang@e.ntu.edu.sg)

This work was supported by the National Natural Science Foundation of China under Grant 51875503 and Grant 51475410.

ABSTRACT With the number of published scientific paper increasing exponentially, scientific document clustering is becoming a challenging task. Therefore, a scientific document clustering model with high quality is needed. In this paper, we propose an extended citation model for scientific document clustering. On the one hand, the proposed model considers that 1) the high frequency and the wide distribution of a scientific document cited in other documents will result in the high similarity between the citing and the cited documents; and 2) the close location of two scientific documents cited in a scientific document will also result in the high similarity between these two documents. On the other hand, the proposed model combines a citation networks and textual similarity network to enhance the performance of scientific document clustering. To evaluate the performance of our proposed model, we collect scientific documents from PMC and PubMed databases in the field of oncology as a case study. It is proved that our proposed model can obtain reasonably clustering results by comparing it with traditional scientific documents clustering models, such as traditional bibliographic coupling model and textual similarity model, according to the indices of precision, recall, and F1-score.

INDEX TERMS Scientific document clustering, citation frequency analysis, citation distribution analysis, citation proximity analysis, textual similarity, random walk algorithm.

I. INTRODUCTION

In recent years, the number of published scientific documents has increased exponentially. It is difficult for researchers, especially for novice researchers, to detect the research fronts by human effort only. Therefore, an effective scientific document clustering model is in great demand for the detection of research fronts. In bibliometrics and informetrics fields, citation analysis is widely used to cluster scientific documents [1]–[5].

The three main citation networks are direct citation, co-citation, and bibliographic coupling, which are illustrated in Fig. 1. If document *A* cites document *B*, then this constitutes a direct citation; if two scientific documents are both cited by other documents, then this constitutes co-citation; and if two scientific documents both cite other documents, then these two scientific documents show bibliographic coupling. The examples of direct citation, co-citation, and bibliographic coupling, between documents *A* and *B* are illustrated in Fig. 1 (a-c), respectively.

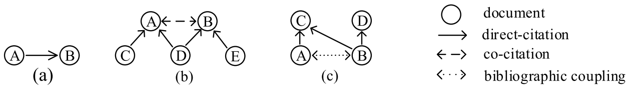


FIGURE 1. Simple examples of citation networks.

As we have described above, citation analyses have been widely used for document clustering. However, as Wan and Liu [6] described that "some citations are very import, but some are trivial," the scientific documents cited in a document are usually not equally important. Therefore, Wan and Liu [6] classified the strength of a citation into five parts, from very trivial to very important, according to the number of times it was cited in a document. In this study, we assume that the high frequency and the wide distribution of the scientific documents cited in other documents will result in high similarity between the citing and the cited document.

In addition, some researchers cluster documents based on their co-citation network [1], [7]–[9]. However, because of

The associate editor coordinating the review of this manuscript and approving it for publication was Dimitrios Katsaros.

the difficulties in obtaining full text documents and extracting useful information from different reference styles, the traditional co-citation model did not consider the proximity of two cited scientific documents, and assumes that two cited scientific documents with co-citation relationship have equal status, wherever they are cited in a document. However, some scientific documents tend to be cited in the same sentence, and others are cited in different sections. It was proved that the similarities between two cited scientific documents with co-citation relationship are related to their proximity [10]. In other words, with the citations becoming close to each other, they are more likely to be related.

To enhance the performance of scientific document clustering, some researchers combined scientific documents' citation network with their textual similarity [1], [4]. In addition, according to the studies of Boyack *et al.* [11], BM25 was proved as an effective approach for scientific document clustering by comparing it with other approaches such as term frequency-inverse document frequency vectors and latent semantic analysis. Therefore, we use BM25 to calculate scientific documents' textual similarity.

In this study, scientific documents form a complex network whose nodes represent scientific documents and whose edge strengths represent the similarity between two corresponding scientific documents. We aim to divide the scientific documents into several clusters so that the scientific documents in the same cluster share a similar theme. Therefore, we propose an extended citation model for scientific document clustering, which considers frequency and distribution of a scientific document cited in other scientific document, proximity of two scientific documents cited in a scientific document, and scientific documents' textual similarity.

The remainder of this article is organized as follows. Section 2 presents a review of some related works. Section 3 describes in detail our extended citation model. Section 4 introduces the data sources and data pre-processing. Section 5 discusses experiments on scientific document clustering. Finally, Section 6 concludes the study and discusses future research directions.

II. RELATED WORK

In previous works, some researchers proposed a weighted direct citation model, which combines direct citation approach with co-citation and bibliographic coupling approaches. For example, Small [12] proposed a combined linkage measure, which combines direct citation with co-citation, longitudinal coupling, and bibliographic coupling; Persson [13] integrated direct citations with shared cited scientific documents and co-citations into one measure of citation strength, and used weighted direct citation to identify research themes. However, both these studies [12], [13] overlooked the fact that the high frequency and the wide distribution of scientific documents cited in other scientific documents will result in high similarity between the citing and the cited documents [6]. Fujita *et al.* [14] applied some measures to weighted citation, such as average publication years, keyword similarity, the cited scientific document similarity, and frequency of citation. Chu and Yeh [15] considered the structure of articles, defined the citation strength by chapter level, and measured the similarity between documents based on cosine similarity. However, the frequency of citation considered by Fujita *et al.* [14] is measured by the total number of citation links, including the numbers of direct citation, co-citation, and bibliographic coupling. Fujita *et al.* [14] did not consider the frequency and distribution of the scientific documents cited in other scientific documents. Chu and Yeh [15] did not consider co-citation strength at the paragraph and sentence levels. Wan and Liu [6] assumed that the importance of citations is not only related to the mentioned times in the document, but also related to the sections they are cited, which indicated that the similarity between the citing and the cited documents relate to both the frequency and distribution of the scientific documents cited in other documents. Then, Wan and Liu [6] used citations with different weights to evaluate the influences of each paper and its author/s. However, they did not use the weighted direct citation approach in document clustering. Therefore, in this study, we assume that the citing and the cited documents are similar if the cited document is mentioned many times or in many sections in the citing document.

