Scientometrics (2020) 124:1923–1943
https://doi.org/10.1007/s11192-020-03576-5

Keyword-citation-keyword network: a new perspective of discipline knowledge structure analysis

Qikai Cheng[1,2] · Jiamin Wang[1,2] · Wei Lu[1,2] · Yong Huang[1,2] · Yi Bu[3]

Received: 11 September 2019 / Published online: 25 June 2020

Abstract
This paper proposes keyword-citation-keyword (KCK) network to analyze the knowledge structure of a discipline. Different from traditional co-word network analysis, KCK network highlights the importance of keywords assigned in different articles, as well as the semantic relationship between keywords in various articles. In this study, we select computer science domain as an example to illustrate the proposed method. Meanwhile, the results of network analysis, PageRank analysis, and research topic analysis are compared with those of traditional co-word analysis. A total of 110,360 articles with 164,146 unique keywords and 1,615,030 references collected from ACM digital library have been used for this empirical study. The results demonstrate that KCK network outperforms in detecting indirect links between keywords with higher semantic relationship, identifying important knowledge units, as well as discovering the topics with greater significance. Findings from this study contribute to a new perspective and understanding for elucidating discipline knowledge structures, and provide guidance for applying this method in various disciplines.

Keywords Keyword-citation-keyword network · Co-word network · Knowledge structure · Cluster analysis · Network analysis · PageRank

Introduction

A disciplinary knowledge structure refers to a hierarchical knowledge system composed of knowledge elements (units) and their interrelationships contained in a specific discipline. In the current research, discipline knowledge structure analysis is discussed in two primary ways. First, domain experts qualitatively describe knowledge structures according to their own professional knowledge and research experience (Hooper 2009; Sluyter et al. 2006). Second, discipline knowledge structures are quantitatively analyzed based on bibliometric methods to reveal interactive associations between information

✉ Jiamin Wang
wangjm@whu.edu.cn

[1] School of Information Management, Wuhan University, Wuhan 430072, China

[2] Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072, China

[3] Department of Information Management, Peking University, Beijing 100871, China

items in textual data (Hou et al. 2018; Khasseh et al. 2017). As bibliometrics have gradually matured, researchers are increasingly able to detect knowledge interrelations and structure in their disciplines through existing objective relationships between academic objects (das Neves Machado et al. 2016; Wang et al. 2016). Elucidation of discipline knowledge structures is critical for understanding discipline connotations, and promoting innovation and development in the discipline (Fortunato et al. 2018; Wang et al. 2016). In addition, numerous analysis methods can be used to identify the knowledge structure of research fields from various perspectives, such as network analysis (Bu et al. 2018; Cobo et al. 2011), PageRank analysis (Song and Kim 2013; Zhao et al. 2018), and research topic analysis (Khasseh et al. 2017; Ravikumar et al. 2015).

Co-word analysis is effective in knowledge representation and has achieved good performance in depicting intellectual structures (Callon et al. 1983), which is considered as one of the main methods in discipline knowledge structure analysis (Khasseh et al. 2017; Sedighi 2016). Since keywords can directly express the topics and main ideas of a particular literature (Zhang et al. 2015), co-word networks are often constructed with keywords in articles and their relations of co-occurrence (Callon et al. 1991). However, some shortcomings also exist. Firstly, count is considered instead of the importance of keywords in the co-word network and, consequently, many general keywords appear in the co-word network. These general keywords may be useful in depicting an approximate overview of a scientific discipline, but are less successful at identifying detailed themes of a research domain (Chen and Xiao 2016). Secondly, the co-occurrence of keywords cannot fully represent the topic or content correlation of a discipline, since it only focuses on keywords that appear in the same article, without accounting for the relationship between different articles (Wang et al. 2012), e.g., the citation relationship.

Extant studies have proposed that a citation implies a topical relatedness between two articles when one cites the other (Bornmann et al. 2018; Garfield 1964; Zhu et al. 2016). Simultaneously, keywords are regarded as main ideas of the topic and content of articles (Hu and Zhang 2015; Khasseh et al. 2017). It is reasonable to assume that a high degree of semantic similarity exists between the keywords of the citing paper and the keywords of the cited paper (Ding et al. 2013; Song et al. 2013). Therefore, we implement an improvement to the traditional keyword co-occurrence relationship by replacing it with the citation relationship in building a knowledge network and analyzing a discipline knowledge structure. Compared with the co-word network, we can discern that the Keyword-Citation-Keyword (KCK) network possesses the following differences. First, there are more links among the nodes, and especially many new links between two keywords which do not appear in the co-word network. This is because the number of references is commonly more than that of keywords, and the citation acts a bridge between keywords in citing and cited papers. Second, keywords with a higher correlation to a specific domain receive more attention, because the KCK network focuses more on the importance of keywords than on counts.

In this study, we extend the citation relationship among articles to keywords, and construct a KCK network based on the premise that there is a higher topical relatedness between keywords in an article and keywords in its citing article. The aim of the present study is to explore the application and advantages of the proposed method in discipline knowledge structure analysis. For this purpose, we compared the performance of our KCK network to a traditional co-word network under the same corpus in the computer science field. Based on previous studies (Khasseh et al. 2017; Ravikumar et al. 2015; Song and Kim 2013), our evaluation is carried out by network analysis, PageRank analysis, and research topic analysis, which constitute the main aspects of discipline knowledge structure analysis.

The remainder of this paper is organized as follows: first, the work related to our study is given in Sect. 2. Section 3 introduces the methodology of this research. Section 4 presents the experiment and the results. Section 5 describes some discussions and implications of this study. Finally, conclusions and directions for future work are presented in Sect. 6.

Related work

Co-word analysis and its improvement

Co-word analysis was first introduced by Callon et al. (1983) to extend co-citation analysis. It is asserted that co-word analysis can penetrate into the literature, and content analysis can be applied to obtain insight into the structure and development of the discipline. The principle of co-word analysis is that, if two professional terms that can express the subject of a particular research area appear in one article at the same time, a certain internal relationship should exist between them, i.e., the more times that they appear in pairs, the closer the relationship and distance between them. When measuring the intensity of the correlation between the words, the research patterns and conceptual structure of corresponding fields can be examined (Yan et al. 2015). Since co-word analysis was proposed, it has been successfully applied in research of knowledge structures in various domains, including informatics (Khasseh et al. 2017; Sedighi 2016), recommendation systems (Hu and Zhang 2015), the Internet of Things (IoT) (Yan et al. 2015), digital libraries (Liu et al. 2012), etc.

However, two prominent issues have been largely overlooked in the extant research. As previously mentioned, traditional co-word analysis considers count instead of the importance of keywords when building the knowledge network, and as a result many general keywords appear in the co-word network. In addition, traditional co-word analysis only focuses on keywords that appear in the same article, without considering the relationship between different articles (Wang et al. 2012); therefore, keywords with higher topical relatedness cannot be connected if they do not appear in the same article simultaneously. In order to overcome these limitations, many scholars have made efforts to improve the co-word network from varied perspectives, which can be summarized into three aspects.

