Journal of Informetrics 10 (2016) 212–223

Contents lists available at ScienceDirect

Journal of Informetrics

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

Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods

Guo Chen [a], Lu Xiao [b,*]

[a] Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094, China
[b] School of Information Management, Nanjing University, Nanjing 210046, China

ARTICLE INFO

Article history:
Received 29 October 2014
Received in revised form 10 January 2016
Accepted 10 January 2016
Available online 30 January 2016

Keywords:
Domain analysis
Keyword analysis
Keyword Activity Index
Digital Library in China

ABSTRACT

Publication keywords have been widely utilized to reveal the knowledge structure of research domains. An important but under-addressed problem is the decision of which keywords should be retained as analysis objects after a great number of keywords are gathered from domain publications. In this paper, we discuss the problems with the traditional term frequency (TF) method and introduce two alternative methods: TF-inverse document frequency (TF-IDF) and TF-Keyword Activity Index (TF-KAI). These two methods take into account keyword discrimination by considering their frequency both in and out of the domain. To test their performance, the keywords they select in China's Digital Library domain are evaluated both qualitatively and quantitatively. The evaluation results show that the TF-KAI method performs the best: it can retain keywords that match expert selection much better and reveal the research specialization of the domain with more details.

1. Introduction

In bibliometric research, publication keywords are considered the basic elements of representing knowledge concepts and have been commonly used to reveal the knowledge structure of research domains (Su & Lee, 2010). Related studies address hotspot detection and trend analysis based on keyword frequency analysis (Chen, 2006; Xie, Zhang, & Ho, 2008), research topic analysis based on co-word clustering (Callon, Courtial, & Laville, 1991; Rip & Courtial, 1984), and knowledge mapping based on co-word networks (Assefa & Rorissa, 2013; Choi, Yi, & Lee, 2011). In these studies, there are two different approaches for using publication keywords, depending on the research aims: (1) using all keywords to explore the structural characteristics of domain knowledge at the macro-level and (2) using some "important" keywords to analyze the details of a domain's major research topics and their relation at the micro-level. Since many empirical studies are carried out at the micro-level, it is necessary to study the selection process of these "important" keywords for bibliometric analysis.

In previous studies, researchers have mainly focused on identifying research topics (for example, research theme clustering and network community discovering) and interpreting the results. Less attention has been given to the process of selecting appropriate keywords for future analysis. The "popular" keywords are usually considered important and are selected based on frequency or by using centrality-based network measures, both of which are proven to select very similar keywords (Choi et al., 2011). However, there are two problems with this selection process. (1) Keywords may be used frequently because

* Corresponding author. Tel.: +86 13601455621.
E-mail address: ahjk_xiaolu@163.com (L. Xiao).

http://dx.doi.org/10.1016/j.joi.2016.01.006
1751-1577/

they are generalizations or represent popular themes. These general keywords may be useful in showing a rough overview of a scientific discipline, but are less successful at displaying detailed themes of a research domain. For example, "library," "information resources," and "services" are high-frequency keywords in the Digital Library (DL) domain, but in fact they are likely universal concepts which also occur within many other research domains in Library and Information Science (LIS). Therefore, they are not good representations of DL research themes. (2) The importance of a keyword in domain analysis should not only be decided by its popularity. It has been generally accepted that the "discrimination" of terms (that is, to what degree they can help distinguish a domain from others) should also be considered.

Meanwhile, there are many well-developed methods for identifying important terms in other research, such as information retrieval, document clustering and so on. For example, TF-IDF is a method typically used for measuring the importance of a term in a document; LDA has been largely used to identify different topics and their representative terms; and the Activity Index has been utilized to identify the research emphasis of research groups. However, the data feature of publication keywords in bibliometrics is unique. As Ferrara and Salini (2012) concluded, harvesting full-text publications is a very complex task, so publication keywords used in previous studies are mostly assigned by author, generated by database, or extracted from publication title or abstract. Comparatively, these types of publication keywords in bibliometric analysis are very sparse, and keywords of a given publication are treated as equally important without appropriate weighting. Therefore, it is still unknown whether these methods can be utilized to select publication keywords for domain analysis in bibliometrics.

The goal of this paper is to resolve this problem by comparing three typical methods, which can be imported to select publication keywords in bibliometric research for domain analysis. By investigating the keywords selected by these methods, we can understand and optimize their utilization.

2. Related work

2.1. Identifying important publication keywords in bibliometrics

As Choi et al. (2011) summarized, efforts to identify important publication keywords in bibliometrics are either popularity-based (mainly using keyword frequency) or network-based (mainly using centrality measures).

High-frequency keywords are usually identified as important research themes for bibliometric analysis. A preset number or frequency threshold is usually used to filter the keywords. For example, Zhao and Wang (2011) selected keywords with frequency above 60 to analyze the research foci of pervasive and ubiquitous computing. Niu et al. (2014) selected the top 30 high-frequency keywords to find significant differences between geosciences, multidisciplinary, and environmental sciences. In some academic databases, high-frequency keywords are utilized to analyze research hotspots and develop domain trends (Su, Deng, & Shen, 2014).

Another common approach to select publication keywords in bibliometrics is to utilize centrality-based network measures. There are many well-defined and widely-used measures for identifying important nodes in a network, including but not limited to: degree centrality, betweenness centrality, closeness centrality (Freeman, 1979), eigenvector centrality (Newman, 2008), and k-core value of nodes (Pittel, Spencer, & Wormald, 1996). However, keywords selected by such metrics are observed to be roughly similar with high-frequency keywords, due to the positive correlations between their frequency and those network measures (Choi et al., 2011; Liu, Guo, Lin, & Ma, 2013). Thus, the problem of high-frequency keywords still exists when using these centrality-based network measures.

Using high-frequency keywords as analysis objects has long been questioned. Early research in information retrieval finds that mid-frequency terms can provide the highest discrimination values (Luhn, 1958; Salton, 1975). Quoniam, Balme, Rostaing, Giraud, and Dou (1998) argued that some low-frequency keywords can express emerging new concepts, and by solely analyzing well-known keywords, classical bibliometric techniques ignore the consideration of innovative aspects. Milojević, Sugimoto, Yan, and Ding (2011) revealed that some high-frequency keywords are nonspecific words, and therefore are of limited value in bibliometric analysis. People are able to recognize a field by certain specific words (Rokaya, Atlam, Fuketa, Dorji, & Aoe, 2008), so these words have been highlighted in many technologies, such as information retrieval, text classification, and term extraction. A basic idea in such research is to show preference to terms that appear frequently in a given document/category/domain and rarely in others (Brunzel & Spiliopoulou, 2007).