Gipp and Beel [10] proposed citation proximity analysis for document clustering. The essence of their study is that with the citations becoming close to each other, they are more likely to be related. They divided the proximity of two cited scientific documents into five levels (i.e., the cited scientific documents in the same sentence, those in the same paragraph, those in the same chapter, those in the same journal or book, and those in the same journal but different editions), and assigned different weights to each level (i.e., 1, 1/2, 1/4, 1/8, and 1/16, respectively). Liu and Chen [16] divided the proximity into four levels (i.e., the cited scientific documents in the same sentence, those in the same paragraph, those in the same section, and those in the same article), and found that the co-citations at the first level (i.e., the cited scientific documents in the same sentence) play a predominant role in forming an overall co-citation network. However, Liu and Chen [16] did not explain the weight of each level of proximity. Boyack *et al.* [17] selected relative distance rather than absolute distance to represent the cited scientific documents' proximity. They compared the proximity-based co-citation model with the traditional co-citation model and concluded that the use of the cited scientific documents' proximity information increases the accuracy of co-citation clustering. Kim *et al.* [18] revealed the implicit relationship in authors' subject disciplines based on the content and proximity of citation sentences. However, Kim *et al.* [18] focused on author clustering and only considered the cited scientific documents in the same section.

Besides the citation analysis, text analysis is also widely used to calculate document similarity. Lin and Wilbur [19] proposed a probabilistic topic-based model for content similarity and proved that it is an effective ranking algorithm for related article research. And Liu *et al.* [20] used probability-based text clustering algorithm to cluster large-scale documents. Boyack *et al.* [11] clustered more than two million biomedical documents based on their textual similarity and proved that BM25 is an effective algorithm for document clustering. In addition, some researchers [1], [4] combined citation analysis with text analysis to enhance the performance of clustering.

Boyack and Klavans [1] compared the accuracies of clustering solutions using four similarity approaches: co-citation, bibliographic coupling, direct citation, and a bibliographic coupling-based citation-text hybrid approach. Ahlgren and Colliander [21] focused on document similarity approaches and compared the text-based approach, citation-based bibliographic coupling approach, and the approach combining both text- and citation-based models. Both these studies [1], [21] concluded that the hybrid model performs better than the citation- or text-based model.

In line with these studies, we extend the citation model for scientific document clustering. This model is extended by considering frequency and distribution of the scientific documents cited in other scientific documents, considering proximity of two scientific documents cited in a scientific document, and combining citation networks with textual similarity network.

III. THE EXTENDED CITATION MODEL

In this study, we propose an extended citation model for scientific document clustering. The similarity between scientific documents based on the proposed model is composed of four networks: direct citation network, co-citation network, bibliographic coupling network, and textual similarity network.

A. DIRECT CITATION NETWORK

A simple approach to calculate the similarity between scientific documents based on direct citation network is to use 1 to represent two documents in direct citation and 0 otherwise. Fig. 1 (a) shows an illustrative example of the traditional direct citation model, which assumes that document A cites document B. Then, the similarity between documents A and B is 1, regardless of the frequency or distribution of document B cited in document A.

In this study, we assume that the similarity between scientific documents with direct citation relationship is reflected in two aspects (i.e. the frequency and distribution of the scientific documents cited in other scientific documents). We use $S^{di} = \left[S^{di}_{i,j}\right]$ to denote an N × N similarity matrix of direct citation network, where N is the number of target scientific documents in this study, and each element $S^{di}_{i,j}$ in the matrix represents the similarity between documents i and j.

$$S^{di}_{i,j} = \frac{t_{i,j}}{T_i}\log_2\left(\frac{m_{i,j}}{M_i}+1\right)+\frac{t_{j,i}}{T_j}\log_2\left(\frac{m_{j,i}}{M_j}+1\right) \tag{1}$$

where $t_{i,j}$ represents the number of times that document j is cited in document i, and $t_{j,i}$ represents the number of times that document i is cited in document j; T_i represents the total number of citations of document i, and T_j represents the total number of citations of document j; $m_{i,j}$ represents the number of sections of document i, which contain the cited document j, and $m_{j,i}$ represents the number of sections of document j, which contain the cited document i; M_i represents the total number of sections of document i, which contain cited scientific documents, and M_j represents the total number of sections of document j, which contain cited scientific documents. It is obvious that the similarity matrix is a symmetric matrix (i.e. the similarity between documents i and j is the same as that between documents j and i). On the other hand, considering that most of scientific documents have many sections, but only one or two of them cites the same scientific documents, to make the document-document similarity more stable, which considers the distribution of scientific documents cited in the document, we use logarithmic function to calculate it. According to (1), the range of $S^{di}_{i,j}$ is from 0 to 1, and the high value of $S^{di}_{i,j}$ represents the high similarity between documents i and j. Our direct citation network reflects the assumption that the high frequency and the wide distribution of document j cited in document i will result in high similarity between documents j and i.

B. CO-CITATION NETWORK

A popular approach to calculate similarity between scientific documents based on the traditional co-citation network is shown in (2).

$$S^{tco}_{i,j} = \frac{\left|C_i \cap C_j\right|}{\left|C_i \cup C_j\right|} \tag{2}$$

where $S^{tco}_{i,j}$ represents the similarity between scientific documents i and j based on the traditional co-citation network, and C_i and C_j represent the collection of documents that cite documents i and j, respectively. Therefore, the similarity between documents A and B shown in Fig. 1 (b) is 1/3. According to (2), the traditional co-citation model neither considers the frequency and distribution of the scientific documents cited in a document, nor considers the proximity of two cited scientific documents.

Therefore, the co-citation network proposed in this study regards the proximity of the cited scientific documents with co-citation relationship as a key factor in calculating the similarities between them. Based on the studies of Gipp and Beel [10] and Liu and Chen [16], we divide the proximity into four levels (i.e., the cited scientific documents in the same sentence, those in the same paragraph, those in the same section, and those in the same document) and assign different weights for each level, which are presented in Table 1. If two cited scientific documents meet more than one level in the document, then the level with a higher value is selected [16]. We use $S^{co} = \left[S^{co}_{i,j}\right]$ to denote an N × N similarity matrix of co-citation network, where each element $S^{co}_{i,j}$ in the matrix represents the similarity between documents i and j with co-citation relationship. The mathematical expression of proposed co-citation network is shown

TABLE 1. Proximity of two cited scientific documents and their corresponding values.