First, word weights can be used to make the co-word analysis method work more effectively. For example, An and Wu (2011) proposed a co-word analysis method based on subject heading weights. Li and Sun (2013) put forward a definition of a weighted co-occurring keywords time gram, and utilized it as a basic unit to analyze temporal information in an existing keywords collection. Second, it can be improved by considering semantic relations. For example, Wang et al. (2012) proposed a semantic-based co-word analysis which can successfully integrate experts' knowledge into co-word analysis. Feng et al. (2017) combined semantic distance measurements with concept matrices generated from ontologically-based concept mapping to improve the co-word analysis method. Third, numerous researchers have attempted to combine citation relationships with the co-word analysis method. For instance, Braam et al. (1991) combined the co-citation word analysis method, and performed a science mapping analysis on a biochemistry and chemoreception dataset. Ding et al. (2013) used Metformin as an example to form an entity–entity citation network based on literature related to Metformin, and demonstrated that the network can connect disconnected scientific entities and discover new knowledge. Song et al. (2013) constructed a Gene-Citation-Gene (GCG) network of gene pairs implicitly connected through citations, and determined that the GCG network can be useful for detecting gene interactions in an

implicit manner compared with the Gene–Gene (GG) network. Furthermore, Bornmann et al. (2018) introduced a new type of keyword co-occurrence network that uses citation context as a data source for generating keyword co-occurrence networks.

From the existing studies, one can discern that the improvement of the co-word analysis method is primarily focused on term frequency weights, the semantic-based method, and combining with the citation analysis method. However, term frequency weights and the semantic-based method optimize co-word analysis at the level of grammar and semantics, but indirect connections between keywords in different articles are not considered. Although the method combined with citation analysis enriches the relationship between entities, few investigations have been performed to improve the co-word network from the perspective of keyword importance. Consequently, a critical question in the co-word analysis field has become precisely how to distinguish the importance of keywords and add links between them in different papers. We attempt to solve the limitation of keyword importance by considering citation counts, and address the indirect linkage between keywords through the citation network between articles.

Topical relatedness between cited and citing articles

A citation represents that a high degree of semantic similarity exists between the content of the citing paper and the content of the cited paper (Zhu et al. 2015). As an essential part of research papers, a citation is a reference to the source of information used in scientific research. Narin (1976) proposed that a reference is an acknowledgment that one document gives to another, while a citation is an acknowledgment that one document receives from another. In general, a citation implies a relationship between a part or the whole of the cited document and a part or the whole of the citing document (Malin 1968). Garfield (1964) presented 15 reasons why authors cite other texts, including paying homage to pioneers, giving credit for related work, correcting the work of others, disputing priority claims of others, etc. Lipetz (1965) identified 29 categories describing relationships between cited and citing articles, which were grouped into the following four clusters: (1) original scientific contribution of the citing paper; (2) other than original scientific contribution of the citing paper; (3) relationship identification between the citing paper and the cited paper; and (4) scientific contribution of the cited paper to the citing paper. He et al. (2009) stated that citations are important inherent elements in scientific literature, which naturally indicate linkages between topics, and proposed an inheritance topic model that conceptually captures how citations can be used to analyze topic evolution.

More recently, citation relationships have been used in research of text content analysis. For example, Ding et al. (2013) developed the entitymetrics approach, based on the assumption that there exists some topical relatedness between two articles when one cites the other, which uses an entity network to discover new knowledge. Song et al. (2013) then utilized the entitymetrics model to construct a Gene-Citation-Gene (GCG) network, and determined that the GCG network can be useful for detecting gene interactions in an implicit manner compared with the Gene–Gene (GG) network. Bornmann et al. (2018) also suggested that citations reflect the cognitive influence of the cited on the citing publication.

Keywords can effectively represent topics and the main ideas of articles (Hu and Zhang 2015; Khasseh et al. 2017). In bibliometric research, publication keywords are considered the basic elements of representing knowledge concepts, and have been commonly used to identify the knowledge structure of research domains (Su and Lee 2010). Related studies concerning keyword analysis focus on hotspot detection and trend analysis (Chen 2006; Li et al. 2009), research topic analysis (Yan et al. 2015; Khasseh et al. 2017), and knowledge

mapping (Ravikumar et al. 2015; Sun and Zhai 2018). Commonly used methods include keyword frequency analysis and co-word analysis. From these researches, keywords used in representing topics and the main ideas of articles are proven to be both effective and feasible, and bibliometric analysis based on keywords also achieves good performance.

Accordingly, it seems reasonable to assume that there is a higher topical relatedness between a keyword in an article and a keyword in its citing article. Based on this assumption, we extend the citation relationship from the article level to the keyword level, and construct a KCK network of keyword pairs implicitly connected through the citation.

Discipline knowledge structure analysis

Currently, discipline knowledge structures are quantitatively analyzed based on bibliometric methods to reveal interactive associations between information items in textual data (Hou et al. 2018; Khasseh et al. 2017). In a narrow sense, keywords represent discipline knowledge, and the combination of keywords from different scopes and in different numbers constitute the structure of discipline knowledge (Wang et al. 2016). Specifically, network analysis (Bu et al. 2018; Cobo et al. 2011), PageRank analysis (Song and Kim 2013), and research topics analysis (Khasseh et al. 2017; Ravikumar et al. 2015) are important perspectives in depicting discipline knowledge structures.

A knowledge network is a type of complex network, in which each node represents a keyword, and each edge represents a correlation (e.g., co-occurrence or citation) between the nodes. The size and characteristics of the network, and the relationship and structure of knowledge, can be identified through an analysis of the knowledge network. For instance, different measures on the network, such as the total number of nodes and edges, average degree, average clustering coefficient, etc., can be obtained. Recently, Bu et al. (2018) used the metrics of network density and average clustering coefficient to measure the performance of an author co-citation network. Compared with degree centrality, the clustering coefficient can determine node importance from a different perspective, i.e., how solidary is a node's neighborhood (Newman 2010).

The PageRank analysis aims to evaluate the network from the perspective of node importance. Compared with thematic detection, which focuses on the community structure of the network, important keywords analysis can reveal the core knowledge of the discipline and reflect its position in the network. This contributes to understanding discipline knowledge in a more detailed manner. Numerous metrics, such as PageRank (Brin and Page 1998), eigenvector centrality, and degree centrality are adopted for this task. Since PageRank value could constitute a superior measure of importance, as it incorporates the node's visibility and authority simultaneously by taking both the number of links and the prestige of the citing nodes into account, it has been widely employed in measuring the impact of entities (Chen et al. 2018; Song and Kim 2013; Sun et al. 2016; Zhao et al. 2018).

