Scientometrics
DOI 10.1007/s11192-017-2290-5

Multi-source data fusion study in scientometrics

Hai-Yun Xu[1] · Zeng-Hui Yue[2] · Chao Wang[1] · Kun Dong[1] · Hong-Shen Pang[3] · Zhengbiao Han[4]

Received: 1 April 2016

Abstract This paper provides an introduction to multi-source data fusion (MSDF) and comprehensively overviews the ingredients and challenges of MSDF in scientometrics. As compared to the MSDF methods in the sensor and other fields, and considering the features of scientometrics, in this paper an application model and procedure of MSDF in scientometrics are proposed. The model and procedure can be divided into three parts: data type integration, fusion of data relations, and ensemble clustering. Furthermore, the fusion of data relations can be divided into cross-integration of multi-mode data and matrix fusion of multi-relational data. To obtain a clearer and deeper analysis of the MSDF model, this paper further focuses on the application of MSDF in topic identification based on text analysis of scientific literatures. This paper also discusses the application of MSDF for the exploration of scientific literatures. Finally, the most suitable MSDF methods for different situations are discussed.

Keywords Data fusion · Relations fusion · Multi-mode analysis · Multi-source data · Scientometrics

✉ Hai-Yun Xu
xuhy@clas.ac.cn

[1] Chengdu Library of Chinese Academy of Sciences, No. 16, South Section 2, Yihuan Road, Chengdu 610041, Sichuan, People's Republic of China

[2] School of Medical Information Engineering, Jining Medical University, Rizhao 276826, People's Republic of China

[3] Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, People's Republic of China

[4] Nanjing Agricultural University, Nanjing 210095, Jiangsu, People's Republic of China

Published online: 15 February 2017

Introduction

Scientific literatures are important research objects in Library and Information Science, in particular in scientometrics, where the association among different entities of scientific and technical literatures, represented in scientific papers and patents, is an important research topic. Within this field, typical association analysis includes entities based on citations, co-authors, and topic terms. According to diverse units, citations and topic terms can be used not only for describing the relationship between literatures but also for describing or analyzing the relationship between authors, journals, and institutions. With the growth of scientific and technical literatures, the types of analyzable relationship have significantly increased and have become valuable data resources for scientometrics analysis. Meanwhile, in addition to those based on citation and co-author relationships, various analysis methods based on coupling relationship have provided new and insightful views.

However, scientometrics analysis currently cannot utilize aggregate analyses effectively, because most are still based on a single-type relationship. Only a few studies have considered two types of relationship, for example, citation relationship and co-author relationship, or three or more types of relationship. There exist some deficiencies in the analytical methods with a single type of relational data used for the literatures. A single relationship usually provides the researchers with partial and unbalanced characteristics of one research field. Scientometric methods based on a single relation can only reflect a limited understanding of one's domain of science from a certain perspective, which easily causes analysis error. Therefore, it is necessary and helpful for researchers to fully understand a research field from different perspectives.

Multi-source data fusion (MSDF) refers to comprehensively fusing different types of information sources or relational data in the analysis procedure by using specific methods in order to reveal the characteristics of the research object and obtain more comprehensive and objective measurement results. The study of MSDF techniques, which has been concentrated mostly in the sensor field, has also become a key topic in bioinformatics, artificial intelligence, face recognition, and other disciplines. In recent years, with the development of data science and complex networks, the clustering of studies that involve the fusion of different networks has also received increasing attention (Monti et al. 2003; Wang et al. 2014).

With the increase in the amount and types of scientific literature, MSDF has become more necessary and meaningful. The aim of this study is to fully explore the status quo of the method and applications of MSDF in scientometrics.

Overview of MSDF in scientometrics

Hot research topic: scientometrics

Van Raan (1997) categorized the core research activities of scientometrics into four interrelated areas: (1) science and technology indicators, (2) science and technology information systems, (3) the interaction between science and technology, and (4) cognitive as well as socio-organizational structures in science and technology. The journal *Scientometrics* addresses the quantitative features and characteristics of science and scientific research. Emphasis is placed on investigations in which the developments and mechanisms of science are studied using statistical mathematical methods (*Scientometrics* editorial

board). *Scientometrics* covers all aspects of scientometrics and has published 46.31% of the scientometrics research papers in the world (Mooghali et al. 2011). Ravikumar et al. (2015) explored the intellectual structure of scientometrics from 2005 to 2010 using text mining and co-word analysis. Additionally, they found that the themes from the text in *Scientometrics* either have well defined genealogies, such as citation analysis, author productivity, and bibliometric analysis, or appear to emanate from multiple preceding themes, such as h-indicator, co-citation map, and co-citation link (Ravikumar et al. 2015).

On the basis of the above analysis, we can infer that the development of scientometrics depends on a considerable amount of valuable information together with effective quantitative analysis methods to handle it. Currently, in the big data era, multi-source heterogeneous data will become increasingly common; thus, finding better means of managing these data is an important task in the development of scientometrics.

Basic types of relation

We propose that MSDF can be divided into the fusion of data types and fusion of data relationships. The fusion of data types involves merging different data types into the same analysis object. Currently, data types include mainly journal articles, conference information, dissertations, patent information, books, information of scientific research project, etc. Hua (2013) divided multi-source data into homogeneous information with heterologous sources, heterogeneous information, and multilingual information, which mainly used the techniques of field mapping, field splitting, filtering repeated data, and weighting heterogeneous data. All these types of technique are inevitable steps of data processing in future scientometrics analysis. This article focuses on the fusion of data relations and thus contains few descriptions of the data type fusion method.

The aim of data relations fusion is to merge different data relations into a new one, which is used to characterize the relationship among various entities. The most widely used MSDF methods in scientometrics is the linear mode which is simple and has a random fusing process (Su and Zhang 2010; Su 2011). Table 1 summarizes the advantage and

Table 1 Types of relational data fusion in scientometrics

Type of data fusion	Subtype of data fusion	Advantages	Disadvantages
Linear mode	–	Logical simplicity	–
Cross-integration of multi-mode data	Multi-mode data visualization	Takes advantage of cross co-occurrence technique	Ignores the same types of entities
	Mapping method of multi-mode network	Logical simplicity and easy processing	Leads to information loss, increases the number of edges of the entire network, and increases noise information
	Non-mapping method of multi-mode network	More accurate and reliable	Algorithm is complicated
Matrix fusion of multi-relational data	–	Presents all relations; logical simplicity	Algorithm is complicated

disadvantages of the three main types of data fusion in scientometrics. Next, we give a more detailed overview of the data fusion types in scientometrics.

Shibata et al. (2009) conducted a comparative study of three networks based on citation relationships: co-citation, bibliographic coupling, and citation networks. The results showed that citation network can find new topics faster and are the most effective means of identifying research fronts, whereas co-citation network are the least effective. In addition, content clustering based on the citation network both yields results with the highest similarity and shows the least risk of omitting emerging research fronts. Klavans and Boyack (2006) discovered that a clustering network built by direct citations has more content similarities than a co-citation network. While Couto et al. (2006) indicated that the bibliographic coupling approach is more effective than the textual approach in the empirical analysis of computer-related literature, Ahlgren and Colliander (2009) proved the textual approach to be better in the empirical analysis of the literature on information retrieval.