Level	Value	Note
1	1	Cited in the same sentence
2	1/2	Cited in the same paragraph
3	1/4	Cited in the same section
4	1/8	Cited in the same document

in (3)-(6).

$$C_1 = C_i \cap C_j \tag{3}$$

$$C_2 = C_i - C_j \tag{4}$$

$$C_3 = C_j - C_i \tag{5}$$

$$S_{i,j}^{co} = \frac{\sum_{k \in C_1} min\left(S_{k,i}^{di}, S_{k,j}^{di}\right) P_{i,j,k}}{\sum_{k \in C_1} max\left(S_{k,i}^{di}, S_{k,j}^{di}\right) P_{i,j,k} + \sum_{k \in C_2} S_{k,i}^{di} + \sum_{k \in C_3} S_{k,j}^{di}} \tag{6}$$

where C_1 represents the collection of documents that cite documents i and j, C_2 represents the collection of documents that cite documents i but do not cite document j, C_3 represents the collection of documents that cite documents j but do not cite document i, k represents the document in the corresponding collection, and $P_{i,j,k}$ represents the proximity of the cited documents i and j in document k. The value of $P_{i,j,k}$ is calculated according to Table 1. According to (6), the similarity between scientific documents not only depends on the number of documents that cite both of them and the number of documents that cite at least one of them, but also on the proximity between them in a document.

C. BIBLIOGRAPHIC COUPLING NETWORK

A popular approach to calculate similarity between scientific documents based on the traditional bibliographic coupling network is shown in (7).

$$S_{i,j}^{tbi} = \frac{\left|C_i^{'} \cap C_j^{'}\right|}{\left|C_i^{'} \cup C_j^{'}\right|} \tag{7}$$

where $S_{i,j}^{tbi}$ represents the similarity between documents i and j based on the traditional bibliographic coupling network, and $C_i^{'}$ and $C_j^{'}$ represent the collection of documents that are cited by documents i and j, respectively. Therefore, the similarity between documents A and B shown in Fig. 1 (c) is 1/2. According to (7), the traditional bibliographic coupling model does not consider the frequency and distribution of a scientific document cited in a document.

Therefore, we extend the traditional bibliographic coupling and use $S^{bi} = \left[S_{i,j}^{bi}\right]$ to denote an N × N similarity matrix of bibliographic coupling network. Each element $S_{i,j}^{bi}$ in the matrix represents the similarity between documents i and j based on bibliographic coupling network. The mathematical expression of the proposed bibliographic coupling network is shown in (8)-(11).

$$C_1^{'} = C_i^{'} \cap C_j^{'} \tag{8}$$

$$C_2^{'} = C_i^{'} - C_j^{'} \tag{9}$$

$$C_3^{'} = C_j^{'} - C_i^{'} \tag{10}$$

$$S_{i,j}^{bi} = \frac{\sum_{k \in C_1} min\left(S_{i,k}^{di}, S_{j,k}^{di}\right)}{\sum_{k \in C_1} max\left(S_{i,k}^{di}, S_{j,k}^{di}\right) + \sum_{k \in C_2} S_{i,k}^{di} + \sum_{k \in C_3} S_{j,k}^{di}} \tag{11}$$

where C_1 represents the collection of documents that are cited by documents i and j, C_2 represents the collection of documents that are cited by documents i but are not cited by document j, C_3 represents the collection of documents that are cited by documents j but are not cited by document i, and k represents the document in the corresponding collection.

It is obvious that the procedure of calculating citation similarity has quadratic complexity. To reduce the computing memory consumption and the clustering algorithm's CPU time, a blocking technique is used. For example, we use a series of lists, in which each element represents the cited scientific document of the corresponding document. In addition, only the scientific documents which have citation relationship will be considered. It has greatly saved the computer processing time because of the sparse citations among scientific documents.

D. TEXTUAL SIMILARITY NETWORK

Although BM25 is widely used to rank matching documents, it is also suitable for scientific documents clustering, especially for clustering the documents with large document set [11]. For example, Boyack *et al.* [11] proved BM25 is an effective approach to cluster scientific documents by comparing it with other well-known analytical techniques, including cosine similarity using term frequency-inverse document frequency vectors, latent semantic analysis, and topic modeling. Because of the limited space, other text similarity metrics, such as cosine similarity, edit distance, and jaccard index, are not considered in this study. Therefore, we collect the documents' abstracts and adopt BM25 to calculate their textual similarity. We use $S^t = \left[S_{i,j}^t\right]$ to denote an N × N similarity matrix of text analysis. Each element $S_{i,j}^t$ in the matrix represents the similarity between documents i and j based on BM25. The mathematical expression of $S_{i,j}^t$ is shown in (12) and (13) [11].

$$S_{i,j}^t = \sum_{x=1}^{n} \left(IDF_x \frac{n_x (k_1 + 1)}{n_x + k_1 \left(1 - b + b\, |L| / \bar{L}\right)} \right) \tag{12}$$

$$IDF_x = \log_{10} \frac{D - d_x + 0.5}{d_x + 0.5} \tag{13}$$

where n represents the number of terms, n_x represents the frequency of term x in the document j. $|L|$ and $\bar{L}$ represent the length of document j and the average length of documents, respectively, and the length of a document is measured by the sum of terms in the document. k_1 and b are two parameters of BM25, and they are set as 2.0 and 0.75, respectively. IDF_x is the inverse document frequency of term x, which is calculated by (13). In (13), D represents the number of documents and d_x represents the number of documents that contain term x. In this study, only the terms whose inverse document frequency values are greater than 2 are considered [11].

E. THE EXTENDED CITATION MODEL

We use $S^c = \left[S^c_{i,j}\right]$ to denote an N × N similarity matrix of citation networks, including direct citation network, co-citation network, and bibliographic coupling network. The element $S^c_{i,j}$ represents the similarity between documents i and j based on the citation networks, and $S^c_{i,j}$ is calculated as shown in (14).

$$S^c_{i,j} = v_1 S^{di}_{i,j} + v_2 S^{co}_{i,j} + v_3 S^{bi}_{i,j} \tag{14}$$

where v_1, v_2, and v_3 $(v_1 + v_2 + v_3 = 1)$ represent the weights of the direct citation network, co-citation network, and bibliographic coupling network, respectively.

