Multi-Source Data Fusion Study in Scientometrics

Hai-Yun Xu[1,2], Chao Wang[1,3], Hong-shen Pang[4], Li-jie Ru[1,3], and Shu Fang[1]

[1] Chengdu Library of Chinese Academy of Sciences, Chengdu, Sichuan 610041, P. R. China

[2] Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China

[3] University of Chinese Academy of Science, Beijing, 100190, P. R. China

[4] Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530,P.R. China

Abstract: This review provides an introduction to MSDF, and discusses the status quo of the methods and applications of MSDF in scientometrics study. Currently, one of the most widely used methods of MSDF in scientometrics is the linear mode with a simple and random fusing process. In this paper, we assume that the improvement of MSDF methods requires a strong mathematical foundation, and breakthrough of MSDF in future may come from advanced fields in data fusion research and their applications, such as sensors, the automation, etc. Based on this assumption, this review have investigated the main thoughts of MSDF used in those fields and proposed that MSDF could be divided into the fusion of data types and fusion of data relations. Furthermore, the fusion of data relations could be divided into the cross-integration of multi-mode data and matrix fusion of multi-relational data. This paper studied the methods and technological process of MSDF applicable to information analysis, especially in the competitive intelligence of scientific and technological area.

Keyword: Data fusion; Relations fusion; Multi-mode analysis; Multi-source Data; Scientometrics

1 Introduction

The association among different entities, such as scientific papers, patents, news, etc., is an important research topic in scientometrics. Typical associations include those based on citations, co-authors and topic terms.

According to diverse units, citation and topic words can be used to describe the relationship between literatures and also to describe or analyze the relationship between authors, journals and institutions. With the growing of scientific literatures, multi-type ones have become valuable data resources. Meanwhile, the analyzable types of relationship have also been enriched. Apart from citation and co-author relationships, various analysis methods based on coupling relationships have provided new insight. However, most of the current scientometrics analysis is mainly based on the single-type relationship. Only a few studies consider two types of relationships, for example, combining citation relationships with co-author relationships. Yet few studies consider three or more types of relationships.

Multi-source data fusion (MSDF) refers to comprehensively analyzing different types of information sources or relational data by a specific method, and utilizing information together to reveal characteristics of the research object for obtaining more comprehensive and objective measurement results. The MSDF study, which has been mostly concentrated in the sensor field, has also become a key subject of bioinformatics, artificial intelligence, face recognition and other disciplines. In recent

years, with the development of data science and complex network, clustering researches on the fusion of different networks have received more attention.

This paper systematically investigates current research and application on MSDF, then pertinently analyzes research progress of MSDF in areas such as sensor and automation. Subsequently, considering the features of scientometrics, we learn from the methods of MSDF in sensors or some other fields, and innovate these methods to be suitable for scientometrics analysis. In order to make up for the deficiencies of the function of single-type relationship, a function that can reveal correlation between associated entities is proposed and it can facilitate expansion of scientometrics analysis methods.

2 Overview of MSDF in scientometrics

2.1 Basic types of relation

We propose that multi-source data fusion can be divided into two kinds: fusion of data types and fusion of data relations.

The fusion of data types is to merge different data types into the same analysis object. Currently, data types mainly include journal articles, conference information, dissertations, patent information, project information, book information and so on. Hua(2013)[1]divided multi-source data into homogeneous information with heterologous source, heterogeneous information and multilingual information. And he indicated that fusion of data types was a basic work involving field mapping, field splitting, filtering repeated data, weighting heterogeneous data, which was inevitable step of data processing in future scientometrics analysis. This article focuses on the fusion of data relations, so there are few descriptions on the method of data types fusion here.