TF-IDF (Salton & Buckley, 1988) is a typical method of identifying important terms by combining their popularity and discrimination. The TF-IDF value of a term can be calculated as:

$$\text{TF-IDF}_{t,d} = f_{t,d} \times \log \frac{N}{\text{d}f_t} \tag{1}$$

where, $f_{t,d}$ is the frequency of word t in the document d, N is the total number of documents, and $\text{d}f_t$ is the number of documents containing word t. The TF-IDF method can also be applied to bibliometric analysis. For example, Jabłońska-Sabuka, Sitarz, and Kraslawski (2014) used TF-IDF to identify "informative" words from publication keywords of the research domain, so as to predict research trends. Roche, Besagni, François, Hörlesberger, and Schiebel (2010) used TF-IDF to select publication keywords of scientific fields, and then categorized them into unusual terms, established terms, and cross section terms. De Battisti, Ferrara, & Salini, 2015 selected the most relevant publication keywords for different topics according to a measure similar to TF-IDF.

2.2. Identifying research emphasis with Activity Index

The Activity Index (AI) is a version of the economists' Comparative Advantage Index (Frame, 1977; Schubert & Braun, 1986). It measures whether a country/region has alternatively comparative advantages in a particular field according to its share in total world publications. AI is defined as:

$$\mathrm{AI} = \frac{\text{the share of the given country in publications in the given field}}{\text{the share of the given country in publications in all science fields}} \tag{2}$$

AI > 1 indicates that the country/region emphasizes a given scientific field in comparison with its average research level. AI < 1 indicates that the country/region has loose research in the field in comparison with its average research level.

AI has been widely utilized to reveal the research profile of countries/regions. Thijs and Glänzel (2008) used AI to describe the national profile of eight European countries' research fields. López-Illescas, de Moya-Anegón, and Moed (2011) used AI as an indicator to investigate university research performance, taking into account discipline specialization. Pouris and Ho (2014) used AI to identify the emphasized and underemphasized research fields of Africa. Harzing and Giroud (2014) used AI to reveal the competitive advantages of nations in different academic disciplines.

Recently, Chen, Xiao, Hu, and Zhao (2015) have extended AI to identify the topic emphasis of different research institutions. This idea is also useful when identifying emphasized research themes in a domain. By viewing the researchers in a given domain as a group, their research theme preferences can be identified by measuring the AI of different keywords in the domain. Accordingly, we define the Keyword Activity Index (KAI) of a keyword in a given domain as:

$$\mathrm{KAI} = \frac{\text{the share of the given domain in publications containing the given keyword}}{\text{the share of the given domain in all publications}} \tag{3}$$

Thus, $n(i, j)$ denotes the number of publications containing the keyword i in domain j; $n(i, \mathrm{all})$ denotes the total number of publications containing the keyword i in the background corpus; $n(j)$ denotes the number of publications in domain j; and $n(\mathrm{all})$ denotes the total number of publications in the background corpus. Accordingly, formula (3) can be rewritten as:

$$\mathrm{KAI} = \frac{n(i,j)/n(i,\mathrm{all})}{n(j)/n(\mathrm{all})} \tag{4}$$

The KAI can help us measure the degree to which a keyword is emphasized by researchers in a given domain. KAI > 1 indicates that the research theme is emphasized in the domain as compared to its average level in all domains; KAI < 1 indicates that the research theme is underemphasized in the domain.

2.3. Identifying representative terms of topics with LDA

LDA (Blei, Ng, & Jordan, 2003) is a type of topic modeling technique which uses a generative probabilistic model to assign documents (or any other objects consisting of a set of terms; for example, journals, institutions, authors, and so on) to clusters. In LDA, each document is assumed to be characterized by a particular set of topics, each topic has a probability of generating various words, and the topic distribution is assumed to have a Dirichlet prior. LDA has been broadly applied in topic detection and tracking. Previous studies have found that LDA performs well in analyzing the rich underlying structures of the domain (Griffiths & Steyvers, 2004). Ferrara and Salini (2012) used LDA to build the Multidimensional representation of textual information and topics for bibliometric analysis, in which each keyword is associated with a relevant topic. Piepenbrink and Nurmammadov (2015) used LDA to identify topics in the literature of transition economies and emerging markets; they showed details of each topic by displaying their top ten most representative words. Some extended LDA models have been proposed according to different research needs, such as the Author-Topic model (Ding, 2011; Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004) and the Author-Conference-Topic model (Tang, Jin, & Zhang, 2008). In these, important words are also identified based on their relevance to different topics. For example, Sugimoto, Li, Russell, Finlay, and Ding (2011) ranked the top 20 words by probability value in each topic of different time periods when analyzing North American Library and Information Science dissertations. Recently, Yan (2015) found that topic-based approaches can reveal the research dynamics, impact and dissemination of the selected database, which highlights topics as a valuable unit of analysis.

3. Methodology and data

3.1. Three methods of identifying important keywords for bibliometric analysis

In this paper, we will compare three methods of identifying important keywords for bibliometric analysis: the traditional TF method (using keyword frequency), the TF-IDF method, and the TF-KAI method. The LDA approach is not included because it requires a large amount of text for training in order to generate meaningful outcomes, while the limited use of keyword content lacks enough training content.

The TF method accounts for the status of publication keywords only within the domain, while the latter two consider the status of keywords both inside and outside the domain. A background corpus needs to be constructed to help highlight

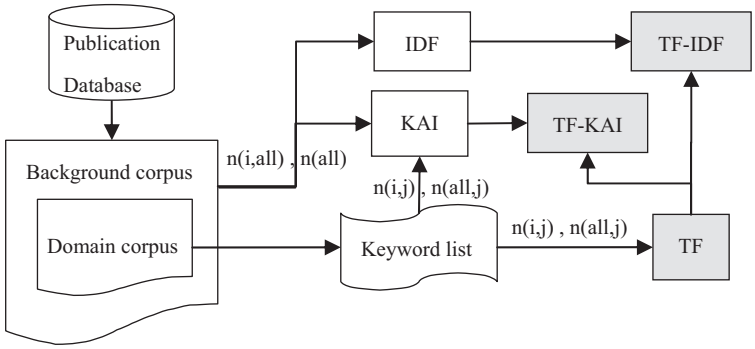

Fig. 1. Procedures of the three methods of identifying important keywords for bibliometric analysis.

these specific keywords. The background corpus consists of publications on a broader scale; for example, the background discipline of the given domain. Fig. 1 demonstrates the procedures of the three methods of identifying important keywords.