Glänzel and Thijs [2] proposed a hybrid similarity model in 2011, which integrated the similarities calculated by the citation model and the content model. In this study, we use $S = \left[S_{i,j}\right]$ to denote an N × N similarity matrix of the extended citation model. Each element $S_{i,j}$ represents the similarity between documents i and j. Due to the different metrics of the citation networks and the textual similarity network, we must normalize the similarity calculated by these two approaches. The normalizing procedure is shown in (15), and $S_{i,j}$ is calculated as shown in (16) [2], [22].

$$S^{nv} = S^{ov} / S^{maxv} \tag{15}$$

$$S_{i,j} = \cos\left(\begin{array}{c} \lambda \arccos\left(S^{nc}_{i,j}\right) + \\ (1-\lambda) \arccos\left(S^{nt}_{i,j}\right) \end{array}\right), \quad \lambda \in [0, 1] \tag{16}$$

where S^{nv} and S^{ov} are similarity values after and before normalizing, S^{maxv} is the maximum similarity in the similarity matrix, λ is the weight of the citation networks, and $S^{nc}_{i,j}$ and $S^{nt}_{i,j}$ represent the normalized similarity between documents based on the citation networks and the textual similarity network, respectively.

IV. DATA SOURCES AND DATA PRE-PROCESSING

A. DATA COLLECTION

To evaluate the practicability of the extended citation model, a case study with open-access and full-text scientific documents is needed. PMC database (https://www.ncbi.nlm.nih.gov/pmc) and PubMed database (https://www.ncbi.nlm.nih.gov/pubmed) are two popular databases in the biomedical field and provide a large number of open-access and full-text scientific documents.

In this study, we collect scientific documents in the field of oncology between the period 2007-2016 from PMC database. Following Kim *et al.* [18], we select 14 journals. As listed in Table 2, we downloaded 12,356 open-access and full-text documents in xml format. Of these, we considered 10,996 scientific documents with article type classified as research article for document clustering. Documents whose article type includes letter, overview, or editorial are not considered in this study, because the formats of letter and editorial are flexible, and the overview contains numerous cited scientific documents that are not suitable for our present model. The details of the collected documents are listed in Table 2, which includes the journal title, the total number of open-access and full-text documents, and the number of open-access and full-text documents classified as research article. In addition, we extract PubMed IDs from 10,996 documents and adopt Batch Entrez (https://www.ncbi.nlm.nih.gov/sites/batchentrez) to collect the abstracts and Medical Subject Headings (MeSH) terms from PubMed database according to their corresponding PubMed IDs. As a result, the scientific documents collected from PubMed and PMC databases are matched.

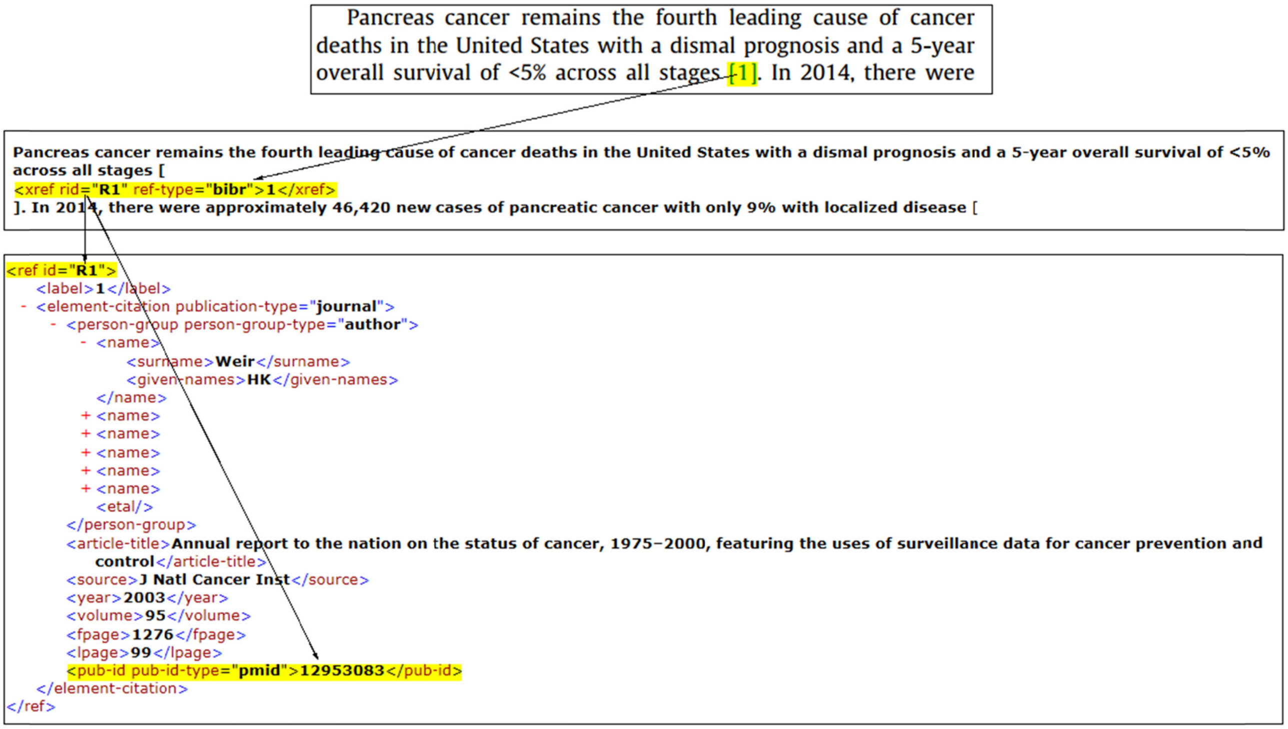
TABLE 2. Information of journals selected from PMC database.

Journal title	Number of full-text documents	Number of research document
Oncotarget	8,119	7,458
Molecular Cancer	1,461	1,340
Oncogene	1,079	1,055
Leukemia	476	376
OncoImmunology	523	194
Annals of Oncology	190	155
Journal of the National Cancer Institute	166	145
Clinical Epigenetics	125	84
Breast Cancer Research	57	45
Neuro-Oncology	43	38
Journal of Thoracic Oncology	40	36
Cancer Cell	35	33
Stem Cells	25	22
Molecular Oncology	17	15
Total	**12,356**	**10,996**

To extract the information of the frequency and distribution of scientific documents cited in other documents and proximity between two cited scientific documents, we use Java to find the location of the scientific documents cited in a document. Most scientific documents in PMC database have a similar reference style, which is shown in Fig. 2. To simplify our work, the scientific documents cited in tables or figures are not considered in this study, and there are seven documents that do not cite scientific documents.