Another important characteristic of complex networks is the community structure, i.e., groups of nodes with high thematic similarity among nodes of the same group and comparatively low thematic similarity among nodes of different groups (Leskovec et al. 2008). The process of discovering these groups is known as community detection. In the current study, community detection is known as research topics identification, and is a task of fundamental importance in knowledge network analysis. Community detection frequently discloses deeper properties of networks and provides meaningful insights about the network's internal structure, as well as its organizational principles (Dakiche et al. 2019). Moreover, discipline connotations can be easily identified from the structure of research topics.

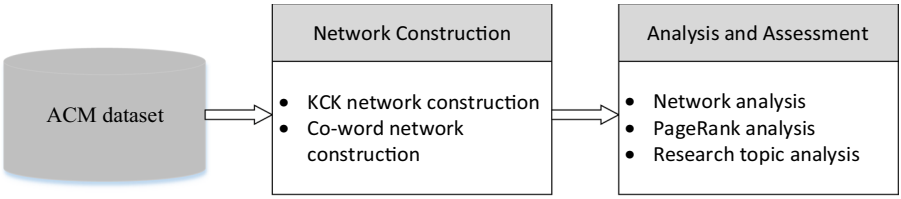

Fig. 1 Methodology of the research

Table 1 Descriptive statistics of the dataset

Statistic	Value
Publication year	1971–2012
Number of articles	110,360
Number of keywords	479,743
Number of unique keywords	164,146
Number of references	1,615,030

Methodology

Figure 1 is the overall research design of the present study. We collected the data from the ACM digital library. Keywords of the papers and citation links were then extracted to construct the KCK network and co-word network. Next, we analyzed and assessed the KCK network from network analysis, PageRank analysis and research topic analysis, in which the co-word network was used for comparison.

Data

The dataset used in this research is 110,360 conference proceedings extracted from the ACM digital library in the period 1971–2012. These conference proceedings are from more than 170 conferences, which capture innovation across the spectrum of computing fields. These papers contain 1,615,030 citation links. Meanwhile, 479,743 keywords are extracted from these articles, with an average of 4.35 for each article. The descriptive statistics of the dataset are presented in Table 1.

KCK network construction

In this paper, we consider keywords as the basic knowledge elements to construct the KCK network. Simultaneously, the traditional co-word network is also built for comparison. Figure 2 shows the construction of the two types of networks. In the co-word network, let us suppose that Paper A, B, and C have two, three, and two keywords, respectively. The keywords in Paper A form a sub-network by the co-occurrence pairs (K1, K2); and the sub-network of Paper B and C can be constituted in the same manner. Since K2 appears in Paper A and B, and K4 appears in Paper B and C, the three isolated sub-networks are linked together, and a co-word network is constructed. In this network,

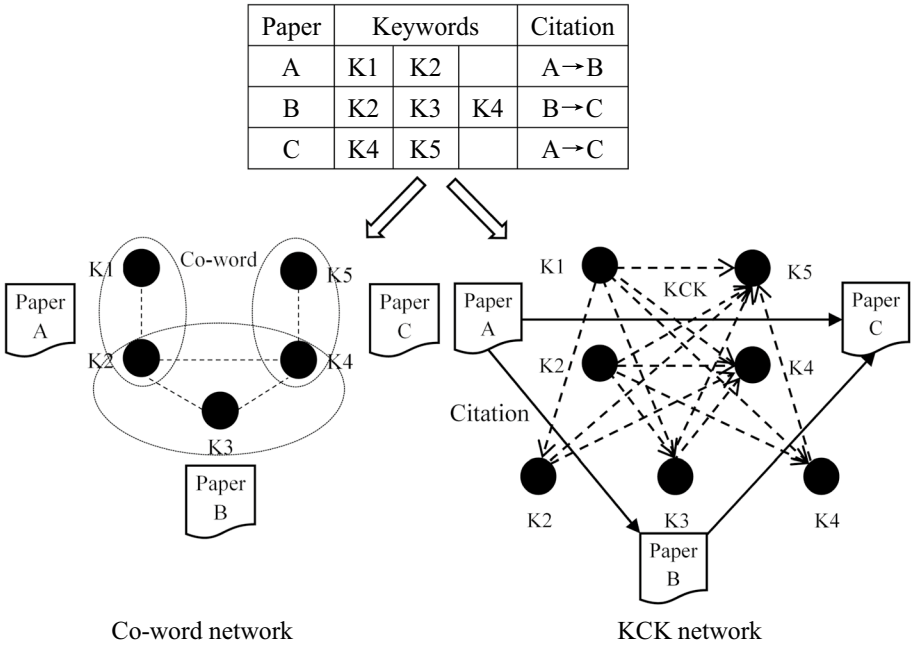

Fig. 2 Construction of two types of networks

since the co-occurrence of keywords has no direction, the generated co-word network is an undirected network.

Regarding the construction of the KCK network, let us assume that Paper A, respectively, cites Paper B and C, and Paper B cites Paper C. The KCK pairs existing in these papers can be represented as (K1, K2), (K1, K3), (K1, K4), …, (K3, K5), and (K4, K5). It is worth noting that citations between the same keywords are excluded. As a result, a KCK network with 14 links between keywords in the three papers is formed. Compared with the co-word network, which only contains five edges, the KCK network provides more information for depicting discipline knowledge structures. Furthermore, not all of the nodes (keywords) in this network have the same importance. Specifically, the higher citation number that a paper has, the more links point to the keywords in this paper; in this way, the importance of keywords can be determined. Since the co-occurrence of keywords has direction, the KCK network constitutes a directed network.

Analysis and assessment methods

Network analysis

Network analysis aims to depict the topology of the constructed network. From the topology, structural features, such as connectivity, sparsity and aggregation, can be discerned. Connectivity concerns how strongly vertices connect with each other, and common metrics of connectivity include metrics related to edges, such as the number of edges whose weights are over five (Zhao et al. 2018). Sparsity focuses on network degree, which is often revealed by average

degree. Aggregation reflects how closely nodes are connected with each other, which is usually measured by average distance and clustering coefficient (Zhao et al. 2018).

Nodes and edges are the basic components of the network, and describe the size of the network. In the case of the same number of nodes, more edges indicate more interactions among nodes, and more relationships are established between keywords in a given network. Average Degree (AD) refers to the average of the degree of all nodes. Network Diameter (ND) is defined as the the longest of all of the calculated shortest paths in a network. The diameter is representative of the linear size of a network. AD and ND are frequently utilized to describe the overall properties of a co-word network (Zhao et al. 2018). The clustering coefficient of a node is the ratio of existing links connecting a node's neighbors to each other to the maximum possible number of such links. The Average Clustering Coefficient (ACC) is the average of the clustering coefficients of all of the nodes. ACC describes the degree of association between neighbors of a node, which also reflects the degree of aggregation of nodes in the network. A network with greater ACC exhibits a better clustering performance in depicting a scientific intellectual structure (Bu et al. 2018). The Average Path Length (APL) refers to the average of the shortest path between all pairs of nodes. Based on network science theories, the smaller is the APL, the larger is the semantic association between nodes in the community, and the better is the performance of clustering. The Modularity algorithm is one of the most common algorithms for measuring the strength of a network community structure (Blondel et al. 2008). In keyword networks, the Modularity algorithm is commonly employed to divide a network community in order to identify research topics (Wang et al. 2014). Generally, in practice, the value of modularity greater than approximately 0.3 appears to indicate a significant community structure (Newman 2004).