Firstly, we will select a given number (recorded as n) of high-frequency keywords according to their distribution in the dataset; then, the same number of keywords will be selected separately by using the other two methods.

In the TF-IDF method, the domain will be viewed as a document whose content consists of all the keywords of the domain publications, so that each keyword can be weighted using TF-IDF. Accordingly, in formula (1), $f_{t,d}$ can be rewritten as $n(i,j)$, and $\log(N/df_t)$ can be rewritten as $\log(n(\text{all})/n(i,\text{all}))$, so the calculation of TF-IDF can be modified into formula (5). Then the top n keywords with the highest TF-IDF value will be selected for future comparison.

$$\text{TF} = \text{IDF}_{i,j} = n(i,j) \times \log \frac{n(\text{all})}{n(i,\text{all})} \tag{5}$$

In the TF-KAI method, the idea of weighting keywords is very similar to TF-IDF. As discussed in Section 2.3, a higher KAI value indicates that the domain has a stronger preference for the research theme, meaning that keywords with high KAI value can represent the topical specialization of a research domain. Therefore, we use KAI instead of IDF for keyword discrimination. Combined with the TF factor, the TF-KAI weights are represented in formula (6). Note that the total number of papers in the given domain is a constant value so that $n(j)$ has no effect on ranking keywords with TF-KAI value. The top n keywords with the highest TF-KAI value will be selected for later comparison.

$$\text{TF} - \text{KAI}_{i,j} = n(i,j) \times \frac{n(i,j)/n(i,\text{all})}{n(j)/n(\text{all})} = \frac{n(i,j)^2}{n(j)} \times \frac{n(\text{all})}{n(i,\text{all})} \sim n(i,j)^2 \times \frac{n(\text{all})}{n(i,\text{all})} \tag{6}$$

Comparing formula (7) with formula (6), we can see that the two factors in the TF-IDF and TF-KAI methods are similar: popularity factors for both are computed based on $n(i,j)$ and discrimination factors for both are computed based on $n(\text{all})/n(i,\text{all})$. However, the weight schemes are different. To make the comparison more clearly, we can square the TF-IDF formula ($\text{TF} - \text{IDF}^2_{i,j} = n(i,j)^2 \times \log^2 \frac{n(\text{all})}{n(i,\text{all})}$). This will not change the keyword rank in the TF-IDF method, so that only one difference between these two methods is retained. Compared to the TF-KAI formula, we can see that the discrimination factor in the TF-IDF method is heavily weakened after the logarithmic operation of $n(\text{all})/n(i,\text{all})$.

3.2. Data

To achieve the best result evaluation, we chose a research domain with which we are familiar, in order to select a representative data set of China's Digital Library (DL) research domain for empirical study. The Chinese Journal Full-Text Database (CJFTD) was used as the data source. Similar to prior research, we use papers published in all core journals of LIS in China to represent a discipline, and papers containing given keywords to represent a research domain. To construct the background corpus of LIS in China, we collected related papers by setting the data source category as "core journals" in the LIS field. This includes all 18 Chinese journals of LIS research, with a set time span from 2000 to 2013. To construct the domain corpus, we selected papers by retrieving "digital library" (in Chinese) within the title or keywords in these core journals.

To calculate the KAI and IDF of a keyword, we need a background publication set which can provide information of its occurrence outside the domain. In scientific research, a sensible background range of a domain should be its background discipline. In this study, we will evaluate the keyword selection results of DL research in the context of LIS in China. The background corpus is composed of all publications in the LIS core journals in China.

The dataset features of author keywords, indexer keywords and title/abstract extracted keywords are quite similar in bibliometric analysis: they are all very sparse and counted only once in each publication. Therefore, we will consider author keywords to be representations in our study. There are two reasons for this. First, author-assigned keywords were carefully selected to identify the distinctive research focus of scientific papers (Abrahamson, 1996; McCloskey, 1998), and researchers

Table 1
Basic information of two corpuses.

Corpuses	Number of papers	Number of keywords	Accumulated keyword frequency
Background corpus (LIS in China)	65,653	67,786	277,720
Domain corpus (DL in China)	3560	5610	15,880

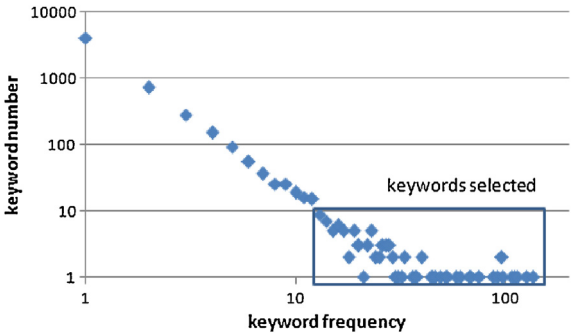

Fig. 2. Keyword selection based on the distribution of keyword frequency.

believe that author keywords can represent a paper's key concepts (Ding, Chowdhury, & Foo, 2001; Yi & Choi, 2012). Second, these keywords are easy to obtain and have been widely used in previous bibliometric studies (Wang, Li, Li, & Li, 2012). Our research can also be applied to two other kinds of keywords because of their similar data feature.

Before using author keywords, we have to manually remove keywords that are mistakenly assigned (Law & Whittaker, 1992), and eliminate meaningless words such as "research," "counter measure," "problem," and so on. Since different authors may use various keywords when describing the same concept, we map keywords with the same meaning into a standard form. After this preprocess, the domain corpus and the background corpus are constructed; their basic information is listed in Table 1.

4. Results analysis

The keyword number is set at 97. This number was chosen based on previous empirical studies on keyword bibliometrics, as well as the keyword frequency distribution in the domain corpus. After investigating prior keyword bibliometric research, we found most had selected no more than 100 keywords for analysis. Combined with keyword frequency distribution in the domain corpus (see Fig. 2), the keyword frequency threshold in the TF method is set at 13, and 97 keywords are selected. Then, an equal number of keywords are selected using the two other methods. The results of the three keyword selection methods are listed in Appendix A. Note that these keywords are translated from Chinese into English according to the Chinese "Great Dictionary of Library and Information Science" (2014).