The fusion of data relations is to merge different data relations into a new one to characterize the relationship among entities. Based on citation relationship, Shibata et al. (2009) [2] conducted a comparative research on three networks, co-citation network, bibliographic coupling network and citation network. And results showed that citation network could find new topics earlier and was the most effective way of identifying research fronts. By contrast, co-citation network was the worst. In addition, the content clustering based on the citation network had both the highest similarity and the least risk of omitting a new research field. Klavans (2006) [3] discovered the clustering network built by direct citation, comparing with co-citation, it had more similar content. Couto et al. (2006) [4] indicated that bibliographic coupling approach was more effective than textual approach in the empirical analysis of computer-related literatures. Nonetheless, Ahlgren et al. (2009) [5] considered textual approach to be better in the empirical analysis for literatures on information retrieval.

Since the above analysis has their respective fields of application, it is hard to say which method is the best, although each has its own unique side and inadequate side. Morris et al. (2008) [6] pointed out that the scientometrics methods based on a certain relation could reflect only a limited understanding of science from one perspective. In other words, every relationship can help researchers to analyze only the partial characteristics of a research field from a certain perspective. It is only when observing a research field from different perspectives that we can fully understand it. And the integration of multiple relationships contributes to a comprehensive understanding of the problem.

2.2 Fusion of data relations

According to the different ways of fusion, the fusion of multi-source data relations can be divided into the cross- integration of multi-mode data and matrix fusion of multi-relational data.

Cross-integration of multi-mode data shows associations among different types of entities taking advantage of cross co-occurrence technique, but ignoring 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 relations, and then performs multivariate statistical analysis, such as cluster analysis, factor analysis, etc.

Both the cross- integration of multi-mode data and matrix fusion of multi-relational data eventually form a comprehensive matrix. For cross- integration of multi-mode data, the dimension of the newly formed matrix is the sum of dimensions of all the original matrixes. But for matrix fusion of multi-relational data, the dimension of the new matrix dose not change.

2.2.1 Cross- integration of multi-mode data

In social network analysis, mode refers to the collection of actors, specifically, the number of the types of these collections. 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 multi-mode network corresponds to more than three 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 only exist among different types of nodes, and vice versa. Taking a 2-mode network as an example, a 2-mode network data can be used to generate a bipartite graph whose vertices can be divided into two disjoint sets. By using mathematical language, this can be expressed as: let G=<V, E> be a graph, and V be the set of vertexes 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, $(x,y) \in E$, we can get $x \in V1$, $y \in V2$ or $x \in V2$, $y \in V1$.

Some researchers combined bibliography and words to discover research topics. We find out there are two approaches for the combination: one is to use the bibliography as a qualification of the relationship between the words, and the other is to use citation to build the relationship between bibliography and words. When using indexing terms in the study of domain description, we may encounter some inherent problems because of linguistic phenomenons, such as polysemy. However, bibliography provides a specific context for indexing terms which, to a certain extent, can avoid the above problems. Based on this consideration, Besselaar et al. (2006) [7] used the method of word-reference co-occurrence to describe research objects and cluster literature collections called "research front".

Leydesdorff (2010) [8] applied the theory of heterogeneous network to 3-mode network. He put feature items of authors, journals, keywords linked together and showed different types of nodes in the same network, which could analyze relationships between nodes and have a more realistic reflection of research networks. Morris (2002, 2004) [9] [10] used the association of two co-occurrence matrixes with same feature items to conduct an application study on the cross-chart and timeline method, and solved the problem of visualization revealing association between two feature items. Pang (2012) [11] improved the presentation method of Morris' cross-chart. The improved cross-chart not only could reveals the association between two feature items, but also could presents the co-occurrence among three feature items. During a patent-based tech mining analysis, Xu et al. (2014) [12] used association rules between technological dimension and functional dimension in patent technology-effect matrix to