Among the commonly used metrics for co-word networks and other complex networks, this study selects seven metrics, including Nodes, Edges, Average Degree (AD), Network Diameter (ND), Average Clustering Coefficient (ACC), Average Path Length (APL), and Modularity (M). These metrics well represent different perspectives in comparing the similarities and differences between the networks generated by the two methods. In addition, we utilize Gephi to obtain these metrics since it can well support the calculation of undirected graphs and directed graphs. The assessment of the two networks will be conducted according to these metrics.

PageRank analysis

The measurement of important keywords aims to identify the core nodes of the network and rank them according to their importance in the network. In this way, the main focus and relative importance of a domain can be determined.

This paper uses the PageRank algorithm to identify important knowledge points in the two networks, in order to compare the differences between the two networks in elucidating the discipline knowledge structure from the perspective of node importance. The PageRank value of keywords in the KCK network is calculated as follows:

$$S(v_i) = (1-d) + d * \sum_{j \in \mathrm{In}(v_i)} \frac{S(v_j)}{\left|\mathrm{Out}(v_j)\right|} \tag{1}$$

where $S(v_i)$ and $S(v_j)$ represent the PageRank values of keywords $\mathrm{v_i}$ and v_j, respectively; $\mathrm{In}(v_i)$ indicates the set of keywords which point to keyword v_i; $\mathrm{Out}(v_j)$ indicates the set of keywords which go out of keyword v_j; $\left|\mathrm{Out}(v_j)\right|$ is the number of elements in a set; and d is

a damping coefficient, which is generally set to 0.85. In the co-word network, the out-degree of a node is equal to the in-degree of the node.

To evaluate the results, we utilize a blind selection experiment for quantitative evaluation. The procedure is as follows: first, we randomize the keywords obtained from the two experiments and obtain 47 unique keywords. Then, we invite the experimenters to select the important ones that can represent the computer science field from keyword sets. In order to ensure the validity of the experiment, the invitees are three experts with many years of research experience in computer science. After the experiments, we count the number and proportion of keywords selected by each experimenter. Since the number of initial keywords provided by the two methods in the candidate word set is equal, it can be concluded that the more keywords selected by experts in one method, the better the performance of the method.

Research topic analysis

Research topic identification aims to sort keywords into groups or clusters, so that the degree of correlation is strong between members of the same cluster and weak between members of different clusters. From the clustering results, the research topics, the size, and the representative keywords of each cluster can be revealed. As mentioned previously, in this paper, we employed the Modularity algorithm (Blondel et al. 2008) to divide each network into clusters. The keywords grouped into the same cluster indicate that they are more likely to have an identical research topic. The subject of each cluster can then be labeled according to the representative keywords contained in each cluster.

The Computing Classification System (CCS) was utilized to evaluate the results of topic identification. The CCS was formulated by the Association for Computing Machinery, which has served as the de facto standard classification system for the computing field. Since the dataset used in this paper was also collected from the ACM digital library, the CSS is appropriate to evaluate the rationality of topic identification results. Besides, we also analyzed the similarities and differences of the main clusters identified in two networks by several cases.

Experiment and results

Network analysis results

According to the principle of network construction presented in Sect. 3.2, all of the co-occurrence pairs and KCK pairs were extracted by a Java program developed by the authors. Next, the file containing nodes and edges was imported into Gephi (an open source network exploration and manipulation tool) (Bastian et al. 2009), and the metrics were also calculated by Gephi.

The statistical results are shown in Table 2. It can be seen that they both contain 163,529 nodes, but the edges of the KCK network are 3,457,161, which is 3.8 times that of the co-word network. This indicates that there are more connections between nodes in the KCK network than those in the co-word network. The higher AD of the KCK network also demonstrates that this network is denser than the co-word network. In terms of ND, there is not much difference between the two methods, indicating a similar linear size of the networks. Concerning ACC and APL, significant differences exist between

Table 2 Statistical data of network metrics in the two methods

Method	Nodes	Edges	AD	ND	ACC	APL	M
Co-word network	163,529	909,774	11.127	11	0.838	42.799	0.563
KCK network	163,529	3,457,161	30.081	10	0.202	7.938	0.537

the two methods. In the co-word network, any two nodes are connected through 42.799 nodes; however, the ACC is quite high. In the KCK network, the average node-to-node distance is only 7.938; however, the ACC is also low. Regarding the modularity (M) value, both the co-word network and the KCK network are at a high level, which indicates that the quality of cluster detection is higher. Moreover, we plotted the keyword degree distribution of each network, as shown in Fig. 3. The fitting results via powerlaw (a Python package) (Alstott et al. 2014) show that the keyword degree distribution of both types of networks follows a power law distribution, although with different alphas and starting points. This indicates that a small number of nodes have more connections in each network, respectively, which are core knowledge of the domain, and they will be the focus of the following analysis.

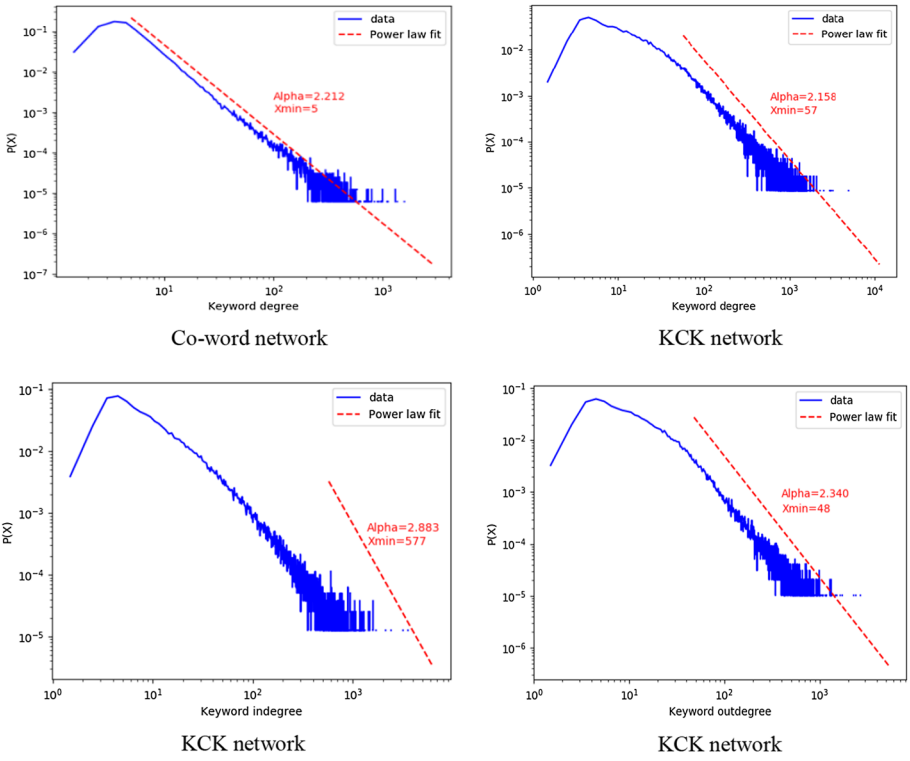

Fig. 3 Keyword degree distribution of two types of networks

PageRank analysis results

The PageRank value of keywords in each network was calculated via Gephi. The top-30 keywords of the PageRank value from each network are listed for illustration and comparison, as shown in Table 3.