4.1. The differences between the results of three methods

We can evaluate the similarities between different methods by comparing the overlapping keywords with their results. The number of overlapping keywords among the TF, TF-IDF, and TF-KAI results are calculated according to Appendix A and shown in Fig. 3 as a Venn diagram (Oliveros, 2007).

Fig. 3 indicates that there are 58 common keywords in the three results (59.8% of the total). The TF-KAI result has the most unique keywords (31) and the TF-IDF result has the least (only 2). In the TF-IDF result, there are only 10 different keywords from the TF result, in which 8 keywords are also contained in the TF-KAI result. As shown in Appendix A, the top-ranked keywords in the TF-IDF result are mostly contained in the TF result, while there are more differences in the top-ranked keywords in the TF-KAI result. For example, the TF-IDF result has only one different keyword ("SAN") within the top 79 keywords, while the TF-KAI result has 24 different keywords.

To explain the different results between the three methods, we demonstrate the correlation between keyword frequency with TF-IDF and TF-KAI weights in Fig. 4. From this, we can see that the TF-IDF weights are more relevant to keyword frequency than the TF-KAI weights. The TF-IDF weight has a linear positive correlation with keyword frequency, while there is no apparent correlation between TF-KAI weight and keyword frequency. This may be because of the data feature of publication keywords. As discussed in Section 1, in bibliometric analysis, keywords of each publication are very sparse and counted only once per publication. In the TF-IDF method, $n(\text{all})$ is far greater than $n(i,\text{all})$, according to the data feature. Thus, the discrimination of the IDF factor ($\log(n(\text{all})/n(i,\text{all})) = \log n(\text{all}) - \log n(i,\text{all})$) will be heavily weakened after applying the logarithm. In the TF-KAI method, $n(\text{all})/n(i,\text{all})$ is more distinct than IDF. Therefore, TF-KAI is more independent from keyword frequency than TF-IDF.

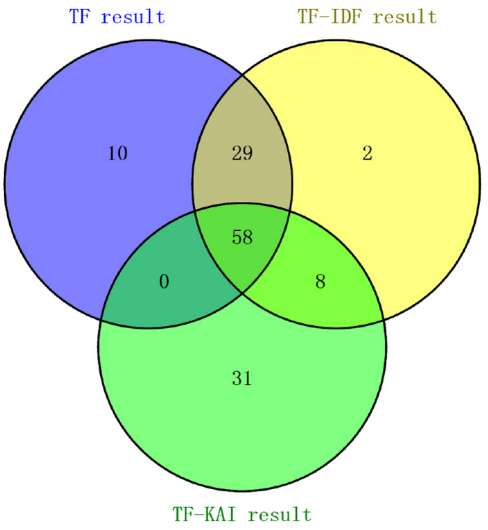

Fig. 3. Keyword overlapping among TF, TF-IDF and TF-KAI results.

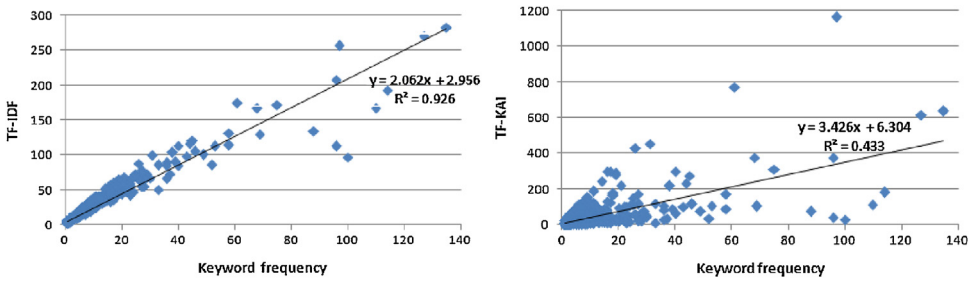

Fig. 4. Correlations between keyword frequency with TF-IDF and TF-KAI.

4.2. Result evaluation

We evaluate the results both qualitatively and quantitatively. In the qualitative evaluation, the results are manually compared to see which can better reveal the knowledge structure of the research domain. To identify it more clearly, we mainly evaluate the TF and TF-KAI methods, because most keywords in the TF-IDF result are included in the TF result, and their non-overlapping keywords are almost all included in the TF-KAI result.

In the quantitative evaluation, a blind selection test is designed, in which several experts are asked to single out a keyword set as a representation of DL research in China. Overlapping keywords between the experts' selections and the above three results are counted, by which we can quantitatively evaluate which method better coincides with the experts' selections.

4.2.1. Qualitative evaluation: Which result can better reveal the knowledge structure of DL in China?

To highlight the differences between the TF result and TF-KAI result, we manually cluster their similar keywords by topic, based on which of the two results can be compared more intuitively (see Table 2). Note that the aim of keyword clustering is to provide a framework for keyword comparison, so we choose manual clustering instead of using automatic clustering algorithms to achieve a better clustering result.

(1) **Data storage.** The TF-KAI method has selected six unique keywords"SAN," "network storage," "NAS," "DAS," "storage device," and "storage system." These keywords represent the subject of digital data storage, which is a core problem in DL. It is also a problem specific to the DL domain compared to other LIS domains. These keywords should not be omitted when analyzing the knowledge structure of DL research in China, despite the fact that the TF method has ignored these keywords because their frequencies are not high enough.

(2) **Basic theory and practical application.** The TF-KAI method has eliminated some general keywords, such as "library," "college library," and "public library." These concepts appear in the DL domain with probabilities far below the baseline regardless of their high frequency, indicating that they have been overlooked more in DL than in the LIS field in China. On the other hand, the TF-KAI method has retained some domain-specific keywords which can better reveal the research feature of DL in China. For example, combining its unique keywords "post digital library" and "ubiquitous library" with the overlapping keywords "traditional library," "hybrid library," "ubiquitous knowledge environment," and "Google," we

Table 2
The unique and overlapping keywords in TF result and TF-KAI result.