Since the above analysis methods have their respective fields of application, it is difficult to determine which method is the best, which reflects the fact that each method has its own advantage and disadvantages. Morris and Van der Veer Martens (2008) noted that the scientometrics methods based on a single relation can provide only a limited understanding of the analyzed objects from one perspective. In other words, each relationship can help researchers analyze only partial characteristics of a research field from a certain point of view. It is only when observing a research field from different perspectives that we can fully understand it. The issue we need to further examine is how to integrate multiple relationships to gain a more comprehensive understanding of the objects we are studying.

Fusion of data relations

According to the different techniques used to achieve fusion, the fusion of multi-source data relations can be divided into the cross-integration of multi-mode data and the matrix fusion of multi-relational data. Cross-integration of multi-mode data shows associations among different types of entities. It takes advantage of the cross co-occurrence technique, but ignores the same types of entities. Matrix fusion of multi-relational data merges similar matrixes or distance matrixes of multi-source data into a new matrix that seeks to present all the relations, and then multivariate statistical analysis, such as cluster analysis, factor analysis, etc., is performed (Fig. 1).

Both the cross-integration of multi-mode data and matrix fusion of multi-relational data finally form a comprehensive matrix. For the cross-integration of multi-mode data, the dimension of the newly formed matrix is the sum of the dimensions of all the original

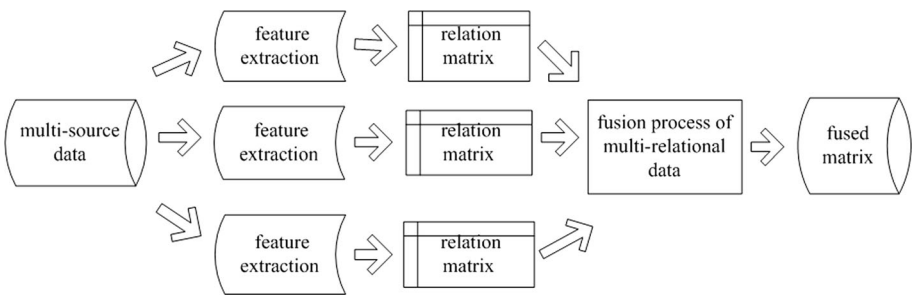

Fig. 1 Matrix fusion of multi-relational data

matrixes; however, for the matrix fusion of multi-relational data, the dimension of the new matrix does not change.

Cross-integration of multi-mode data

In social network analysis, a mode refers to a collection of actors, specifically, the number of types of actors in the collection. The network of relationships between two collections is called a 2-mode crossing network (2-mode network). Similarly, the network of relationships among three collections is called a 3-mode crossing network (3-mode network), and a multi-mode network corresponds to relationships among three or more collections.

Cross-integration of multi-mode data can be defined as the process of combining multi-source data to form a multi-mode matrix, where connections exist only among different types of node, and vice versa. For example, data for a 2-mode network can be used to generate a bipartite graph, the vertices of which can be divided into two disjoint sets. This can be expressed using mathematical language as follows. Let $G = <V, E>$ be a graph, and V be the set of vertices such that $V1 \cup V2 = V$, $V1 \cap V2 = \Phi$. G is called a bipartite graph; if two end points of any edge in G are such that $(x, y) \in$ E, we can obtain $x \in V1$, $y \in V2$ or $x \in V2$, $y \in V1$.

Some researchers have combined bibliography and topic terms to discover research topics, and two approaches have emerged from this combination. One consists of using the cited references as another reinforced restriction of the relationship between the topic terms, and the second is using citations to build the relationship between bibliography and topic terms. When indexed terms are used to study the domain structure, problems such as polysemy may be encountered. Under these circumstances, the cited references provide a specific context for indexing terms, which, to a certain extent, can reduce lexical ambiguity. Taking this into consideration, Van Den Besselaar and Heimeriks (2006) used the co-occurrence of topic terms with cited references to describe the research objects and further applied cluster analysis to literature collections to identify the research front.

In an application study on the cross-chart and timeline method, Morris et al. (2002) and Morris and Yen (2004) used the association of two co-occurrence matrices with identical feature items, and solved the problem by using visualization to reveal an association between two feature items. To integrate textual information into the modeling process on an equal footing with statistical data, Antal et al. (2002) investigated whether similarities or dependencies between variables quantized from textual information agree with those identified by expert assessment. Antal et al. (2004) presented an application method for the extraction of relationships among domain variables from text to support Bayesian network structure learning.

Leydesdorff (2010) applied the theory of heterogeneous networks to a 3-mode network, where he linked the feature items of authors, journals, and keywords and showed these different types of node in one network, which could be used to analyze the relationships between nodes and provide a more realistic reflection of research networks. Pang (2012) improved the presentation mode of Morris' cross-chart. The improved cross-chart not only revealed the association between two feature items, but also could present the co-occurrence among three feature items. In a patent-based technology mining analysis, Xu and Fang (2014) used association rules between technological and functional dimensions in a patent technology-effect matrix to acquire the correlative degree between the technology subject and the effect subject in a specific field. Then, they identified core patent clusters containing either the same technology or the same effect, or both, in the 2-mode network consisting of technology-effect subject terms. During this analysis process, they merged technology subjects and effect subjects, and the method they used to process data was also

a 3-mode network analysis. Wei and Li (2014) and Wei et al. (2014) used research articles on knowledge management to build a 3-mode network including authors, keywords, and periodicals, which revealed, via visualization, the characteristics and development trends of knowledge management. To address the excessive outliers problem involved in authors co-occurrence analysis, Teng (2015) analyzed the co-occurrence of authors, institutions, and countries to build a mixed co-occurrence network that was based on the theory of a super-network. His empirical research showed that the method could not only eliminate the isolated nodes in the authors co-occurrence network, but also enhance the amount of information in the network by adding the coupling relations between authors and institutions or countries.

Matrix fusion of multi-relational data

Braam et al. (1991) combined cluster analysis of co-citations with topic words analysis to identify research topics. In their method based on clusters of co-citation analysis, they counted the frequency of words contained in cited references, which was used to test whether the clustering method could gather literature possessing similar topic terms into a cluster. Calero-Medina and Noyons (2008) analyzed the knowledge creation and flow process between scientific publications by combining co-occurrence of topic terms and citation network analysis. They used co-occurrence of topic terms to find related topic terms and theories, and used citation network analysis to discover key literature in the field. Zhang et al. (2007) indicated that cluster analysis of co-occurrence topic terms and strategic coordinate analysis were both powerful tools to discover discipline hotspots and that the combination of co-occurrence topic terms and citing frequency could lead to better results. The rest of this section discusses two main types of relation fusion: the fusion between two types of relations and the fusion among the triple relations.