From Table 3, it can be found that security, privacy, information retrieval, XML and visualization, in the top-10 keywords based on the co-word network, are all important vocabularies; whereas, design, collaboration, usability, evaluation, and education are general vocabularies with broad semantics which cannot represent the important knowledge points of the computer science domain. In the top-30 keywords set of the co-word network, a large number of unimportant keywords still remain, such as performance, simulation, user experience, etc. Among the top-10 keywords identified by the KCK network, non-photorealistic rendering, ubiquitous computing, sensor networks, augmented reality, CSCW, social networks, privacy, and information retrieval represent important sub-areas or research topics in computer science. In the top-30 keywords set of the KCK network, only a few keywords, such as design, ethnography, evaluation and collaboration, are not important vocabularies in the field.

Several cases are listed for illustrating the difference. For example, non-photorealistic rendering is the highest-ranking keyword in the KCK network. This is an area of computer graphics that focuses on enabling a broad variety of expressive styles for digital art. The second highest-ranking keyword, ubiquitous computing, is related to distributed computing, mobile computing, mobile networking, sensor networks, human–computer interaction, etc. Indeed, they are well-known knowledge points and with greater impact in computer science domain. Similarly, augmented reality, CSCW, information visualization, transactional memory, etc., are all representative words in computer science. However, they do not appear in the top-30 keywords set of the co-word network.

Table 4 presents the results of the blind selection experiment.

From Table 4, one can see that the coincidence proportion of selected keywords with the KCK network is higher than that of the co-word network. The average coincidence ratio of the KCK network reaches 66.98%, while it is only 56.42% for the co-words network, and thus the KCK network can better fit the results of manual selection by experts. This also indicates that the KCK network can not only identify the discipline vocabularies with greater importance, but also solve the problem that a large number of words with broad semantics achieve a high ranking in the traditional co-occurrence methods. This demonstrates that the KCK network is superior to the co-word network in elucidating discipline knowledge points.

Research topic analysis results

High-frequency words are normally identified to map the network because they represent the main topics of text contents. In our study, we filtered the nodes by the node degree, and obtained the most frequent 200 keywords in each model to map the network due to a balance between analysis scale and research aim. Gephi was utilized to detect the clusters of each network and visualize the created networks. The distance between nodes is determined by ForceAtlas2 (Mathieu et al. 2014), a frequently-used layout algorithm for network spatialization. In terms of rendering method, the nodes are assigned different colors based on modularity classes. In this way, nodes in the same color show that their research

Table 3 Keyword ranks based on PageRank

Co-word network				KCK network			
Rank	Keywords	Rank	Keywords	Rank	Keywords	Rank	Keywords
1	Security	16	Data mining	1	Non-photorealistic rendering	16	Routing
2	Design	17	Multimedia	2	Ubiquitous computing	17	Wireless sensor networks
3	Privacy	18	Scheduling	3	Sensor networks	18	Design
4	Collaboration	19	Routing	4	Augmented reality	19	Ethnography
5	Usability	20	Software engineering	5	CSCW	20	Texture mapping
6	Information retrieval	21	Sensor networks	6	Social networks	21	Web characterization
7	Evaluation	22	Performance	7	Privacy	22	Image-based rendering
8	Education	23	Peer-to-peer	8	Information retrieval	23	Animation
9	XML	24	Web services	9	Children	24	Evaluation
10	Visualization	25	Social networks	10	Awareness	25	Collaboration
11	Interaction design	26	Simulation	11	Information visualization	26	Visualization
12	Ubiquitous computing	27	Accessibility	12	Transactional memory	27	Aspect-oriented programming
13	Machine learning	28	User experience	13	Java	28	Texture synthesis
14	Children	29	Virtual reality	14	Collaborative filtering	29	CS1
15	Java	30	Genetic algorithms	15	Web search	30	Security

Table 4 Blind selection results

Experimenter ID	Number of selected keywords	Coincidence with co-word network		Coincidence with KCK network	
		Number	Proportion (%)	Number	Proportion (%)
1	36	22	61.11	23	63.89
2	27	13	48.15	19	70.37
3	30	18	60.00	20	66.67
Average	31	18	56.42	21	66.98

topics are similar, while those in different colors indicate that their research topics are distinct. Figures 4 and 5 present the knowledge structures in the computer science domain by using the co-word network and the KCK network, respectively.

Figure 4 shows the subjects of the top-10 clusters detected in the co-word network, and the representative keywords of each cluster in the co-word network can be found in Table 5. Figure 5 presents 10 subjects in the KCK network, and the representative keywords of each cluster in the KCK network can be found in Table 6. Through consultations with experts in the computer science field, we hold that the subjects of the clusters manually given by the representative keywords are reasonable, and they constitute a good representation of the knowledge structure of computer science.

From the results, several similarities and differences between the two networks can be found. First, the topics of education and collaboration, algorithm optimization, XML and semantic web occurred in the current co-word network but not in the KCK network. This indicates that a large amount of papers related to these topics have been published in this