B. DATA PRE-PROCESSING

In this sub-section, the details of data pre-processing will be discussed, including numbering scientific documents, constructing citation networks without isolated scientific documents, and text pre-processing.



Pancreas cancer remains the fourth leading cause of cancer deaths in the United States with a dismal prognosis and a 5-year overall survival of <5% across all stages [1]. In 2014, there were

```
Pancreas cancer remains the fourth leading cause of cancer deaths in the United States with a dismal prognosis and a 5-year overall survival of <5%
across all stages [
<xref rid="R1" ref-type="bibr">1</xref>
]. In 2014, there were approximately 46,420 new cases of pancreatic cancer with only 9% with localized disease [
```

```
<ref id="R1">
   <label>1</label>
 - <element-citation publication-type="journal">
   - <person-group person-group-type="author">
     - <name>
          <surname>Weir</surname>
          <given-names>HK</given-names>
       </name>
     + <name>
     + <name>
     + <name>
     + <name>
     + <name>
       <etal/>
     </person-group>
     <article-title>Annual report to the nation on the status of cancer, 1975–2000, featuring the uses of surveillance data for cancer prevention and
       control</article-title>
     <source>J Natl Cancer Inst</source>
     <year>2003</year>
     <volume>95</volume>
     <fpage>1276</fpage>
     <lpage>99</lpage>
     <pub-id pub-id-type="pmid">12953083</pub-id>
   </element-citation>
</ref>
```

FIGURE 2. **Example of reference style in xml format.**

As some scientific documents do not contain PubMed IDs, we endow each of them with a unique integer as their identification number, such that only the same documents have the same identification numbers. In this study, two scientific documents are regarded as different documents if they meet one of the following constraints: 1) if both of them contain PubMed IDs, but their PubMed IDs are different; 2) if both of them contain PMC IDs, but their PMC IDs are different; 3) if both of them contain DOI, but their DOI are different; and 4) if their titles or authors are different. Finally, we obtain a database that contains 287,701 unique scientific documents, which includes 10,996 target documents and 276,705 support documents (i.e. the documents that are cited by target documents but are not regarded as target documents).

To simplify the work, we find and eliminate isolated scientific documents by constructing the direct citation network, co-citation network, and bibliographic coupling network. An isolated scientific document is a document that does not have citation relationship with other scientific documents, and there are 30 isolated scientific documents in our case study. We eliminate these isolated documents and obtain a database that contains 286,895 unique scientific documents, including 10,966 target documents and 275,929 support documents. Then, we reconstruct direct citation network (number of edges: 8,930), co-citation network (number of edges: 8,103), and bibliographic coupling network (number of edges: 1,288,609), according to (1), (6), and (11), respectively. It is interesting that the number of edges of bibliographic coupling is higher than direct citation and co-citation. We think this phenomenon is reasonable, and it is caused by the following three aspects. 1) The data we collect cover the last 10 years, however some support documents in the database were published 10 years ago. 2) In this study, we select 14 journals in the oncology field as a case study, but there are many related papers that were published in other journals and were not considered. 3) Each edge of direct citation requires the target document is cited by other target documents, each edge of co-citation requires two target documents are cited by at least one of target documents, and if there exists one document in the database that is cited by two target documents, then these two documents are bibliographic coupling. However, the number of target documents (10,966) is much lesser than the scale of database (286,895).

In addition, we use R programming to pre-process the abstract. We eliminate figures, punctuations, duplicate whitespaces, and stop-terms, such as "the," "then," and "and," according to the SMART information retrieval system (http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop), and we transform alphabets from uppercase to lowercase. To maintain uniformity among words that are almost similar but used in different styles, such as "like," "likes," and "liked," the porter stemming algorithm [23] is applied in this study. We calculate the inverse document frequency for each term in abstract, and retain terms whose inverse document frequency values are greater than 2. Finally, we construct the textual similarity network (number of edges: 2,634,389) according to (12). We integrate the citation network, co-citation network, bibliographic coupling network, and textual similarity network according to (16). In addition, we upload the raw data used in this study into figshare.com (https://doi.org/10.6084/m9.figshare.7634489) to facilitate experimental verification.

V. EXPERIMENTS AND RESULTS

In this section, we prove the practicability of our proposed extended citation model by comparing it with the traditional bibliographic coupling model and the textual similarity model (i.e., λ in (16) is set as 0) for scientific document clustering. We use random walk [24] in R programming on a personal computer with Windows 7 64-bit, 1.60 GHz Intel (R) Core CPU, and 4 GB RAM to cluster the scientific documents and display the experimental results.

The random walk algorithm [24] is a popular community detection algorithm whose input is the similarity network and output is the clustering results. Therefore, we construct the network whose nodes represent scientific documents and whose edge strengths represent the similarity between two corresponding scientific documents. Because of the limited computer memory and the large target documents set, the list of three elements $< N_i, N_j, S_{i,j} >$ instead of the adjacency matrix is used to represent the network, where N_i, N_j, and $S_{i,j}$ represent ith document, jth document, and their similarity, respectively, and $i < j$. In addition, to reduce the clustering algorithm's CPU time and achieve a high level of clustering performance, the similarity threshold is used in this study by disregarding the document-document similarity which is less than 0.0001.

After a series of experiments, we set v_1 (i.e., the weight of direct citation network), v_2 (i.e., the co-citation network), and v_3 (i.e., the weight of bibliographic coupling network) as 0.3, 0.3, and 0.4, respectively, and we set λ as 0.3 for our presented model.

A. EVALUATION INDICES

As MeSH terms are annotated by experts and not used to construct similarity networks, we regard MeSH terms of each scientific document as the classifications of documents and use them to evaluate the clustering solutions based on different models [25]. To simplify the problem, only the descriptor terms are used in this study, and all MeSH terms are ignored in more than 5% of the documents. Then, the top three most frequent terms are regarded as themes in each cluster. In addition, we use precision, recall, and F1-score [26] as the evaluation indices for scientific document clustering.