Fusion between two types of relations

In the field of information retrieval, Weiss et al. (1996) developed a prototype system of a hierarchical network search engine, the HyPursuit system, which was used for retrieving and browsing by detecting the clusters of content-links to hypertext documents. The content-link clustering algorithm was based on a literature-similarity function, which considered the similarity of the topic terms and the hyperlinks similarity factor. The results of the function constituted the maximum value of similarities. Using the approaches based on reference links and co-words, Small (1998) identified the direct and indirect connection relationship between literature items. Antal and Millinghoffer (2006) proposed that the relatedness of two variables can be based on direct relations in probabilistic graphical models and on direct–indirect literature similarity. Direct literature similarity means that domain variables are related if their descriptions are similar, and indirect literature similarity means that domain variables are related if the same documents are similar to their descriptions. Janssens (2007) and Janssens et al. (2008) conducted research that involved combining Web content with hyperlinks and merged the relationship based on topic terms and the relationship based on bibliographic coupling. They used transformations obtained by using Fisher's inverse Chi square method for constructing new relational data sets. The results of the empirical study showed that their method was effective in finding the structures of a specific research field. Moreover, Janssens et al. (2009) also integrated cross-citation of journals with text mining, and then validated and improved the existing classification scheme of topics.

Fusion in triple relations

Wang and Kitsuregawa (2002) proposed a clustering algorithm based on content-link coupling to retrieve Web pages. They integrated outbound links, inbound links, and terms to improve retrieval performance. He et al. (2001, 2002) proposed a method of Web text clustering to merge the structure of text-based hyperlinks, co-citations, and text content. They used the structure of text-based hyperlinks to calculate similarities, the intensity of which was moderated by text similarities, and then integrated both the similarity of hyperlink structures and text similarities with co-citations by using linear weighting to build a weighted adjacency matrix.

Evaluation of relational fusion results

As compared with multi-relationship clustering results, the evaluations of single-relationship clustering results differ in studies. For instance, the results of an experiment conducted by Calado et al. (2006) showed that Web page retrieval based on link relations was superior to text classification, but the results of other experiments showed that similarity clustering based on topic terms was superior to similarity clustering based on citations (Janssens 2007; Janssens et al. 2008). All the experiments indicated that the clustering results after the fusion were better than the single-type relationships clustering results.

In the field of scientometrics, linear fusion is the primary algorithm used in relational fusion research. However, MSDF is a complex issue, and the three main types of data relationship are frequently correlated instead of being independent. For this reason, a simple linear operation is not sufficient to solve the problem of data fusion. Still, we can learn the methods of MSDF from other research fields, such as sensors and automation, to improve and enrich the MSDF methods for scientometrics analysis.

Research and application of relational fusion

Cross-integration of multi-mode data

Cross-integration of multi-mode data is typically used to visualize the associations among different data. Currently, the module identification of multi-mode data is a hot research topic in complex network research. In the case of 2-mode network, community detection methods can be roughly divided into two types: mapping and non-mapping methods. The mapping method consists of converting 2-mode networks into 1-mode networks, which leads to information loss and an incorrect reflection of the nature of the original networks. Latapy et al. (2008) summed up three drawbacks of the mapping method: (1) information loss, (2) an increase in the number of edges in the entire network, and (3) an increase in unnecessary noisy information that did not exists in the original network.

To ensure the accuracy of the analysis process, the non-mapping method, which can directly identify the module on the original 2-mode network, is more reliable. Guimerà et al. (2007) and Barber (2007) respectively defined the modularity based on a 2-mode network and proposed a community discovery algorithm and both algorithms are aimed to maximize the modularity, but differ in the manner in which they accomplish this.

Matrix fusion of multi-relational data

Matrix fusion of multi-relational data is an important type of MSDF. Guo (2005) summarized four main and basic data fusion algorithms in the sensor field: probability theory, evidence theory, fuzzy set approach, and neural networks.

Probability theory

In probability theory, data with low confidence, which is determined by analyzing the compatibility of various sensor data, is removed. Then, according to the known prior probabilities, obtaining optimal fusion results involve utilizing Bayesian probability to estimate data with higher confidence. The merit of probability theory is that the procedure is concise and easy to handle. The drawback is that prior probabilities are not always easy to obtain and if these probabilities are not consistent with the facts, the fusion results are inaccurate.

Evidence theory

As an uncertainty reasoning method, evidence theory was first proposed by Dempster and was established formally and gradually evolved and matured by Shafer. Consequently, this method is also known as the Dempster–Shafer or D–S evidence theory. D–S evidence theory is an extension of probability theory through the introduction of belief, basic assignment, and likelihood functions to handle uncertain problems. Unlike probability theory, evidence theory uses intervals to determine the likelihood function of evidence, and can also calculate the value of the likelihood function when the hypothesis is true. Evidence theory is an effective information fusion method that can meet a weaker condition than Bayesian probability and can distinguish unknown or uncertain information and perform the fusion process at different levels with stronger fault-tolerant capabilities.

However, when a conflict occurs between items of evidence, D–S evidence theory may produce results inconsistent with the facts, thus leading to its failure. Another deficiency of evidence theory is that it is not easy to determine the basic assignment function and composition formula; in particular, there is no unified method in the basic assignment function.

Fuzzy set approach

The fuzzy set approach addresses the uncertainty in information fusion with specific models and uses fuzzy reasoning to complete the data fusion. Fuzzy clustering is a process for classifying samples into groups, merging characteristic parameters, and classifying the samples. The fuzzy set approach has the advantage of logical inference. As compared with probability statistics, the fuzzy set approach is closer to people's mode of thinking, more easily overcomes some of the problems that probability theory faces, and is more suitable for fusion at a higher level. However, logical inference method is not mature and its construction is not sufficiently systematic, and it is strongly influenced by subjective factors in describing and processing information.

Neural networks

The neural network algorithm was proposed based on studies by subject specialists that examined how animals process information. It has the characteristics of strong fault tolerance, hierarchy, self-learning, adaptability, and parallel processing capabilities and is able also to simulate complex nonlinear functions. Currently, neural networks have been successfully applied to the state evaluation of information fusion. Although the neural network has a strong nonlinear processing capability that can be effectively used in information fusion technology, it may lead to local minimization, slow convergence, and sample-dependent or other issues.

Meta-analysis

In addition, meta-analysis is a systematic evaluation method. By an integration of multiple studies having the same purpose, meta-analysis uses quantitative method to produce the final evaluation results. Therefore, meta-analysis is also a data fusion analysis method. The biggest advantage of meta-analysis is an increase in the sample size, which helps to improve the accuracy of research findings and to eliminate the inconsistencies between individual studies. Han and Zhu (2014) systematically summarized the four methods of meta-analysis, that is, integration methods based on *P* value, rank, effect size, and counting.