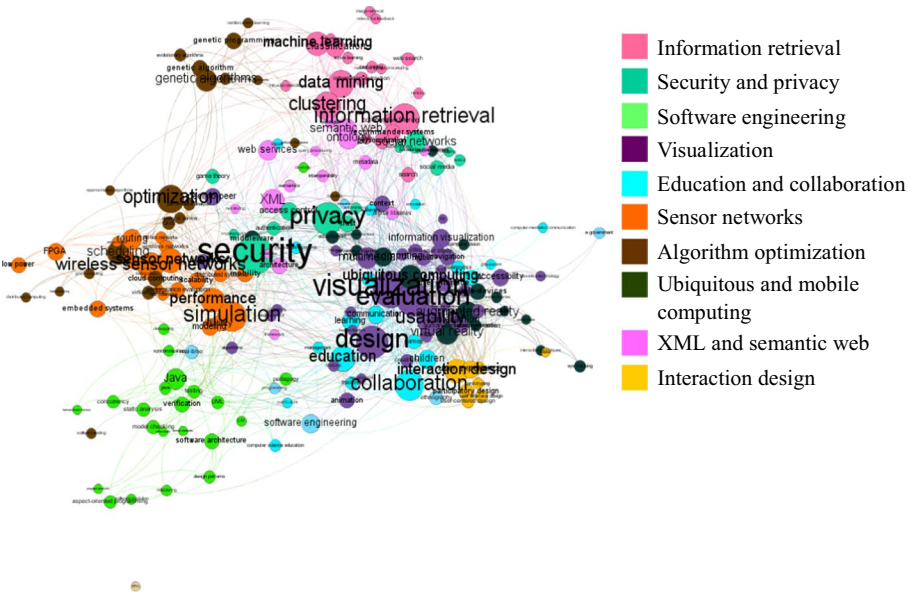

Fig. 4 Co-word network

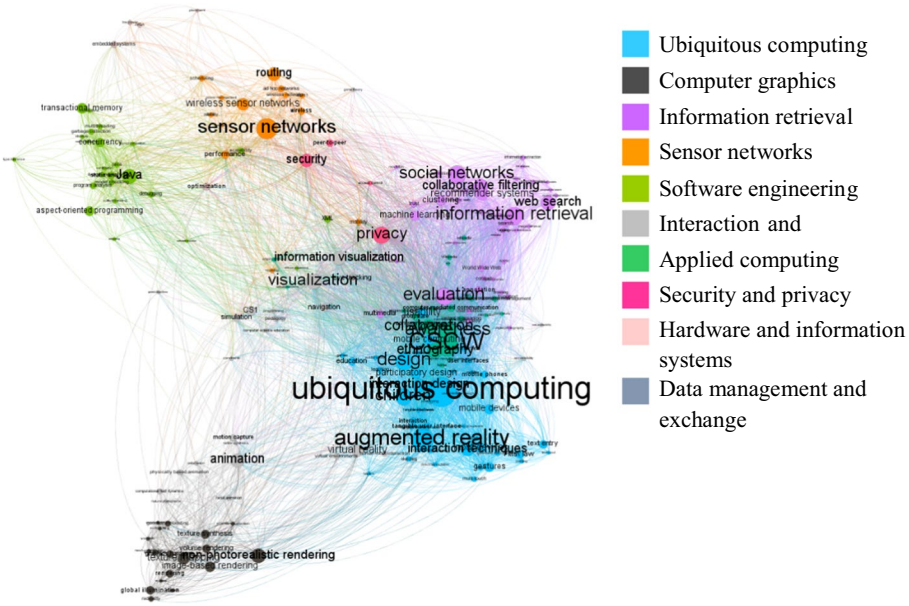

Fig. 5 Keyword-citation-keyword (KCK) network

domain. They tend to be mature and are with close internal correlation, but may have little influence on other topics. For example, compared to the topics like information retrieval and security and privacy, XML and semantic web has less academic influence on the CS discipline. The academic influence of papers with the topic of XML and semantic web mainly comes from papers sharing the same topic.

Second, the distinct topics identified in KCK network includes computer graphics, applied computing, hardware and information systems, and data management and exchange. This illustrates that these topics have more links with other topics, in other words, they have a greater impact on other themes. Since the citation counts play an important role in the KCK network. In this way, even for some topics with relatively fewer articles, they may have greater impact on other topics. Besides, it is noted that most of these topics are applied research topics. In the computer science domain, the progresses of these topics are easier to be applied in various researches. For instance, computer graphics refers to computer-generated image data created with the assistance of specialized graphical hardware and software, which constitutes a vast and developed area of computer science; it is the second-largest cluster of the KCK network.

Furthermore, the two models share six topics, including information retrieval, security and privacy, software engineering, sensor networks, ubiquitous computing, and visualization, respectively. These subjects are the main research interests or sub-areas in computer science (Sun et al. 2016; Uddin et al. 2015), and they can all be found in the Computing Classification System (CCS). They are not only hot topics but also topics with great influence in the domain. Besides, it is noted that these topics' ranking in two networks vary widely. For example, ubiquitous computing is the largest community detected in the KCK network but ranks only eighth in the co-word network. With the rapid development of mobile computing and human–computer interaction, ubiquitous computing has been widely employed in various studies. Consequently, the influence of this topic was

strengthened in the KCK network. This also indicates that there may be much room for further research and application of this topic.

These findings indicate that the KCK network is superior in identifying some important topic clusters, and it assists researchers to find the topics with higher impact on other topics and the overall domain but have not yet become mature.

Discussion

To identify the characteristics and evaluate the performance of the KCK network, we compared it with the co-word network in elucidating the discipline knowledge structure from the aspects of network analysis, PageRank analysis, and research topic analysis.

The constructed KCK network consisted of 163,529 nodes and 909,774 edges, while the co-word network consisted of 3,457,161 edges under the same nodes. This indicates that the KCK network could be effective in building indirect links between keywords contained in different papers and provide more information for the knowledge network. This also confirms the viewpoint proposed by Ding et al. (2013), that entity citations can establish more indirect links between entities and discover new knowledge. The topology structure of the two models was also analyzed by average degree, network diameter, average clustering coefficient, average path length, and modularity. It was found that the network diameter and modularity are similar in both networks, and exhibit a similar network size and higher strength of division of a network into communities. However, significant differences exist in the metrics of average degree, average clustering coefficient, and average path length. The main reason for this is that the edges between any two keywords are limited in the same paper in the co-word network, and thus the probability that the neighbors of a keyword tend to link to each other is higher. This likelihood tends to be greater than the average probability of a connection randomly established between two nodes in the KCK network. Simultaneously, the citation relationships provide more linkages between the keywords in the KCK network, and thus the distance from one node to another will be greatly decreased.

Our experiment that investigates the importance of keywords indicates that the KCK network is more effective in discovering important knowledge units of the computer science domain. In addition, it is found that the important keywords have a higher ranking in the KCK network compared to the co-word network. The main reason for this result is that the KCK network considers the importance of keywords. In this way, keywords with a significant influence on the domain will be distributed on key nodes of the network. This is because the co-word network is a type of network based on word frequency, and keywords with a high frequency are not necessarily of high importance. Although the KCK network offers obvious advantages compared to the co-word network, much room remains for improvement. For example, considering that the full text of articles (Lu et al. 2018) can provide more semantic entities and connections among them, the KCK network based on full-text data may improve the accuracy of the model.

The topic analysis results revealed the existence of similarities, as well as differences, between the KCK network and the co-word network. Since the establishment of the two networks is based on different relationship, the identified topics depict the knowledge structure from different perspectives. For the co-word network, the topic intensity is depending more on the amount of papers and the impact of one keyword to others mainly occurs within the same topic. However, the KCK network emphasizes the citation relationship, thus some potentially important topics in the domain can be revealed. This provides an opportunity for

the researchers to combine these topics with other studies. Furthermore, combing the co-word network and KCK network can provide a more complete picture for a specific domain.