Clusters	Unique keywords in TF-KAI result	Unique keywords in TF result	Overlapping keywords
Data storage	SAN, network storage, NAS, DAS, storage device, storage system		
Basic theory and practical application	Post DL, ubiquitous library, the DL Promotion Project in China, National DL of China, regional DL, Tsinghua Tongfang DL TPI System	College library, library, public library, National library of China, college, Library Science, sustainable development	Traditional library, personal DL, mobile DL, college DL, electronic library, hybrid library, ubiquitous knowledge environment, virtual library, Google
Resource building	Characteristic resources, resource discovery, streaming media, virtualization	Internet, network information resources, literature resources, e-book, library consortia, co-construction and sharing, information resources sharing, networked	Network, information resource, digitization, digital object, digital resources, database, characteristic database, digital collections, DL Alliance, DL construction, resource integration, library construction, construction mode, information resources organization, information organization, knowledge organization
Technology in DL	Cloud service platform, semantic grid, semantic interoperation, middleware, SOA, SOAP, CORBA, OAI, union search, Fedora	Search engine, information retrieval, information system, data mining, domain ontology	Information technology, DL system, DL architecture, library automation, cloud computing, interoperation, grid technology, semantic web, ontology, metadata, XML, DC, Open-source software
User and service	DL service, user interface, human interaction, usability evaluation, usability, recommendation system, personalized customization, knowledge service capacity	Service mode, library service, user demand, information demand, service, reader research, reference service, quality of service, librarian	DL user, DL portal, information service, user service, personalized information service, personalization, knowledge service, personalized service
Intellectual property	Digital watermarking		Copyright, intellectual property, fair use, copyright protection, legal permission, The right of communication through information network, copyright law, copyright owner, intellectual property protection
Others	Information security management, access control, knowledge organization systems, Party School	Library management, network security, knowledge management, e-commerce, China, evaluation, standard, index system, electronic reading room, America	Library development, information security, standard specification

can see the Chinese researchers' efforts to explore the development orientation of DL in China. Additionally, its unique keywords "Digital Library Promotion Project in China," "Chinese National Digital Library," "regional digital library," and "Tsinghua Tongfang Digital Library TPI System" demonstrate the important practical applications of DL in China.

(3) **Resource building.** The TF-KAI method has eliminated some relatively redundant keywords, including "electronic resource," "network information resources," "library consortia," and "information resources sharing," to name a few. Similar concepts are already covered by overlapping keywords, so these keywords can be omitted. On the other hand, the TF-KAI method has retained some specific keywords such as "characteristic resources," "streaming media," and "resource discovery."

(4) **Technology in DL.** The TF method retains "search engine," "information retrieval," "information system," "data mining," and "domain ontology," which are high-frequency words with very low domain relevance. These refer to basic technologies which have received more attention in other domains other than DL. The TF-KAI method retains some specific concepts such as "cloud service platform," "semantic grid," "semantic interoperation," "middleware," "SOA," "CORBA," "union search," and so on, which mainly describe the technical model and architecture in DL. These specific concepts can reveal the technical feature of DL in more detail.

(5) **User and service.** The TF-KAI method has eliminated some concepts which are more relevant to traditional library research; for example, "library service," "librarian," "reference service," etc. It retains "DL service," "user interface," "human interaction," "usability evaluation," "usability," "recommendation system," "personalized customization," and so on, from which we can perceive that user interaction and experience are emphasized in DL research in China.

The above analysis shows that, on one hand, the TF-KAI method retains many useful high-frequency keywords; on the other hand, it also highlights some important domain-specific keywords which are conducive to revealing research details and features of a research domain.

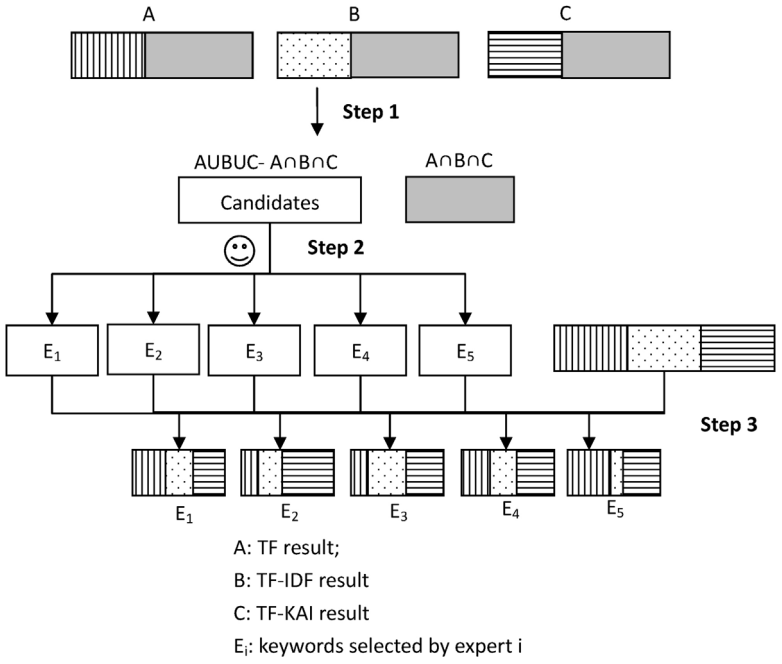

Fig. 5. The procedure of blind-selection test by experts.

Table 3
Results of blind-selection test by experts.

Expert ID	Number of selected keywords	Overlapping Keyword with the TF result		Overlapping Keyword with the TF-IDF result		Overlapping Keyword with the TF-KAI result	
		Number	Percentage (%)	Number	Percentage (%)	Number	Percentage (%)
1	44	14	31.82	16	36.36	30	68.18
2	40	17	42.50	20	50.00	23	57.50
3	36	13	36.11	16	44.44	22	61.11
4	40	13	32.50	15	37.50	26	65.00
5	36	14	38.89	17	47.22	22	61.11
Average	38.8	14.2	36.60	16.8	43.30	24.6	63.40

4.2.2. Quantitative evaluation: Which result fits better with the experts' selection?

To quantitatively evaluate the results, we designed a blind-selection test in which several experts are invited to independently select keywords for mapping the knowledge structure of the DL domain in China. The TF, TF-IDF, and TF-KAI results will be compared with the experts' selections; with the superior result being the one that better corresponds with the experts' selection. The procedure of blind-selection testing can be divided into three steps (see Fig. 5).

Step 1: Construct the candidate set by mixing the non-overlapping keywords[1] in the TF, TF-IDF, and TF-KAI results (see Appendix A). Keywords in the candidate set are still grouped by topic according to Table 2; two unique keywords in the TF-IDF result ("image retrieval" and "MARC") are added in the topic "Technology in DL".