Khaleghi et al. (2013) provided a generic view of multisensor data fusion methodologies, including the potential advantages, challenging aspects, and existing methodologies, as well as the recent developments and emerging trends. In particular, they discussed the existing data fusion methods relying on a data-centric taxonomy, and explored each method based on the specific data-related challenging aspects. Safari et al. (2014) studied the data fusion problem for asynchronous, multirate, multisensor linear systems and presented a comprehensive state space model that includes all the sensor systems. A linear system is observed by multiple sensor systems, each having a different sampling rate.

In addition to the data fusion algorithms in the sensor field, recently more matrix fusion methods for multi-relational data in other fields have been developed. Snidaro et al. (2015) provided the comprehensive status of recent and current research on context-based information fusion systems, tracing back the roots of the original thinking behind the development of the concept of "context." Yaqoob et al. (2016) provided a comprehensive view of information fusion in social big data, including the foundations, state-of-the-art research, applications, challenges, and future research directions. Xu and Zhao (2016) presented an overview of the existing intuitionistic fuzzy decision-making theories and methods from the perspective of information fusion, involving the determination of attribute weights, the aggregation of intuitionistic fuzzy information, and the ranking of alternatives. They also described the potential challenges in future research, as well as providing a survey of recent applications of the discussed theories and methods in various fields.

Analysis of clustering ensembles

Clustering analysis

Clustering analysis is derived from data mining and statistics, which is the core issue of knowledge discovery, machine learning, artificial intelligence, pattern recognition, and other application areas. The clustering technique divides data objects into several clusters,

in which objects in the same cluster are more similar, while objects in different clusters are more different.

Kogan et al. (2006) mapped recent clustering algorithms to their applications. According to the specific clustering rules and the method used to apply these rules, clustering algorithms can be divided into four types, respectively based on partitioning, hierarchy, density, and grids. Important recent clustering algorithms include the clustering algorithm based on computational intelligence (Boryczka 2009; Bharne et al. 2011), the semi-supervised clustering algorithm (Tang et al. 2007; Zhao et al. 2012), the spectral clustering algorithm (Wang et al. 2011), and so on.

Ensemble clustering

The cross-integration of multi-mode data and matrix fusion of multi-relational data forms a comprehensive matrix. When clustering analysis is applied to the comprehensive matrix, different clustering algorithms may lead to different clustering results. Under the circumstances, we need ensemble clustering to obtain a more reasonable clustering through integrating these different clustering results. Ensemble clustering involves using a- specific fusion function to cluster different relational data respectively and merge the results into one cluster. Prior to ensemble clustering, an assessment of each clustering result is necessary to help select the appropriate clustering method.

Ensemble clustering is an active machine learning research area (Yi et al. 2012). However, the literature shows that few research efforts have used ensemble clustering algorithms in scientometrics to analyze scientific structure. Clustering ensembles use different parameters to obtain a large number of cluster memberships, and then fuse the cluster memberships to obtain the final clustering results (Strehl et al. 2009).

Clustering ensemble algorithms can be divided into two processes. First, they use different clustering algorithms to generate a large amount of initial cluster memberships. Then, they use a fusion function to integrate initial cluster memberships into the final clustering results. As compared with the single clustering algorithm, clustering fusion can reflect the characteristics of a data set from different aspects, and combine these characteristics to improve performance, robustness, and accuracy. In addition, clustering fusion is suitable for parallel data processing, especially for distributed data sets, in that it clusters each distributed data set in parallel and integrates the clustering results into a final clustering result.

Clustering fusion functions

Fusion functions, also known as consensus functions, are the key to ensemble clustering research. Common fusion functions include the co-association matrix method (Fred 2001; Fred and Jain 2002a, b), the Hypergraph-Partition method (Strehl and Ghosh 2003; Fern and Brodley 2004; Ayad and Kamel 2003), the methods based on information theory (Ayad et al. 2004), hybrid models (Topchy et al. 2004), voting (Zhou and Tang 2006), evidence accumulation (Fred and Jain 2002a, b), and neural networks (Yang et al. 2006).

The differences among cluster memberships can significantly affect the final fusion result. Therefore, the selection strategy based on clustering differentiation usually consists of selecting the cluster memberships with a large differentiation in the members participating in the fusion, referred to as fusion members. The key to this method is to measure the degree of difference between the cluster members. Common measurement methods include normalized mutual information (NMI) (Fred and Jain 2002a, b; Strehl et al. 2002;

Zhou and Tang 2006), the rand index (RI) (Banerjee and Langford 2004), the Jaccard index (JI) (Zhang et al. 2012), the adjusted rand index (ARI) (Hubert and Arabie 1985), clustering error (CE) (Meila and Heckerman 2001), variation of information (VI) (Meila 2005), and so on. The methods used to measure the quality of cluster memberships include the F-measure, RI, JI, the cophenetic correlation coefficient, and the composed density between and within clusters index (CDbw), etc. (Yang et al. 2008).

Future methods of MSDF in scientometrics

A future research model of MSDF in scientometrics

In this paper, a future research application model and procedure of MSDF in scientometrics is proposed. As shown in Fig. 2, the method can realize fusion among different types of information, data relations, and clustering results. The model and procedure can be divided into three procedures: data type integration, fusion of data relations, and ensemble clustering. Following the conceptual framework of the study of Pavlidis et al. (2001), here we divide the complete MSDF model and process in scientometrics into three procedures: early fusion, mid fusion and late fusion. Early fusion constitutes the fusion of data types, mid fusion the fusion of data relations, and late fusion ensemble clustering.

Here, we discuss the most suitable methods for each condition (Fig. 2). Since this an ideal model, in some cases of actual analysis, only one data type or relation can already achieve a satisfactory analysis result; in this case, it is not necessary to consider all the procedures shown in Fig. 2.

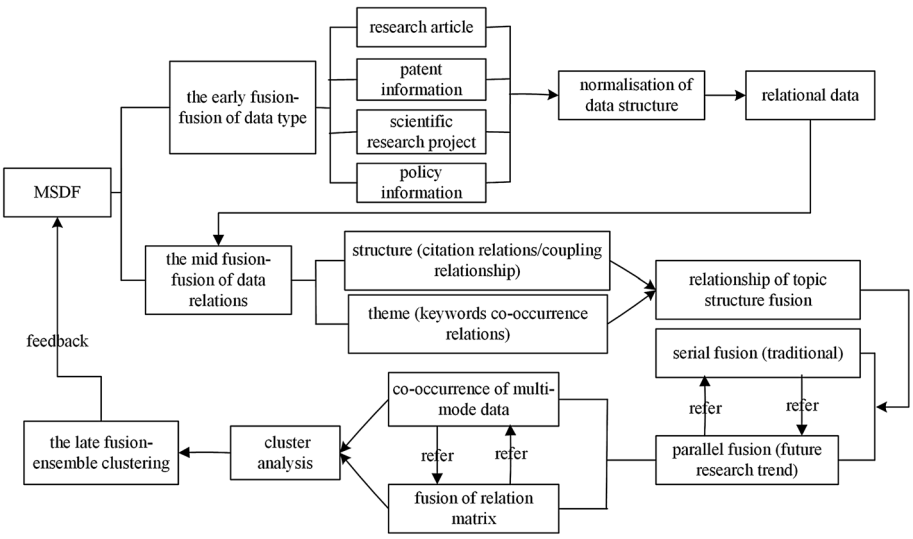

Fig. 2 Application model and procedure of MSDF in scientometrics

Early fusion: fusion of data types

First, a collection of a variety of data sources, such as journal articles, conference information, dissertations, patent information, project information, book information, and industrial economic data, should be included in the scope of scientometrics analysis, because different data types reflect different technology contents. For example, scientific papers focus on basic scientific research efforts, patent information focuses on technological innovations, and industrial economic data aim to grasp technology market information. Therefore, comprehensively considering multi-source information can lead to a more objective output from scientometrics analysis.