acquire the correlative degree between the technology subject and the effect subject in a certain field. Then they identified core patents or patent clusters of the same technology-effect, same technology or same effect in the 2-mode network consisting of technology-effect subject terms. During this analysis process, they merged technology subjects and effect subjects together, and the method they used to process data was also a kind of 3-mode network analysis. Wei et al., (2014) [13] [14] taking knowledge management as an example, built a 3-mode network including authors, keywords and periodicals, and reveal the inherent law and development trend of knowledge management by the empirical analysis and visualization. To address the excessive outliers problem involved in authors co-occurrence analysis, Teng (2015) [15] analyzed the co-occurrence of authors, institutions and countries, and built a mixed author-institution-country co-occurrence network based on the theory of supernetwork. His empirical research showed that the method could not only eliminate the isolated nodes in the authors co-occurrence network, but also enrich the amount of information in the network by adding the coupling relations between authors and institutions or countries. Pang (2012) [11] indicated that cross co-occurrence of 3-mode could reveal not only the knowledge of 3-mode but also the knowledge of 1-mode or 2-mode, and could therefore reveal deeper and broader knowledge than usual co-occurrence did.

2.2.2 Matrix fusion of multi-relational data

Braam et al. (1991) [16] combined co-citation cluster analysis with content words analysis to identify research topics. Based on co-citation clusters, they counted the frequency of words contained in references to test whether the cluster could gather literatures possessing similar terms into a class. Calero-Medina et al. (2008) [17] analyzed the knowledge creation and flow process between scientific publications by combining word co-occurrence and citation network analysis. They used word co-occurrence to find related terms and theories, and used citation network analysis to discover the key literatures in this field. Zhang (2007) [18] indicated that co-word clustering and strategic coordinates analysis were powerful tools to research discipline hotspots, and the combination of co-words and cited 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.

(1) Fusion between two types of relations

In the field of information retrieval, Weiss et al. (1996) [19] developed a prototype system of hierarchical network search engine—HyPursuit system, which was used for retrieval and browsing by detecting the content-link clustering of hypertext documents. The clustering algorithm of content-link was based on a literature similarity function which considered the similarity of the terms and hyperlinks similarity factor. The result of the function was the maximum value of similarities. Using the approaches based on reference links and co-words, Small (1998) [20] identified the direct and indirect connection relationship between literatures. Janssens et al. (2007, 2008)[21][22] learnt from the method of combining web content and hyperlink, and merged the relationship based on words and the relationship based on bibliographic coupling together. They used Fisher's inverse chi-square method for constructing new relational data sets, and conducted an empirical study, the result of which showed that the method was applicable to find the structure of research field on bioinformatics and information science. Moreover, Janssens et al. (2009) [23] also integrated cross-citation of journal with text mining, then validated and improved the existing classification scheme of topics.

(2) Fusion in the triple relations

Wang and Kitsuregawa (2002) [24] 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）[25] [26] proposed a web text clustering method merging the structure of text-based hyperlink, co-citation and text content. They used the structure of text-based hyperlink to calculate similarities, whose intensity was moderated by the text similarities, and then integrated both the hyperlink structure similarity and text similarity with co-citation by linear weighting to build a weighted adjacency matrix.

(3) Evaluation of relational fusion results

Compared with multi relationship, the evaluations of single-relationship clustering results differ in researches. The experiment conducted by Calado (2006, 2007, 2008)[21][22][27] showed that web page retrieval based on link relations was superior to text classification, but some other experiments showed that the similarity clustering based on words was superior to the similarity clustering based on citations. To sum up, all the experiments indicated the clustering results after the fusion was better than the one based on single-type relationships.

In the field of scientometrics, linear fusion is the mainly used algorithm in the relational fusion research. However, MSDF is complex, and the three main types of data relationships are often not independent but are correlated with each other, which is why a simple linear operation is not enough to solve the problem of data fusion. Still, we can learn the methods of MSDF from other research fields, such as sensor, automation and so on, to improve and enrich the MSDF methods in scientometrics analysis

3 Research and Application on relational Fusion

3.1 Cross- integration of multi-mode data

The cross-integration of multi-mode data is usually used to visualize the association among different data, which has no difficulties with visualization techniques. Currently, the module identification of multi-mode data is a research hotspot in complex network researches. In the case of 2-mode network, community detecting methods can be roughly divided into two types: the non-mapping method and mapping method. The mapping method is to convert 2-mode networks into 1-mode networks, which will lead to information loss and cannot reflect the nature of all original network. Latapy (2008)[28] summed up three drawbacks of mapping method: leading to information loss, increasing the number of edges of entire network and increasing unnecessary new information that does not exist in the original network.