In this study, first we assume that if two scientific documents share at least one MeSH term, they are considered to belong to the same category and are regarded as similar documents. Then, we calculate *TP*, *FP*, *FN*, and *TN* according to Table 3, where *TP* represents two similar documents belonging to the same cluster; *FP* represents two dissimilar documents belonging to the same cluster; *FN* represents two similar documents belonging to different clusters; and *TN* represents two dissimilar documents belonging to different clusters. Finally, we calculate precision, recall, and F1-score according to (17)-(19), respectively [26].

TABLE 3. Relationship between TP, FP, FN, and TN.

Two documents belong to the same cluster?	Two documents belong to the same category?	
	Yes	No
Yes	*TP*	*FP*
No	*FN*	*TN*

TABLE 4. Evaluation of citation networks with different weights of direct citation network, co-citation network, and bibliographic coupling network.

Models	Precision	Recall	F1-score
Citation networks (v_1=0.6, v_2=0.2, v_3=0.2)	0.1229	0.1794	0.1459
Citation networks (v_1=0.5, v_2=0.3, v_3=0.2)	0.1175	0.1887	0.1448
Citation networks (v_1=0.5, v_2=0.2, v_3=0.3)	0.1269	0.1843	0.1503
Citation networks (v_1=0.4, v_2=0.4, v_3=0.2)	0.1200	0.1773	0.1431
Citation networks (v_1=0.4, v_2=0.3, v_3=0.3)	0.1179	0.1758	0.1411
Citation networks (v_1=0.4, v_2=0.2, v_3=0.4)	0.1233	0.1504	0.1355
Citation networks (v_1=0.3, v_2=0.5, v_3=0.2)	0.1149	0.1921	0.1438
Citation networks (v_1=0.3, v_2=0.4, v_3=0.3)	0.1121	0.2025	0.1443
Citation networks (v_1=0.3, v_2=0.3, v_3=0.4)	**0.1242**	0.1989	**0.1529**
Citation networks (v_1=0.3, v_2=0.2, v_3=0.5)	0.1237	0.1606	0.1398
Citation networks (v_1=0.2, v_2=0.6, v_3=0.2)	0.1045	**0.2244**	0.1426
Citation networks (v_1=0.2, v_2=0.5, v_3=0.3)	0.1191	0.1936	0.1475
Citation networks (v_1=0.2, v_2=0.4, v_3=0.4)	0.1179	0.1412	0.1285
Citation networks (v_1=0.2, v_2=0.3, v_3=0.5)	0.1211	0.2048	0.1522
Citation networks (v_1=0.2, v_2=0.2, v_3=0.6)	0.1162	0.1970	0.1462

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN}, \tag{18}$$

$$\mathrm{F1-score} = \frac{2PrecisionRecall}{Precision + Recall} \tag{19}$$

B. PARAMETER SENSITIVITY ANALYSIS

Table 4 lists the precision, recall, and F1-score values based on citation networks with different weights of direct citation network, co-citation network, and bibliographic coupling network. As F1-score based on the citation networks with $v_1 = 0.3$, $v_2 = 0.3$, and $v_3 = 0.4$ is higher than others, we set v_1, v_2, and v_3 as 0.3, 0.3, and 0.4, respectively.

Table 5 lists the precision, recall, and F1-score values for different values of λ. According to Table 5, the F1-score based on the model with $\lambda = 0.3$ is higher than others. Therefore, we set $\lambda = 0.3$ in this study.

TABLE 5. Evaluation of extended citation model with different values of λ.

Models	Cluster number	Precision	Recall	F1-score
Extended citation model ($\lambda = 0.2$)	19	0.1213	0.3537	0.1806
Extended citation model ($\lambda = 0.3$)	11	0.1179	**0.4092**	**0.1831**
Extended citation model ($\lambda = 0.4$)	13	0.1183	0.3850	0.1810
Extended citation model ($\lambda = 0.5$)	17	0.1199	0.3598	0.1799
Extended citation model ($\lambda = 0.6$)	16	**0.1234**	0.3190	0.1780
Extended citation model ($\lambda = 0.7$)	20	0.1208	0.2089	0.1531
Extended citation model ($\lambda = 0.8$)	18	0.1151	0.2646	0.1604

TABLE 6. Evaluation of proposed model, traditional bibliographic coupling model, and textual similarity model.

Models	Precision	Recall	F1-score
Proposed extended citation model	0.1179	**0.4092**	**0.1831**
Traditional bibliographic coupling model	**0.1367**	0.1449	0.1407
Textual similarity model	0.1190	0.2573	0.1627

C. A COMPARATIVE ANALYSIS OF THE PROPOSED MODEL WITH OTHER MODELS

Before comparing the clustering solutions based on different models with the evaluation indices, we find that the number of clusters (8,295) and the number of clusters with one document (7,629) based on the traditional co-citation model are higher than others. Thus, using precision or recall to evaluate the performance of this model is meaningless. Therefore, we do not compare the proposed model with the traditional co-citation model in this study.

Table 6 lists the precision, recall, and F1-score values based on different models. According to Table 6, precision based on the traditional bibliographic coupling is higher than others, but recall based on it is lower than others. This phenomenon shows that there are many similar documents that are divided into different clusters, and their high values of precision are based on their small scale of clusters. And F1-score based on citation networks with $v_1 = 0.3$, $v_2 = 0.3$, $v_3 = 0.4$ (shown in Table 4) is higher than the F1-score based on traditional bibliographic coupling model. This means that considering frequency and distribution of scientific documents cited in other scientific documents and proximity of two scientific documents cited in a document does improve the quality of scientific document clustering model. In addition, our extended citation model has the highest F1-score, which means our proposed model does cover the shortage of traditional bibliographic coupling model and textual similarity model. Thus, it is proved that our proposed extended citation model can obtain reasonable clustering results.

TABLE 7. Evaluation of random walk and K-Means.