Mid fusion: fusion of data relations

Second, various data relations should be obtained and fused. There are basically two means of accomplishing multi-source data fusion. One consists of obtaining a variety of associated data types respectively and fusing relation matrices of different data types by mapping, and the other consists of directly identifying the community topics of multi-mode data. Both methods can enhance the data relationship strength by acquiring complementary information. In fact, cross-integration of multi-mode data and matrix fusion of multi-relational data each has its own fusion method, and they can be combined in actual application.

Cross-integration of multi-mode data identifies modules by learning from non-mapping identification methods of complex heterogeneous networks, and considers the relationships among different dimensions in specific problems of scientometrics analysis. However, visualization for cross co-occurrence of multi-mode data, for example, visualization of the information and association of more dimensions, which is important for knowledge discovery, still has considerable room for improvement.

Matrix fusion of multi-relational data can introduce existing fusion methods from the field of sensors,automation, and so on. According to the objects and characteristics of scientometrics, we can improve these methods and eventually form fusion methods for application in scientometrics. Multiple matrix fusion techniques exist, and each has a unique set of advantages, and therefore, a hybrid approach that combines them will lead to an enhanced matrix fusion technique. For example, evidence theory method is simple and effective, but the typical D–S theory method is sensitive to highly conflicting evidence. Neural network algorithms are relatively complex, but are characterized by strong fault tolerance, hierarchy, self-learning, self-adaptation, and parallel processing capabilities. Thus, we can combine evidence theory method with neural networks to achieve the fusion of relational data according to the data characteristics and the complexity of the analysis problem.

Late fusion: ensemble clustering

The final procedure of MSDF is ensemble clustering to improve the accuracy of cluster analysis. With the continuous improvement of clustering analysis methods, an ensemble clustering algorithm can be introduced to merge a variety of clustering results into a final one, and therefore, it is crucial to choose the most effective clustering algorithm and consensus function so that the ensemble clustering is more effective. Figure 3 shows the ensemble clustering process.

As shown in Fig. 3, the clustering fusion algorithm can be divided into two steps. First, many initial cluster memberships can be generated by different clustering algorithms. Given data

cluster object sets $X = \{x_1, x_2, \ldots, x_N\}$, which comprise N data objects, after H times clustering for data sets X, we can obtain cluster member set $P = \{p_1, p_2, \ldots, p_H\}$, containing H clustering results, of which the i-th clustering results are the set p_i, $p_i = \left\{C_i^1, C_i^2, \ldots, C_i^{k_i},\right\}$, $(I = 1, 2, \ldots, H)$, where k_i is the count of cluster member p_i. Clustering algorithms include partition clustering, hierarchy clustering, density clustering, grid clustering, computational intelligence clustering, semi-supervised clustering, and spectral clustering algorithms.

In the next step, an integration process accesses the initial cluster memberships and uses a fusion function to create a final cluster. The purpose of the clustering fusion is to design a fusion function Γ, which merges all cluster members, namely, $p_1, p_2, \ldots, p_H$, into the final clustering result P^f.

As compared with a single clustering algorithm, a clustering fusion algorithm can reflect the characteristics of a data set from different aspects, combine these characteristics to improve the clustering performance, and also improve the robustness and accuracy of the clustering. A clustering fusion algorithm, which can cluster all the distributed data in parallel and then integrate the clustering results into a final clustering result, is suitable for parallel data processing, especially for distributed data.

Data fusion focused on topic identification

To obtain a clearer and deeper analysis of the MSDF model, we further focus our research on topic identification based on text analysis of scientific literature. We give more descriptions of the process of accessing the relation matrix and accomplishing fusion. Since our study concerns the fusion of data relations, here we do not further explain the methods of data type fusion, but pay more attention to data relation fusion in the process of topic identification based on text analysis.

In this section, we describe two types of relational data fusion processes. Type I consists of relation matrix integration, and Type II consists of community identification through multi-mode cross-integration. The two types of analysis results can complement each other.

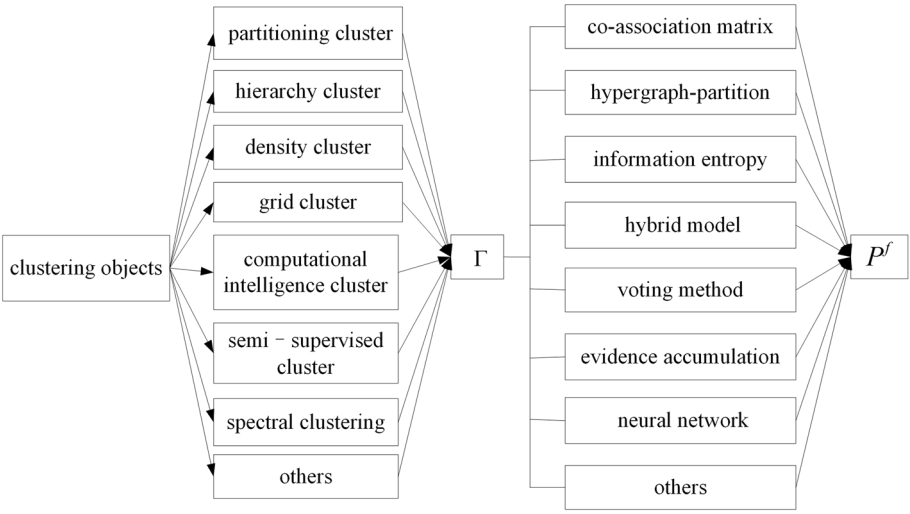

Fig. 3 Ensemble clustering process

Type I

Relation matrix integration includes (1) obtaining various types of data relationship that relate to the analysis target, (2) projecting the variety of relational data to the target relation, and (3) generating the target analysis matrix through the matrix fusion operation.