In addition to the method itself and its advantages compared with the traditional co-word network, this study provides several implications for future researchers. First, the KCK network provides a novel perspective for discipline knowledge structure analysis, which transcends conventional methods to map the knowledge domain based on the co-occurrence of keywords. The findings in this paper inform us to focus on the importance of keywords and indirect relationships between them. This constitutes a significant foundation for future improvement of the traditional co-word method. Second, one of the advantages of considering citation networks in keyword co-occurrence analysis is that a bridge between research topics and publication year of papers can be constructed. This can inspire future researchers to duplicate this method in other scholarly text analyses, such as research fronts analysis (Huang and Chang 2014; Morris, et al. 2003), evolution of research topics (Chandra 2018), and research path analysis.

Conclusion

This paper introduces a citation network to extend traditional co-word analysis, and a novel perspective, called the Keyword-Citation-Keyword (KCK) network, is proposed. In particular, we have examined the computer science domain, and compared the similarities and differences between the two networks in elucidating the discipline knowledge structure. The results showed that the KCK network possesses unique advantages for discipline knowledge structure analysis in some aspects. First, the KCK network are effective in building indirect links between keywords contained in different papers and provide more information for the knowledge network, which have been demonstrated to be of importance in discipline knowledge structure analysis. Second, the KCK network is more effective in discovering important knowledge units of the computer science domain. Besides, the KCK network has advantages in discovering the topics with great impact on other topics but hardly identified in co-word network.

This research possesses several limitations worth noting. First, our sample is only valid for articles published from 1971 to 2012. Although our sample of articles covers the majority of conference papers of computer science in this period, it fails to include papers published in recent years. As a result, the topics of clusters are unable to reflect some of the latest topics. This limitation, however, will be addressed in future work. Second, we utilized the keywords of the papers to build the networks. Although the keywords directly express the main ideas of a particular literature, they are still not adequate for a full paper. Consequently, we will attempt to extract keywords and semantic entities from titles, abstracts, and full text to build the networks in order to provide more information.

Acknowledgements This work was partially supported by Major Projects of National Social Science Foundation of China (No. 17ZDA292) and National Natural Science Foundation of China (No.71704137).

Appendix

See Tables 5 and 6.

Table 5 Top 10 clusters and representative keywords based on the co-word network

Cluster	Cluster theme	Representative keywords
1	Information retrieval	Information retrieval; machine learning; data mining; classification; recommender systems; personalization; web search; active learning; ranking; search
2	Security and privacy	Security; privacy; social networks; social media; trust; access control; collaborative filtering; game theory; authentication; web 2.0
3	Software engineering	Java; verification; software architecture; concurrency; testing; model checking; aspect-oriented programming; static analysis; pedagogy; debugging
4	Visualization	Visualization; design; evaluation; usability; information visualization; accessibility; multimedia; navigation; animation; video
5	Education and collaboration	Collaboration; education; children; learning; communication; training; awareness; management; coordination; computer-mediated communication
6	Sensor networks	Simulation; performance; wireless sensor networks; sensor networks; routing; FPGA; low power; modeling; embedded systems; scalability
7	Algorithm optimization	Optimization; scheduling; genetic algorithms; genetic algorithm; genetic programming; cloud computing; evolutionary algorithms; reinforcement learning; grid computing; approximation algorithms
8	Ubiquitous and mobile computing	Ubiquitous computing; augmented reality; virtual reality; mobile computing; mobile devices; middleware; adaptation; digital government; pervasive computing; integration
9	XML and semantic web	XML; semantic web; ontology; web services; architecture; metadata; digital libraries; functional programming; interoperability; semantics
10	Interaction design	Interaction design; participatory design; user-centered design; user interface design; ethnography; prototyping; design automation; human–computer interaction; user research; product design

Table 6 Top 10 clusters and representative keywords based on the KCK network

Cluster	Cluster theme	Representative keywords
1	Ubiquitous computing	Ubiquitous computing; augmented reality; design; interaction design; children; interaction techniques; participatory design; mobile device; gestures; usability
2	Computer graphics	Non-photorealistic rendering; texture mapping; image-based rendering; texture synthesis; volume rendering; global illumination; rendering; graphics hardware; geometric modeling; radiosity
3	Information retrieval	Information retrieval; social networks; evaluation; collaborative filtering; web search; recommender systems; machine learning; clustering; annotation; multimedia
4	Sensor networks	Sensor networks; routing; wireless sensor networks; performance; wireless; scheduling; wireless networks; ad hoc networks; mapping; mobility
5	Software engineering	Java; transactional memory; concurrency; aspect-oriented programming; static analysis; XML; debugging; garbage collection; program analysis; scalability
6	Interaction and visualization	Visualization; information visualization; animation; virtual reality; CS1; simulation; navigation; motion capture; physically based virtual environments; eye tracking
7	Applied computing	CSCW; awareness; collaboration; ethnography; mobile computing; computer-mediated communication; communication; groupware; user interface; user studies
8	Security and privacy	Privacy; security; peer-to-peer; access control; phishing; anonymity; authentication; role engineering; intrusion detection; usable security
9	Hardware and information systems	Optimization; embedded systems; low power; FPGA; placement; reliability; power management; energy efficiency; distributed algorithms; compilers
10	Data management and exchange	Web services; data exchange; data integration; SVG; interoperability; schema mapping; version control; XSLT; subtyping; adaptive layout

References

Alstott, J., Bullmore, D. P., & Plenz, D. (2014). powerlaw: A Python package for analysis of heavy-tailed distributions. *PLoS ONE, 9*(1), e85777.

An, X. Y., & Wu, Q. Q. (2011). Co-word analysis of the trends in stem cells field based on subject heading weighting. *Scientometrics, 88*(1), 133–144.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the 3rd international AAAI conference on weblogs and social media, May 17–19, 2009, San Jose, CA, USA* (pp. 361–362).

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.

Bornmann, L., Haunschild, R., & Hug, S. E. (2018). Visualizing the context of citations referencing papers published by Eugene Garfield: A new type of keyword co-occurrence analysis. *Scientometrics, 114*(2), 427–437.

Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis I. Structural aspects. *Journal of the American Society for Information Science and Technology, 42*(4), 233–251.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th international conference on world wide web, April 14–18, 1998, Brisbane, Australia*.

Bu, Y., Wang, B., Huang, W.-B., Che, S., & Huang, Y. (2018). Using the appearance of citations in full text on author co-citation analysis. *Scientometrics, 116*(1), 275–289.

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 chemistry. *Scientometrics, 22*(1), 155–205.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information, 22*(2), 191–235.

Chandra, Y. (2018). Mapping the evolution of entrepreneurship as a field of research (1990–2013): A scientometric analysis. *PLoS ONE, 13*(1), e0190228.

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, C., Song, M., & Heo, G. E. (2018). A scalable and adaptive method for finding semantically equivalent cue words of uncertainty. *Journal of Informetrics, 12*(1), 158–180.