Step 2: Invite several experts to select keywords[2] from the candidate set. The selection criterion is that they believe those keywords can be used as analysis objects for mapping the DL domain in China, together with those common keywords in Table 2.

Step 3: Compare the experts' selection with the TF, TF-IDF, and TF-KAI results. Their overlapping proportions are calculated based on whether we can evaluate the methods quantitatively.

In practice, we invited five experts who have published related papers about bibliometric analysis of DL in China in the past three years, in order to ensure their understanding of the knowledge structure of DL research in China. The statistics of the result comparison is shown in Table 3, from which we can see that the experts tended to select keywords in the TF-KAI result. Therefore, we can conclude that the TF-KAI method performs better than the two other methods.

[1] The overlapping keywords are the 58 keywords occurring in the results produced by all three methods (see Fig. 3), and the rest are non-overlapping keywords.

[2] At first, we suggested each expert select just 36 keywords, but most experts assert that fixing the number will disturb their decision, so we canceled the limitation of keyword number in the blind-selection.

5. Discussion and conclusions

In bibliometric research, if a researcher aims to reveal the details of a domain's major research topics and their relations at the micro-level, it is necessary to select a few keywords as a representation of important research themes in the domain. Previous studies tend to select keywords by frequency or network-based measures, which have been proven to be highly correlated to keyword frequency. In actuality, these methods consider keyword status only within the domain, which may ignore specific keywords which are good representations of the domain specialization. Therefore, we introduced two other methods (TF-IDF and TF-KAI) to select important domain publication keywords. With the support of a background corpus, we can use IDF and KAI to measure a keyword's discrimination to a domain. Domain-specific keywords, which are more concentrated in the given domain than in other domains, can be highlighted.

The empirical study of DL in China shows that the TF-KAI method performed the best. The keywords it selected are most similar with the experts' selection and can better reveal the knowledge structure of the domain. We found that the KAI index can help identify more specific keywords than the IDF index. This can be attributed to the characteristics of bibliometric data. In most bibliometric datasets, keywords are sparse and un-weighted in each publication, which is quite uncharacteristic for full-text data. In this case, the discrimination of IDF is heavily weakened after applying the logarithm, while KAI performed better as a discrimination factor. In a further study, we tried to optimize the TF-IDF method by replacing $n(\text{all})$ as $n(\max)$, which is the maximum keyword frequency in the background corpus, so as to enlarge the discrimination effect of IDF. However, there are still 81 overlapping keywords between the new result of TF-IDF and the TF result, and no additional unique keywords have been selected.

With the support of a background corpus, we can analyze a research domain in the context of its background disciplines and investigate its research themes in a holistic perspective. This same idea has been widely accepted in information retrieval, document clustering and so on (Liu, Wang, Yi, Xu, & Wang, 2005). However, in bibliometrics, domains are usually analyzed independently without considering their relation to background disciplines. In reality, research in a domain will share a common knowledge base with other domains in the same discipline; they also have particular preferences which lead to the specialization of the domain, as well as the diversities among domains. If we investigate a domain in the context of its background disciplines, we will discover that certain general topics are actually underemphasized, even despite popularity, because they are less active in the domain publications when compared to their average distribution in other domains. On the other hand, domain-specific keywords can help people recognize the domain more efficiently, because they can better represent the topical specialization of the given domain. Thus, it is imperative to highlight the topical specialization of a domain when revealing its knowledge structure. Utilizing KAI with the support of a background corpus is a useful method to identify domain-specific keywords in bibliometric data. We believe this can also be applied when using database-provided keywords or title/abstract-extracted keywords.

Our research has some limitations. One major problem is that we have not yet tested other revised weighting schemes based on the TF-KAI method. In Section 4.1, we discuss the effects of the two factors of weighting keywords in bibliometric data (popularity and discrimination). The improved performance of the TF-KAI method is mainly due to its strengthening of the discrimination factor. Furthermore, we can also weaken the factor of keyword popularity; for example, by applying the logarithm or square root of frequency (TF). Our study raises an open-ended question: how can important bibliometric objects be identified in the context of a very different data feature with full-text data? We believe that there is still room for improvement and we are looking for more appropriate methods in the following study.

Acknowledgments

This study was supported by The Major Project of the National Social Science Foundation of China (12&ZD221). The authors would like to give special thanks to Dr. Shuguang Han (University of Pittsburgh) for his valuable comments and editorial help, and to Prof. Changping Hu, Prof. Gaoyong Liu, Dr. Jiming Hu, Dr. Xin Lin, and Dr. Weiwei Yan for their help in the result evaluation. The authors are grateful to anonymous referees and editors for their invaluable and insightful comments.

Appendix A.

Top 97 keywords selected by each method. The unique keywords in each result are underlined. In the TF-IDF and TF-KAI results, non-overlapped keywords with the TF result are in bold.

Rank	TF result	Freq.	TF-IDF result	*w*	TF-KAI result	*w*
1	Copyright	135	Digital resources	302.4	Traditional library	1162.1
2	Metadata	127	Copyright	282.8	DL construction	768.7
3	Information resource	114	Metadata	270.9	Copyright	634.1
4	Information service	110	Traditional library	256.5	Metadata	614.9
5	Digital resources	107	Intellectual property	207.5	Digital resources	456.1
6	Traditional library	97	Information resource	193.0	Personal DL	453.5
7	College library	96	DL construction	174.9	Digital library system	428.8
8	Intellectual property	96	Personalized service	171.6	Intellectual property	373.4

Appendix A (*Continued*)