The first step is to obtain various matrices of topic relationships. In scientific literature, there is a link of topic content between topic terms, article authors, and the cited reference. Therefore, an integrated analysis of the relationship between topic terms, article authors, and cited reference can overcome the lack of information that results from a single relationship and obtain a more accurate discovery of topics. In this study, we choose to acquire three types of relational matrix: a co-occurrence matrix of topic terms, a coupling matrix of author-topic terms, and a coupling matrix of citations-topic terms. In the present study, "coupling" is used in its broad sense; it is not confined to the traditional literature coupling concept, but instead represents a means of measurement by two or more interdependent entities and interrelated factors between items. Figure 4 shows the schematic of the topics-related multi-relationship used in our study.

First, a co-occurrence matrix of topic terms from documents constitutes the most direct expression of inter-related topics, and therefore, we first access a co-occurrence matrix of topic terms, as the basis matrix for fusion of the relation matrices in topics identification. Additionally, the topic relevance relationship includes a reinforcing association and a new added association. Here, we define various topics associations within the original literature sets as a strengthened type of the co-occurrence matrix of topic terms (the basis one), while those beyond the original literature set belong to the new added association. The reinforcing association can strengthen the basis matrix, because the new added association introduces additional information, which did not exist in the original data set. Both these types of association are important for complementing the base associated matrix, and they jointly constitute an enhanced relation matrix for topic identification.

Second, in the big science era, with its transdisciplinary characteristics, collaborative research is primarily employed. The process of collaboration is accompanied by the diffusion of innovation and knowledge; in particular, tacit knowledge diffusion occurs through research collaborations. Therefore, it is reasonable to mine topic similarity through author collaboration. In the co-authorship analysis, the co-occurrence matrix of topic terms covers topics related to information provided by the author collaboration, and therefore it belongs to the reinforcing association.

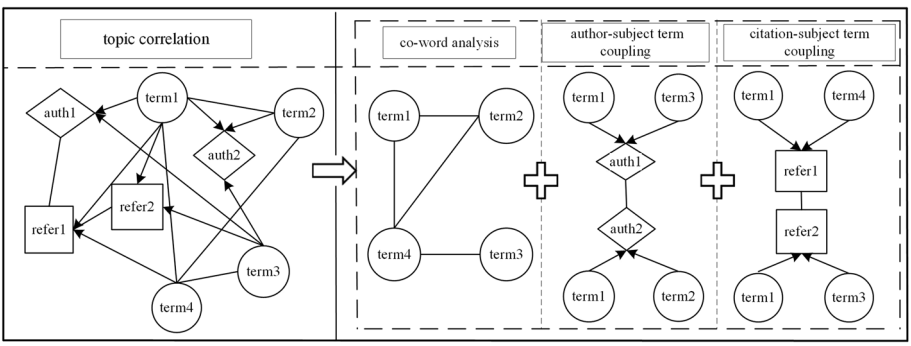

Fig. 4 Various types of data relation matrix fusion

However, not all literatures are the result of author collaborations. A small amount of collaboration reflects the joint research interests of the co-authors. In particular, collaborations of authors from different disciplines significantly facilitates the identification of interdisciplinary topics. That is, the coauthor phenomenon occurs in authors working in the same or similar discipline or on interdisciplinary research. Therefore, to explore topics associated with the non-collaboration efforts of those collaboration ones belongs to the new added association. In particular, obtaining author-topic terms coupling relationships is an effective means of identifying interdisciplinary topics.

Third, a citation network reflects the longitudinal continuity and succession in the scientific knowledge dissemination, including the horizontal cross and penetration. It is also an excellent means of studying knowledge dissemination. Thus, the coupling matrix of citations-topic terms can also reflect the topic similarity in the knowledge diffusion process.

Similarly, the co-occurrence matrix of topic terms covers topics-related information provided by the coupling of citations-topic terms, and therefore, it belongs to the reinforcing association. However, the coupling relationship of topic terms in different literatures belongs to the new added association. It reflects citations as an intermediary of the dissemination of knowledge innovation.

When three kinds of relational matrix, the co-occurrence matrix of topic terms, coupling matrix of author-topic terms and coupling matrix of citations-topic terms, are acquired, we can choose the appropriate matrix fusion method to realize the three relationship fusion.

Type II

For Type II topic identification, which through a discovery from a multi-relation community, more complex networks are needed. Cross-integration of multi-mode data can directly identify the relation communities of the multi-mode network that can be formed, which can save more relationships of variables and avoid data distortion during projection (Fig. 5).

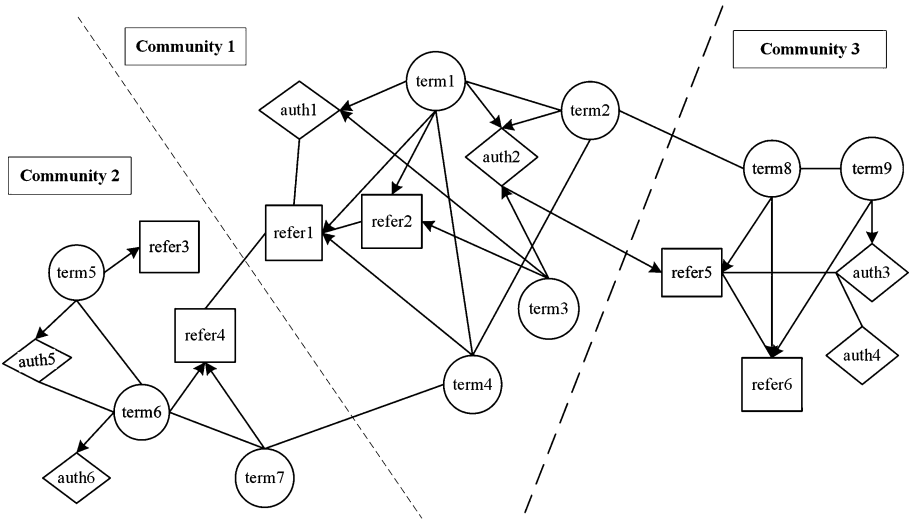

Fig. 5 Community identification in cross-integration multi-mode data

Discussion of the application of MSDF in scientometrics

In this article, we proposed a future research model and procedure of MSDF in scientometrics. The model and procedure can be divided into three parts: data type integration, fusion of data relations, and ensemble clustering. Here, we discuss the most suitable methods for different conditions.

In the early fusion stage (data integration), different data types reflect different contents of technology, according to the section "Future Method of MSDF in Scientometrics." Therefore, it is suggested that different data types be collected according to the specific analysis target. When the objective is to discover the hot topics or the forefront topics of basic research areas, the collection and integration of the data types should be focused more on conference papers, journal articles, and dissertations, while when the objective is to explore the industry competitive intelligence, such as predicting the future technology trends, the collection and integration of the data types should place more emphasis on patent documents, project information, industry and economic data.

In the mid fusion stage (fusion of data relations), the type of fusion method that should be chosen depends on the complexity and analysis targets of the relationship types. If the data type is simple and the researcher is familiar with the importance of the various relationships in the evaluation, he/she can use linear mode integration. However, with the growth in the availability of data relationships, prior to analysis it is difficult to acquire a perfect understanding of a variety of data relationships. In this situation, a visualization of the cross-integration of multi-mode data can approximately show the relationship of objects, which will facilitate the acquisition of more information about various data relationships. On this basis, we can further perform the relation fusion through probability theory, evidence theory, the fuzzy set approach, neural networks, or meta-analysis.