Chen, G., & Xiao, L. (2016). Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods. *Journal of Informetrics, 10*(1), 212–223.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology, 62*(7), 1382–1402.

Dakiche, N., Benbouzid-Si Tayeb, F., Slimani, Y., & Benatchba, K. (2019). Tracking community evolution in social networks: A survey. *Information Processing and Management, 56*(3), 1084–1102.

das Neves Machado, R., Vargas-Quesada, B., & Leta, J. (2016). Intellectual structure in stem cell research: Exploring Brazilian scientific articles from 2001 to 2010. *Scientometrics, 106*(2), 525–537.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., et al. (2013). Entitymetrics: Measuring the impact of entities. *PLoS ONE, 8*(8), e71416.

Feng, J., Zhang, Y. Q., & Zhang, H. (2017). Improving the co-word analysis method based on semantic distance. *Scientometrics, 111*(4), 1–11.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., et al. (2018). Science of science. *Science, 359*(6379), eaao0185.

Garfield, E. (1964). Can citation indexing be automated? In M. E. Stevens, V. E. Giuliano, & L. B. Heilprin (Eds.), *Statistical association methods for mechanized documentation: Symposium proceedings Washington 1964* (pp. 189–192). Department of Commerce National Bureau of Standards.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009). Detecting topic evolution in scientific literature: How can citations help? In *Proceedings of the 18th ACM conference on information and knowledge management, November 2-6, 2009, Hong Kong, China* (pp. 957–966).

Hooper, R. P. (2009). Towards an intellectual structure for hydrologic science. *Hydrological Processes, 23*(2), 353–355.

Hou, J., Yang, X., & Chen, C. (2018). Emerging trends and new developments in information science: A document co-citation analysis (2009–2016). *Scientometrics, 115*(2), 869–892.

Hu, J., & Zhang, Y. (2015). Research patterns and trends of recommendation system in china using co-word analysis. *Information Processing and Management, 51*(4), 329–339.

Huang, M. H., & Chang, C. P. (2014). Detecting research fronts in OLED field using bibliographic coupling with sliding window. *Scientometrics, 98*(3), 1721–1744.
Khasseh, A. A., Soheili, F., Moghaddam, H. S., & Chelak, A. M. (2017). Intellectual structure of knowledge in iMetrics: A co-word analysis. *Information Processing and Management, 53*(3), 705–720.
Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2008). Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th international conference on the world wide web, April 21-25, 2008, Beijing, China* (pp. 695–704).
Li, L. L., Ding, G., Feng, N., Wang, M. H., & Ho, Y. S. (2009). Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006. *Scientometrics, 80*(1), 39–58.
Li, S., & Sun, Y. (2013). The application of weighted co-occurring keywords time gram in academic research temporal sequence discovery. In *Proceeding of the 76th ASIS&T annual meeting: Beyond the cloud: Rethinking information boundaries, November 1–5, 2013, Montreal, Quebec, Canada*.
Lipetz, B. A. (1965). Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *Journal of the Association for Information Science and Technology, 16*(2), 81–90.
Liu, G. Y., Hu, J. M., & Wang, H. L. (2012). A co-word analysis of digital library field in china. *Scientometrics, 91*(1), 203–217.
Lu, W., Huang, Y., Bu, Y., & Cheng, Q. (2018). Functional structure identification of scientific documents in computer science. *Scientometrics, 115*(1), 463–486.
Malin, M. V. (1968). The science citation index^{R}: A new concept in indexing. *Library Trends, 16,* 374–387.
Mathieu, J., Tommaso, V., Sebastien, H., & Mathieu, B. (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE, 9*(6), e98679.
Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Timeline visualization of research fronts. *Journal of the Association for Information Science and Technology, 54*(5), 413–422.
Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity* (p. 334, 337). Cherry Hill: Computer Horizons Inc.
Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, 69*(6 Pt 2), 066133.
Newman, M. (2010). *Networks: An introduction*. Oxford: Oxford University Press.
Ravikumar, S., Agrahari, A., & Singh, S. N. (2015). Mapping the intellectual structure of scientometrics: A co-word analysis of the journal scientometrics (2005–2010). *Scientometrics, 102*(1), 929–955.
Sedighi, M. (2016). Application of word co-occurrence analysis method in mapping of the scientific fields (case study: The field of informetrics). *Library Review, 65*(1/2), 52–64.
Sluyter, A., Augustine, A. D., Bitton, M. C., Sullivan, T. J., & Wang, F. (2006). The recent intellectual structure of geography. *Geographical Review, 96*(4), 594–608.
Song, M., Han, N. G., Kim, Y. H., Ding, Y., & Chambers, T. (2013). Discovering implicit entity relation with the gene-citation-gene network. *PLoS ONE, 8*(12), e84639.
Song, M., & Kim, S. Y. (2013). Detecting the knowledge structure of bioinformatics by mining full-text collections. *Scientometrics, 96*(1), 183–201.
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.
Sun, X., Ding, K., & Lin, Y. (2016). Mapping the evolution of scientific fields based on cross-field authors. *Journal of Informetrics, 10*(3), 750–761.
Sun, Y. W., & Zhai, Y. (2018). Mapping the knowledge domain and the theme evolution of appropriability research between 1986 and 2016: A scientometric review. *Scientometrics, 116*(1), 203–230.
Uddin, A., Singh, V. K., Pinto, D., & Olmos, I. (2015). Scientometric mapping of computer science research in Mexico. *Scientometrics, 105*(1), 97–114.
Wang, X., Cheng, Q., & Lu, W. (2014). Analyzing evolution of research topics with NEViewer: A new method based on dynamic co-word networks. *Scientometrics, 101*(2), 1253–1271.
Wang, H., Deng, S., & Su, X. (2016). A study on construction and analysis of discipline knowledge structure of Chinese LIS based on CSSCI. *Scientometrics, 109*(3), 1725–1759.
Wang, Z. Y., Li, G., Li, C. Y., & Li, A. (2012). Research on the semantic-based co-word analysis. *Scientometrics, 90*(3), 855–875.
Yan, B. N., Lee, T. S., & Lee, T. P. (2015). Mapping the intellectual structure of the Internet of Things (IoT) field (2000–2014): A co-word analysis. *Scientometrics, 105*(2), 1285–1300.
Zhang, W., Zhang, Q., Yu, B., & Zhao, L. (2015). Knowledge map of creativity research based on keywords network and co-word analysis, 1992–2011. *Quality and Quantity, 49*(3), 1023–1038.
Zhao, W., Mao, J., & Lu, K. (2018). Ranking themes on co-word networks: Exploring the relationships among different metrics. *Information Processing and Management, 54*(2), 203–218.

Zhu, Y., Song, M., & Yan, E. (2016). Identifying liver cancer and its relations with diseases, drugs, and genes: A literature-based approach. *PLoS ONE, 11*(5), e0156091.
Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology, 66*(2), 408–427.