Rank	TF result	Freq.	TF-IDF result	*w*	TF-KAI result	*w*
9	Library	91	Cloud computing	167.2	Cloud computing	372.9
10	Network	88	Information service	167.2	Personalized service	305.7
11	Personalized service	75	Network	133.7	Mobile DL	295.5
12	Database	69	Digitization	131.1	Interoperation	294.7
13	Cloud computing	68	Database	129.6	DL user	294.3
14	DL construction	61	Fair use	120.8	College DL	288.6
15	Digitization	58	Hybrid library	115.7	DL Alliance	276.8
16	Ontology	53	Interoperation	112.7	Fair use	273.8
17	Knowledge management	52	Ontology	112.5	DL portal	240.4
18	Knowledge service	49	College library	112.2	Hybrid library	231.3
19	Information organization	46	Information organization	105.0	Copyright protection	216.2
20	Fair use	45	Copyright protection	103.6	**The DL Promotion Project in China**	185.5
21	Hybrid library	44	Knowledge service	100.5	Information resource	179.3
22	Interoperation	40	Personal DL	100.0	Legal permission	175.4
23	Information technology	40	Library	96.8	Digitization	171.6
24	Copyright protection	38	Semantic web	88.5	Digital object	165.7
25	Information retrieval	37	Digital library system	87.2	**National DL of China**	149.0
26	Semantic web	36	Knowledge management	85.5	The right of communication through information network	147.4
27	Public library	33	Personalized information service	85.3	**Regional DL**	130.8
28	Personalized information service	33	Information technology	84.8	Ubiquitous knowledge environment	127.1
29	Librarian	32	Copyright law	74.1	Standard specification	124.3
30	Personal DL	31	The right of communication through information network	73.1	**Network storage**	123.6
31	Service mode	30	Resource integration	72.7	Grid technology	121.0
32	Resource integration	29	Information retrieval	72.3	**SAN**	119.7
33	Knowledge organization	29	DL architecture	70.0	Personalized information service	117.4
34	Library service	28	XML	68.6	Information organization	114.2
35	XML	28	**SAN**	68.2	Copyright law	113.4
36	Library construction	28	User service	67.5	**DL service**	112.7
37	Data mining	27	Knowledge organization	67.2	Information service	111.5
38	Search engine	27	Service mode	66.7	**NAS**	111.3
39	Copyright law	27	Library construction	66.6	DL architecture	110.7
40	Digital library system	26	Librarian	66.1	**Cloud service platform**	110.4
41	Personalization	26	College DL	65.7	**Semantic grid**	108.2
42	Characteristic database	26	DL Alliance	65.3	**Semantic interoperation**	107.1
43	The right of communication through information network	25	Virtual library	64.8	User service	104.8
44	DL architecture	25	Personalization	64.7	Semantic web	104.3
45	Library management	24	Characteristic database	64.2	Ontology	104.2
46	User service	24	Intellectual property protection	62.7	Database	101.4
47	User demand	23	Mobile DL	60.6	**Information security management**	100.2
48	Library consortia	23	Data mining	60.1	Virtual library	97.1
49	Library Science	23	Information security	59.2	Intellectual property protection	95.6
50	Virtual library	23	Legal permission	59.1	**Storage device**	94.6
51	Information security	23	Digital object	58.7	**Usability evaluation**	90.1
52	Information demand	22	DL user	57.8	Copyright owner	83.9
53	Intellectual property protection	22	Library development	57.3	**DAS**	82.8
54	Library development	22	Library automation	56.2	Electronic library	81.1
55	Library automation	21	User demand	55.9	**User interface**	78.0
56	Co-construction and sharing	20	Search engine	54.7	Google	76.7
57	Evaluation	20	Library service	54.7	Knowledge service	76.0
58	Network information resources	20	Grid technology	54.0	Resource integration	75.6
59	National library of China	19	Ubiquitous knowledge environment	52.0	**Storage system**	73.7
60	Information resources sharing	19	DL portal	51.0	**Union search**	73.7
61	Service	19	Library consortia	50.7	**Post DL**	73.6
62	College DL	19	Public library	49.8	**Knowledge service capacity**	73.6
63	DL Alliance	19	Electronic library	48.9	Network	71.0
64	Legal permission	18	Co-construction and sharing	48.0	Open-source software	70.2
65	Digital object	18	Open-source software	47.9	**Middleware**	69.4
66	e-Book	17	National library of China	47.7	**Party school**	68.1
67	College	17	Evaluation	47.7	Information Resources Organization	63.5
68	e-Commerce	17	Information demand	47.6	XML	62.0
69	Mobile DL	17	Library management	47.0	**Tsinghua Tongfang DL TPI System**	60.1
70	Grid technology	17	Google	46.3	Information technology	58.8

Appendix A (*Continued*)

Rank	TF result	Freq.	TF-IDF result	*w*	TF-KAI result	*w*
71	Literature resources	16	e-Book	46.1	Library automation	58.7
72	China	16	Information resources sharing	44.8	Personalization	58.6
73	DL user	16	Network information resources	43.5	**Knowledge organization systems**	57.6
74	Ubiquitous knowledge environment	16	Service	43.3	**SOA**	57.5
75	Electronic library	16	Copyright owner	42.2	DC	56.2
76	Open-source software	16	DC	42.2	Characteristic database	56.0
77	Standard	15	Construction mode	42.1	Construction mode	55.6
78	Networked	15	Library science	41.8	Information security	55.2
79	Internet	15	Standard	41.8	Digital collections	54.6
80	reference service	15	**The DL Promotion Project in China**	41.1	**Usability**	54.2
81	Google	15	Standard specification	41.0	Library development	54.2
82	Network security	14	Information resources organization	40.7	**Virtualization**	53.3
83	Domain ontology	14	College	40.1	Library construction	52.7
84	Quality of service	14	Digital collections	39.8	**Ubiquitous library**	52.6
85	Index system	14	**User interface**	39.4	**Recommendation system**	51.9
86	DL portal	14	**Network storage**	39.2	**Streaming media**	51.1
87	DC	14	Domain ontology	39.0	**CORBA**	50.1
88	Construction mode	14	**NAS**	38.7	Knowledge organization	49.2
89	Reader research	13	Literature resources	37.9	**Characteristic resources**	47.3
90	Electronic reading room	13	**Knowledge organization systems**	37.9	**Personalized customization**	47.3
91	Sustainable development	13	Networked	37.8	**Human interaction**	45.0
92	America	13	**Recommendation system**	37.3	**Digital watermarking**	42.9
93	Information system	13	e-Commerce	37.2	**Fedora**	38.4
94	Standard specification	13	**Semantic grid**	35.9	**Resource discovery**	36.9
95	Copyright owner	13	Network security	35.8	**SOAP**	35.4
96	Information resources organization	13	**Image retrieval**	35.4	**Access control**	34.6
97	Digital collections	13	**MARC**	35.4	**OAI**	33.5

References

Abrahamson, E. (1996). Management fashion. *Academy of Management Review*, *21*(1), 254–285.

Assefa, S. G., & Rorissa, A. (2013). A bibliometric mapping of the structure of STEM education using co-word analysis. *Journal of the American Society for Information Science and Technology*, *64*(12), 2513–2536.