When the relationships to be analyzed are simple and clear, it is suggested that probability theory or meta-analysis be used. If we have a sufficient understanding of the attributes of various analyzed relationships before fusion analysis, we can easily apply a fuzzy set approach. However, when we lack detail information about the various analyzed relationships, we need to seek more automated methods, such as evidence theory and neural networks. D–S evidence theory may produce results that are inconsistent with the facts, whereas neural networks may lead to a local minimization result. Therefore, a combined application of D–S evidence theory and neural networks is a better choice for executing the relation fusion.

In the late fusion stage (ensemble clustering), the clustering results of a relation matrix can be optimized by ensemble clustering. It is important to choose the appropriate cluster algorithm and fusion function according to the data characteristics. We can choose several cluster algorithms simultaneously; for example, hierarchy and spectral clustering algorithms can be a complementary pair or computational intelligence and semi-supervised clustering can be performed simultaneously.

Conclusion

This paper provided an introduction to MSDF. Through investigation, we found that the most widely used methods of MSDF in scientometrics is the linear mode with its simple and random fusing process. Since we assume that the improvement of MSDF methods requires a solid mathematical foundation, the breakthrough in MSDF in the future may

come from advanced fields of data fusion research, such as sensor theory and bioinformatics. According to this assumption and considering the features of scientometrics, we proposed a novel MSDF research model and procedure for scientometrics. The model and procedure can be divided into three parts: data type integration, fusion of data relations, and ensemble clustering. Furthermore, the fusion of data relations can be divided into the cross-integration of multi-mode data and matrix fusion of multi-relational data. To achieve a clearer and deeper analysis of the proposed MSDF model, we performed topic identification based on the text analysis of scientific literatures that uses MSDF methods. Finally, we discussed the most suitable methods for different conditions.

By acquiring complementary information, MSDF methods can enhance the strength of data relationships. In fact, cross-integration of multi-mode data and matrix fusion of multi-relational data each has its own fusion method, such they can be combined in actual application. Cross-integration of multi-mode data identifies modules by learning from non-mapping identification methods of complex heterogeneous networks, and considers the relationships among different dimensions in specific problems of scientometrics analysis. However, the visualization of cross co-occurrence of multi-mode data still has considerable room for improvement, for example, visualizing the information and association of more dimensions, which is important for knowledge discovery.

In the future, it will be necessary to consolidate the theoretical and computational bases of MSDF, since the data-driven model of MSDF has to handle very large heterogeneous data sets. As mentioned above, in our opinion, a significant breakthrough may occur in research areas that have a solid mathematical foundation, such as sensor theory and bioinformatics. As an application oriented methodology, scientometrics could constantly adopt advanced methods of MSDF from these areas, such as biomedicine, where MSDF was promoted by the availability of multiple linked, rich omic level data, while taking into account its own characteristics of scientometrics.

Acknowledgements This work was supported by the National Social Science Fund of China (Grant No. 14CTQ033), Natural Science Foundation of Shandong Province, China (Grant No. ZR2015GL015), and the Youth Innovation Fund of Promotion Association, CAS.

References

Ahlgren, P., & Colliander, C. (2009). Document–document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics, 3*(1), 49–63.

Antal, P., Fannes, G., Timmerman, D., Moreau, Y., & De Moor, B. (2004). Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine, 30*(3), 257–281.

Antal, P., Glenisson, P., & Fannes, G. (2002). On the potential of domain literature for clustering and Bayesian network learning. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 405–414).

Antal, P., & Millinghoffer, A. (2006). Learning causal bayesian networks from literature data. *Periodica Polytechnica Electrical Engineering, 50*(3/4), 201.

Ayad, H., Basir, O., & Kamel, M. (2004). *Multiple classifier systems* (pp. 144–153). Heidelberg: Springer.

Ayad, H., & Kamel, M. (2003). *Advances in intelligent data analysis V* (pp. 307–318). Heidelberg: Springer.

Banerjee, A., & Langford, J. (2004). An objective evaluation criterion for clustering. In: 10th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 515–520).

Barber, M. J. (2007). Modularity and community detection in bipartite networks. *Physical Review E, 76*(6), 1–11.

Bharne, P. K., Gulhane, V., & Yewale, S. K. (2011). Data clustering algorithms based on swarm intelligence. In: 3rd international conference, electronics computer technology (ICECT) (pp. 407–411).

Boryczka, U. (2009). Finding groups in data: Cluster analysis with ants. *Applied Soft Computing, 9*(1), 61–70.

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

Calado, P., et al. (2006). Link-based similarity measures for the classification of Web documents. *Journal of the American Society for Information Science and Technology, 57*(2), 208–221.

Calero-Medina, C., & Noyons, E. C. (2008). Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. *Journal of Informetrics, 2*(4), 272–279.

Couto, T., et al. (2006). A comparative study of citations and links in document classification. In: 6th ACM/IEEE-CS joint conference on digital libraries (pp. 75–84).

Fern, X. Z., & Brodley, C. E. (2004). Solving cluster ensemble problems by bipartite graph partitioning. In: Twenty-first international conference on Machine learning (p 36).

Fred, A. (2001). *In multiple classifier systems* (pp. 309–318). Heidelberg: Springer.

Fred, A. L., & Jain, A. K. (2002). Data clustering using evidence accumulation. In: 16th international conference on pattern recognition (pp. 276–280).

Fred, A., & Jain, A. K. (2002b). *Structural, syntactic, and statistical pattern recognition* (pp. 442–451). Heidelberg: Springer.

Guimerà, R., Sales-Pardo, M., & Amaral, L. A. N. (2007). Module identification in bipartite and directed networks. *Physical Review E, 76*(3), 036102.

Guo, H. Y. (2005). Research and development of multi-sensor information fusion technology. *Bulletin of National Science Foundation of China, 19*(1), 17–21.

Han, M. F., & Zhu, Y. P. (2014). Applications of meta-analysis in multi-omics. *Chinese Journal of Biotechnology, 30*(007), 1094–1104.

He, X., Ding, C. H., Zha, H., & Simon, H. D. (2001). Automatic topic identification using webpage clustering. In IEEE International Conference on Data Mining, ICDM 2001 (pp. 195–202).

He, X., Zha, H., Ding, C. H., & Simon, H. D. (2002). Web document clustering using hyperlink structures. *Computational Statistics & Data Analysis, 41*(1), 19–45.

Hua, B. L. (2013). Research on the methods of multi-source fusion. *Information Studies: Theory & Application, 36*(11), 16–19.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218.

Janssens, F. (2007). *Clustering of scientific fields by integrating text mining and bibliometrics* (pp. 124–129). Leuven: Katholieke Universiteit Leuven.