Algorithms (cluster number)	Precision	Recall	F1-score
Random walk (11)	**0.1179**	0.4092	**0.1831**
K-Means (10)	0.1098	0.4195	0.1740
K-Means (15)	0.1043	**0.5852**	0.1778
K-Means (20)	0.1108	0.3524	0.1686

D. A COMPARATIVE ANALYSIS OF RANDOM WALK ALGORITHM WITH K-MEANS

This study uses random walk algorithm to cluster scientific documents because random walk is one of the popular community detection algorithm and it does not need the preset number of clusters before clustering process. The well-known clustering algorithm K-Means [27] is compared to prove the effectiveness of random walk in clustering scientific documents. Because the numbers of clusters based on random walk ranges from 10 to 20, the numbers of clusters based on K-Means are set as 10, 15, and 20, respectively.

Table 7 lists the precision, recall, and F1-score values based on random walk algorithm and K-Means. According to Table 7, F1-score based on random walk is higher than others, which proves that the random walk algorithm used in this study is an effective clustering algorithm.

E. IDENTIFYING THE THEME AND REPRESENTATIVE DOCUMENTS OF CLUSTERS

Jarneving [28] stated that "making use of core documents for bibliometric mapping is a good choice in view of their perceived impact on current research." Therefore, it is important to find the representative document in the cluster. In this study, we calculate the degree centrality of each scientific document in the cluster, which represents the importance of the document in the cluster [29], [30], and it is calculated as (20).

$$DC_i = \delta\left(v_i, v_j\right) S_{i,j}, \quad \mathrm{i} \neq \mathrm{j} \tag{20}$$

where DC_i represents the degree centrality of ith document, v_i and v_j represents the cluster to which documents i and j has been assigned, respectively, and $\delta\left(v_i, v_j\right) = 1$, if $v_i = v_j$ (documents i and j belong to the same cluster), and 0 otherwise. Glänzel and Thijs [2] suggested that core documents should ideally represent about 0.1-1.0% of the original collection. Therefore, according to their degree centrality, we select 0.5% scientific documents in each cluster as representative documents. In addition, in this study, the top three most frequent MeSH terms in the cluster are selected as the cluster themes [25]. Owing to limited space, Table 8 lists the partial representative scientific documents in the top five largest clusters. Researchers may easily detect the research fronts in their fields by reading the representative documents in each cluster. According to Table 8, cluster A focuses on the research about oncogene protein, cluster B focuses on DNA, cluster C focuses on the outcome of cancer treatment, cluster

TABLE 8. Theme and representative documents of each cluster based on extended citation model.

Clusters (size)	Themes (the number of documents that contain it)	Representative documents' title (their degree centrality)
Cluster A (5157)	neoplasm metastasis (343) proto-oncogene proteins c-akt (329) down-regulation (293)	(a) Germline BRCA1 mutation reprograms breast epithelial cell metabolism towards mitochondrial-dependent biosynthesis: evidence for metformin-based "starvation" strategies in BRCA1 carriers (26.8802) (b) Rhabdomyosarcoma cells show an energy producing anabolic metabolic phenotype compared with primary myocytes (25.9114)
Cluster B (2697)	DNA methylation (184) promoter regions genetic (149) cell transformation neoplastic (145)	(a) Clinical impact of gene mutations and lesions detected by SNP-array karyotyping in acute myeloid leukemia patients in the context of gemtuzumab ozogamicin treatment: results of the ALFA-0701 trial (17.1789)
Cluster C (1955)	treatment outcome (205) kaplan-meier estimate (189) disease-free survival (184)	(a) First-in-human phase I study of copanlisib (BAY 80-6946), an intravenous pan-class I phosphatidylinositol 3-kinase inhibitor, in patients with advanced solid tumors and non-Hodgkin's lymphomas (23.4733)
Cluster D (713)	flow cytometry (74) mice inbred c57bl (71) killer cells natural (44)	(a) The IDO1 selective inhibitor epacadostat enhances dendritic cell immunogenicity and lytic ability of tumor antigen-specific T cells (17.1960)
Cluster E (253)	RNA long noncoding (74) enhancer of zeste homolog 2 protein (53) polycomb repressive complex 2 (50)	(a) Analysis of the polycomb-related lncRNAs hot air and antil in bladder cancer (10.5222)

D focuses on the research about cell, and cluster E focuses on RNA.

VI. DISCUSSIONS AND CONCLUSION

In this study, we not only considered the frequency and distribution of scientific documents that are cited in a document and the proximity between two scientific documents cited in a document, but also integrated text and citation analysis. We propose an extended citation model and demonstrate that it can obtain reasonable clustering results by comparing it with the traditional bibliographic coupling model and the textual similarity model. Following are some discussions of this study:

1) We use random walk algorithm in the *igraph* package of R programming to cluster scientific documents. The random walk algorithm is proved to be an effective algorithm to cluster scientific documents in this study, by comparing it with K-Means. However, there might be a more suitable algorithm for document clustering. Therefore, exploring the possibility of using other effective and efficient algorithms, such as hierarchical and leader clustering, for document clustering is one direction of our future research.

2) There are some limitations of our study that we plan to address through additional research. For example, scientific documents with letters, editorials, and overviews were excluded from the database in the current study. Thus, we aim to extend our model to cover documents with letters, editorials, overviews, and other document styles. Furthermore, this study focus on the scientific documents clustering based on the extended citation model which considers the frequency and distribution of scientific documents cited in other documents, the proximity of two scientific documents cited in a document, and the combination of citation analysis and textual analysis. Due to the limited space, we use the well-known BM25 to calculate scientific documents' textual similarity, but not consider other text similarity metrics such as cosine similarity, edit distance, Jaccard index and so on. Therefore, integrating other well-known text similarity metrics to enhance the performance of our proposed model is one of our future works. In addition, because of the limitation of technology, we do not consider the semantics of the citation sentence, which will again be dealt with in our future research.

SUPPORTING INFORMATION

The raw data for this paper have been uploaded to Figshare (https://doi.org/10.6084/m9.figshare.7634489), including the document collected from PMC database and their corresponding PubMed IDs.

REFERENCES

[1] K. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?" *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2389–2404, May 2010.

[2] W. Glänzel and B. Thijs, "Using 'core documents' for the representation of clusters and topics," *Scientometrics*, vol. 88, no. 1, pp. 297–309, 2011.

[3] W. Glänzel and B. Thijs, "Using 'core documents' for detecting and labelling new emerging topics," *Scientometrics*, vol. 91, no. 2, pp. 399–416, 2012.