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

Brunzel, M., & Spiliopoulou, M. (2007). Domain relevance on term weighting. In Z. Kedad, N. Lammari, E. Métais, F. Meziane, & Y. Rezgui (Eds.), *Natural language processing and information systems* (pp. 427–432). Berlin, Heidelberg: Springer.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics*, *22*(1), 155–205.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, *57*(3), 359–377.

Chen, G., Xiao, L., Hu, C. P., & Zhao, X. Q. (2015). Identifying the research focus of library and information science institutions in china with institution-specific keywords. *Scientometrics*, *103*(2), 707–724.

Choi, J., Yi, S., & Lee, K. C. (2011). Analysis of keyword networks in MIS research and implications for predicting knowledge evolution. *Information & Management*, *48*(8), 371–381.

De Battisti, F., Ferrara, A., & Salini, S. (2015). A decade of research in statistics: A topic model approach. *Scientometrics*, *103*(2), 413–433.

Ding, Y. (2011). Community detection: Topological vs. topical. *Journal of Informetrics*, *5*(4), 498–514.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management*, *37*(6), 817–842.

Ferrara, A., & Salini, S. (2012). Ten challenges in modeling bibliographic data for bibliometric analysis. *Scientometrics*, *93*(3), 765–785.

Frame, J. D. (1977). Mainstream research in Latin America and the Caribbean. *Interciencia*, *2*(3), 143–148.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl 1), 5228–5235.

Harzing, A. W., & Giroud, A. (2014). The competitive advantage of nations: An application to academia. *Journal of Informetrics*, *8*(1), 29–42.

Jabłońska-Sabuka, M., Sitarz, R., & Kraslawski, A. (2014). Forecasting research trends using population dynamics model with Burgers' type interaction. *Journal of Informetrics*, *8*(1), 111–122.

Law, J., & Whittaker, J. (1992). Mapping acidification research: A test of the co-word method. *Scientometrics*, *23*(3), 417–461.

Liu, T., Wang, X. L., Yi, G., Xu, Z. M., & Wang, Q. (2005). *Domain-specific term extraction and its application in text classification*. Proceedings of the 8th joint conference on information sciences (pp. 1481–1484).

Liu, X., Guo, Z., Lin, Z., & Ma, J. (2013). A local social network approach for research management. *Decision Support Systems*, *56*, 427–438.

López-Illescas, C., de Moya-Anegón, F., & Moed, H. F. (2011). A ranking of universities should account for differences in their disciplinary specialization. *Scientometrics*, *88*(2), 563–574.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165.

McCloskey, D. N. (1998). *The rhetoric of economics*. Wisconsin: University of Wisconsin Press.

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

Newman, M. E. (2008). The mathematics of networks. *The New Palgrave Encyclopedia of Economics*, *2*, 1–12.

Niu, B., Hong, S., Yuan, J., Peng, S., Wang, Z., & Zhang, X. (2014). Global trends in sediment-related research in earth science during 1992–2011: A bibliometric analysis. *Scientometrics*, *98*(1), 511–529.
Oliveros, J. C. (2007).Venny. An interactive tool for comparing lists with Venn's diagrams. Available from ⟨http://bioinfogp.cnb.csic.es/tools/venny/index.html⟩.
Piepenbrink, A., & Nurmammadov, E. (2015). Topics in the literature of transition economies and emerging markets. *Scientometrics*, *102*(3), 2107–2130.
Pittel, B., Spencer, J., & Wormald, N. (1996). Sudden emergence of a giant k-core in a random graph. *Journal of Combinatorial Theory Series B*, *67*(1), 111–151.
Pouris, A., & Ho, Y. S. (2014). Research emphasis and collaboration in Africa. *Scientometrics*, *98*(3), 2169–2184.
Quoniam, L., Balme, F., Rostaing, H., Giraud, E., & Dou, J. M. (1998). Bibliometric law used for information retrieval. *Scientometrics*, *41*(1), 83–91.
Rip, A., & Courtial, J. P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, *6*(6), 381–400.
Roche, I., Besagni, D., François, C., Hörlesberger, M., & Schiebel, E. (2010). Identification and characterisation of technological topics in the field of molecular biology. *Scientometrics*, *82*(3), 663–676.
Rokaya, M., Atlam, E., Fuketa, M., Dorji, T. C., & Aoe, J. I. (2008). Ranking of field association terms using co-word analysis. *Information Processing & Management*, *44*(2), 738–755.
Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004, July). *The author-topic model for authors and documents*. Proceedings of the 20th conference on Uncertainty in artificial intelligence (pp. 487–494). AUAI Press.
Salton, G. (1975). *Theory of indexing*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.
Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, *9*(5–6), 281–291.
Su, H. N., & Lee, P. C. (2010). Mapping knowledge structure by keyword co-occurrence: A first look at journal papers in technology foresight. *Scientometrics*, *85*(1), 65–79.
Su, X., Deng, S., & Shen, S. (2014). The design and application value of the Chinese social science citation index. *Scientometrics*, *98*(3), 1567–1582.
Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., & Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation. *Journal of the American Society for Information Science and Technology*, *62*(1), 185–204.
Tang, J., Jin, R., & Zhang, J. (2008, December). *A topic modeling approach and its integration into the random walk framework for academic search*. Proceedings of the eighth IEEE international conference on data minning, ICDM'08 (pp. 1055–1060).
Thijs, B., & Glänzel, W. (2008). A structural analysis of publication profiles for the classification of European research institutes. *Scientometrics*, *74*(2), 223–236.
Wang, Z. Y., Li, G., Li, C. Y., & Li, A. (2012). Research on the semantic-based co-word analysis. *Scientometrics*, *90*(3), 855–875.
Xie, S., Zhang, J., & Ho, Y. S. (2008). Assessment of world aerosol research trends by bibliometric analysis. *Scientometrics*, *77*(1), 113–130.
Yan, E. (2015). Research dynamics, impact, and dissemination: A topic-level analysis. *Journal of the Association for Information Science and Technology*, *66*(11), 2357–2372.
Yi, S., & Choi, J. (2012). The organization of scientific knowledge: The structural characteristics of keyword networks. *Scientometrics*, *90*(3), 1015–1026.
Zhao, R., & Wang, J. (2011). Visualizing the research on pervasive and ubiquitous computing. *Scientometrics*, *86*(3), 593–612.