Janssens, F., Glänzel, W., & De Moor, B. (2008). A hybrid mapping of information science. *Scientometrics, 75*(3), 607–631.

Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and Management, 45*(6), 683–702.

Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion, 14*(1), 28–44.

Klavans, R., & Boyack, K. W. (2006). Quantitative evaluation of large maps of science. *Scientometrics, 68*(3), 475–499.

Kogan, J., Nicholas, C., & Teboulle, M. (2006). *Grouping multidimensional data* (pp. 28–29). Heidelberg: Springer.

Latapy, M., Magnien, C., & Del Vecchio, N. (2008). Basic notions for the analysis of large two-mode networks. *Social Networks, 30*(1), 31–48.

Leydesdorff, L. (2010). What can heterogeneity add to the scientometric map? Steps towards algorithmic historiography. In *Débordements Mélanges Offerts À Michel Callon* (pp. 283–289).

Meilă, M. (2005). Comparing clusterings: an axiomatic view. In: *22nd international conference on machine learning* (pp. 577–584).

Meilă, M., & Heckerman, D. (2001). An experimental comparison of model-based clustering methods. *Machine Learning, 42*(1–2), 9–29.

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning, 52*(1), 91–118.

Mooghali, A., Alijani, R., Karami, N., & Khasseh, A. (2011). Scientometric analysis of the scientometric literature. *International Journal of Information Science and Management, 9*(1), 19–29.

Morris, S., DeYong, C., Wu, Z., Salman, S., & Yemenu, D. (2002). DIVA:a visualization system for exploring document databases for technology forecasting. *Computers & Industrial Engineering, 43*(4), 841–862.

Morris, S. A., & Van der Veer Martens, B. (2008). Mapping research specialties. *Annual Review of Information Science and Technology, 42*(1), 213–295.

Morris, S. A., & Yen, G. G. (2004). Crossmaps visualization of overlapping relationships in collections of journal papers. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5291–5296.

Pang, H. S. (2012). *Research on knowledge discovery method based on multiple co-occurrence* (pp. 116–119). Beijing: University of Chinese Academy of Science.

Pavlidis, P., Weston, J., Cai, J., & Grundy, W. N. (2001). Gene functional classification from heterogeneous data. In *Proceedings of the fifth annual international conference on computational biology* (pp. 249–255).

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.

Safari, S., Shabani, F., & Simon, D. (2014). Multirate multisensor data fusion for linear systems using Kalman filters and a neural network. *Aerospace Science and Technology, 39,* 465–471.

Scientometrics. Journal Description. http://link.springer.com/journal/11192. 2016-7-14.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for Information Science and Technology, 60*(3), 571–580.

Small, H. (1998). A general framework for creating large-scale maps of science in two or three dimensions: The SciViz system. *Scientometrics, 41*(1–2), 125–133.

Snidaro, L., García, J., & Llinas, J. (2015). Context-based information fusion: a survey and discussion. *Information Fusion, 25,* 16–31.

Strehl, A., & Ghosh, J. (2003). Cluster ensembles-a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research, 3,* 583–617.

Su, N. (2011). *On methods of multiple relations fusion in research field analysis* (pp. 16–20). Beijing: University of Chinese academy of science.

Su, N., & Zhang, Z. Q. (2010). On the multiple relation fusion research in scientometrics. *Information Science, 28*(9), 1309–1313.

Tang, W., Xiong, H., Zhong, S., & Wu, J. (2007). Enhancing semi-supervised clustering: A feature projection perspective. In: *13th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 707–716).

Teng, L. (2015). Research on hybrid co-occurrence network of author-Institution-country based on super-network. *Journal of The China Society for Scientific and Technical Information, 34*(1), 28–36.

Topchy, A., Jain, A. K., & Punch, W. (2004). A mixture model for clustering ensembles. Society for industrial and applied mathematics. In *SIAM International Conference on Data Mining* (pp. 379).

Van Den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics., 68*(3), 377–393.

Van Raan, A. (1997). Scientometrics: State-of-the-art. *Scientometrics., 38*(1), 205–218.

Wang, Y., Jiang, Y., Wu, Y., & Zhou, Z. H. (2011). Spectral clustering on multiple manifolds. *IEEE Transactions on Neural Networks, 22*(7), 1149–1161.

Wang, Y., & Kitsuregawa, M. (2002). Evaluating contents-link coupled web page clustering for web search results. In *ACM CIKM international conference on information and knowledge management* (pp. 499–506).

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods, 11*(3), 333–337.

Wei, X. Q., & Li, C. L. (2014). The research of science collaboration behavior based on author-year-keyword network—The case of the library and information science. *Journal of Intelligence, 11,* 117–123.

Wei, X. Q., Li, C. L., & Liu, F. F. (2014). Construction and visualization of 3-mode network: A literature study of knowledge management of library and information science. *Information Studies: Theory & Application, 08,* 74–78.

Weiss, R., Vélez, B., & Sheldon, M. A. (1996). HyPursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In: *7th ACM conference on hypertext* (pp. 180–193).

Xu, H. Y., & Fang, S. (2014). Core patents mining based on cross co-occurrence analysis to patent technology-effect subject terms and citations. *Library and Information Service, 33*(2), 158–166.

Xu, Z., & Zhao, N. (2016). Information fusion for intuitionistic fuzzy decision making: an overview. *Information Fusion, 28,* 10–23.

Yang, Y., Kamel, M. S., & Jin, F. (2006). ART-based clustering aggregation. In *IEEE international conference on granular computing* (pp. 482–485).

Yang, Y., Xin, F., & Mohamed, K. (2008). Survey of clustering validity evaluation. *Application Research of Computers, 25*(6), 1630–1632.

Yaqoob, I., Chang, V., Gani, A., Mokhtar, S., Hashem, I. A. T., Ahmed, E., et al. (2016). Information fusion in social big data: Foundations, state-of-the-art, applications, challenges, and future research directions. *International Journal of Information Management*. doi:10.1016/j.ijinfomgt.2016.04.014.

Yi, J. F, Yang, T. B, Jin, R., Jain, A. K., & Mahdavi, M. (2012). Robust ensemble clustering by matrix completion. In *2012 IEEE 12th international conference on data mining* (pp. 1176–1181).

Zhang, H., Wang, X. Y., & Cui, L. Trend. (2007). Co-word analysis method combined with literature CI research thematic areas. *Information Studies: Theory & Application, 30*(3), 378–380.

Zhang, S., Wong, H. S., & Shen, Y. (2012). Generalized adjusted rand indices for cluster ensembles. *Pattern Recognition, 45*(6), 2214–2226.

Zhao, W. Z., Ma, W. Z., Li, Z. Q., & Shi, Z. Z. (2012). Efficiently active learning for semi-supervised document clustering. *Journal of Software, 23*(6), 1486–1499.

Zhou, Z. H., & Tang, W. (2006). Clusterer ensemble. *Knowledge-Based Systems, 19*(1), 77–83.