[4] X. Liu, W. Glänzel, and B. De Moor, "Optimal and hierarchical clustering of large-scale hybrid networks for scientific mapping," *Scientometrics*, vol. 91, no. 2, pp. 473–493, 2012.

[5] L. Guo, X. Cai, F. Hao, D. Mu, C. Fang, and L. Yang, "Exploiting fine-grained co-authorship for personalized citation recommendation," *IEEE Access*, vol. 5, pp. 12714–12725, 2017.

[6] X. Wan and F. Liu, "Are all literature citations equally important? Automatic citation strength estimation and its applications," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 9, pp. 1929–1938, 2014.

[7] H. Small, "The relationship of information science to the social sciences: A co-citation analysis," *Inf. Process. Manage.*, vol. 17, no. 1, pp. 39–50, 1981.

[8] H.-Y. Li, L. Cui, M. Cui, and Y.-Y. Tong, "Active research fields of acupuncture research: A document co-citation clustering analysis of acupuncture literature," *Alternative Therapies Health Med.*, vol. 16, no. 6, pp. 38–45, 2010.

[9] H. Small, "Maps of science as interdisciplinary discourse: Co-citation contexts and the role of analogy," *Scientometrics*, vol. 83, no. 3, pp. 835–849, 2010.

[10] B. Gipp and J. Beel, "Citation proximity analysis (CPA) : A new approach for identifying related work based on co-citation analysis," in *Proc. 12th Int. Conf. Scientometrics Informetrics*, Rio de Janeiro, Brazil, Jul. 2009, pp. 571–575.

[11] K. W. Boyack *et al.*, "Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches," *PLoS ONE*, vol. 6, no. 3, 2011, Art. no. e18029.

[12] H. Small, "Update on science mapping: Creating large document spaces," *Scientometrics*, vol. 38, no. 2, pp. 275–293, 1997.

[13] O. Persson, "Identifying research themes with weighted direct citation links," *J. Informetrics*, vol. 4, no. 3, pp. 415–422, 2010.

[14] K. Fujita, Y. Kajikawa, J. Mori, and I. Sakata, "Detecting research fronts using different types of weighted citation networks," *J. Eng. Technol. Manage.*, vol. 32, pp. 129–146, Apr./Jun. 2014.

[15] K.-C. Chu and C.-C. Yeh, "Knowledge flow of biomedical informatics domain: Position-based co-citation analysis approach," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, San Francisco, CA, USA, Aug. 2016, pp. 1119–1126.

[16] S. Liu and C. Chen, "The proximity of co-citation," *Scientometrics*, vol. 91, no. 2, pp. 495–511, 2012.

[17] K. W. Boyack, H. Small, and R. Klavans, "Improving the accuracy of co-citation clustering using full text," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 64, no. 9, pp. 1759–1767, 2013.

[18] H. J. Kim, Y. K. Jeong, and M. Song, "Content- and proximity-based author co-citation analysis using citation sentences," *J. Informetrics*, vol. 10, no. 4, pp. 954–966, 2016.

[19] J. Lin and J. W. Wilbur, "PubMed related articles: A probabilistic topic-based model for content similarity," *BMC Bioinf.*, vol. 8, no. 1, p. 423, 2007.

[20] M. Liu, Y. Liu, B. Liu, and L. Lin, "Probability-based text clustering algorithm by alternately repeating two operations," *J. Inf. Sci.*, vol. 39, no. 3, pp. 372–383, 2013.

[21] P. Ahlgren and C. Colliander, "Document–document similarity approaches and science mapping: Experimental comparison of five approaches," *J. Informetrics*, vol. 3, no. 1, pp. 49–63, 2009.

[22] W. Glänzel and B. Thijs, "Using hybrid methods and 'core documents' for the representation of clusters and topics: The astronomy dataset," *Scientometrics*, vol. 111, no. 2, pp. 1071–1087, 2017.

[23] P. Willett, "The porter stemming algorithm: Then and now," *Program*, vol. 40, no. 3, pp. 219–223, 2006.

[24] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Proc. 20th Int. Symp. Comput. Inf. Sci.*, Istanbul, Turkey, Oct. 2005, pp. 284–293.

[25] L. Yeganova, W. Kim, K. Sun, and W. J. Wilbur, "Retro: Concept-based clustering of biomedical topical sets," *Bioinformatics*, vol. 30, no. 22, pp. 3240–3248, 2014.

[26] D. Yu, W. Wang, S. Zhang, W. Zhang, and R. Liu, "Hybrid self-optimized clustering model based on citation links and textual features to detect research topics," *PLoS ONE*, vol. 12, no. 10, 2017, Art. no. e0187164.

[27] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *J. Roy. Stat. Soc. C, Appl. Statist.*, vol. 28, no. 1, pp. 100–128, 1979.

[28] B. Jarneving, "Bibliographic coupling and its application to research-front and other core documents," *J. Informetrics*, vol. 1, no. 4, pp. 287–307, 2007.

[29] L. C. Freeman, "Centrality in social networks conceptual clarification," *Soc. Netw.*, vol. 1, no. 3, pp. 215–239, 1979.

[30] J. Nan, B. Xiao, Z. Lin, and Q. Xu, "Keywords extraction from Chinese document based on complex network theory," in *Proc. 7th Int. Symp. Comput. Intell. Design*, Hangzhou, China, Dec. 2014, pp. 383–386.

SHUAI ZHANG received the Ph.D. degree in mechanical engineering from Zhejiang University, China, in 2005. He is currently a full-time Professor with the School of Information, Zhejiang University of Finance and Economics, Hangzhou, China. He has published more than 30 papers in international journals in the recent ten years, including bibliometrics, supply chain management, business intelligence, data mining, E-government, and manufacturing informatization.

YANGBING XU is currently pursuing the M.S. degree with the School of Information, Zhejiang University of Finance and Economics, Hangzhou, China. His current research interests include data mining and bibliometrics.

WENYU ZHANG received the B.S. degree from Zhejiang University, China, in 1989, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2002. He is currently a full-time Professor with the School of Information, Zhejiang University of Finance and Economics, Hangzhou, China. He has published more than 40 papers in international journals and more than 20 papers in international conference proceedings in the recent ten years, including supply chain management, digital library, bibliometrics, concurrent engineering, distributed manufacturing, business intelligence, business analytics, data mining, multi-agent technology, and semantic Web.