To ensure the accuracy of the analysis process, the non-mapping method is more reliable, which identifies the module directly on the original 2-mode network. Guimerà et al. (2007)[29] and Barber et al. (2007)[30] respectively defined the modularity based on 2-mode network, and proposed the corresponding algorithm of community discovery. Both algorithms aim to maximize the modularity, but what's different is the way that they maximize. These have enriched the theory of community detecting based on 2-mode networks.

3.2 Matrix fusion of multi-relational data

Guo (2005) [31] summarized four main data fusion algorithm in sensor field.

(1) Probability theory

In probability theory, we can filter out these data with low confidence by analyzing compatibility of various sensor data. Then, according to the known prior probabilities, we can utilize Bayesian probability to estimate data with high confidence, thus obtain optimal fusion results. The merit of probability theory is being concise and easy to handle related events. The drawback is that the prior probability is not easy to get. What is more, if the prior probability is not consistent with the fact, the result of fusion will be further from the truth.

(2) Evidence theory

As an uncertainty reasoning method, evidence theory was first proposed by Dempster, and was founded and gradually developed by Shafer, which also known as the Dempster-Shafer evidence theory or D-S evidence theory. D-S evidence theory is an extension of probability theory that introduces belief function, basic assignment function and likelihood function to deal with uncertainties. Unlike probability theory, the 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 condition weaker than Bayesian probability does, distinguish unknown or uncertain information, perform fusion at different levels, and have a strong fault-tolerant ability.

However, when conflict occurs between evidences, D-S evidence theory may produce results inconsistent with the facts, thus become unusable. Another deficiency of this theory is that it is not easy to determine the basic assignment function and composition formula, in particular, the basic assignment function does not have unified methods.

(3) Fuzzy set approach

Fuzzy set approach is to use model to reflect the uncertainty in information fusion and use fuzzy reasoning to complete data fusion. Fuzzy clustering is a classification process of sample group, a process to merge characteristic parameters of sample and classify the sample. Fuzzy set approach has an advantage of logical inference. Compared with probability statistics, fuzzy set approach is closer to people's way of thinking, easier to overcome some of the problems that probability theory faces, and more suitable for fusion at a higher level. However, the method of logical inference is not mature and systematical enough, and is strongly influenced by subjective factors in describing and processing information.

(4) Neural networks

Neural network algorithm was proposed based on the multi-discipline specialists' study of the way in which animals process information. It has strong fault tolerance, hierarchy, self-learning, adaptability and parallel processing abilities. It can also simulate complex nonlinear functions. Currently, neural network has been successfully applied to the state evaluation of information fusion. Although the neural network has a strong nonlinear processing ability to be well used in the information fusion technology, it may lead to local minimization, slow convergence and sample-dependent and other issues.

In addition, meta-analysis is a systematic evaluation method. By an integration of multiple studies for the same purpose, meta-analysis uses quantitative methods to get final evaluation results. Therefore, meta-analysis is also a data fusion analysis. The biggest advantages of meta-analysis are the increasing of sample size, increasing accuracy of findings, and resolving inconsistency in the results of the individual studies. Han (2014) [32] systematically summarized the four methods of meta-analysis including methods based on P-value, rank, effect size and counting.

3.3Analysis of Clustering ensemble

(1) Clustering analysis

Clustering analysis derives from data mining and statistics, which is the core issue of knowledge discovery, machine learning, artificial intelligence, pattern recognition and other researches. Clustering technique divides data objects into several clusters, in which objects in the same cluster have high similarities and objects in different clusters have many differences.

Kogan (2006) [33] summarized the recent clustering algorithms suitable for different applications. According to the clustering rules and the method to apply these rules, clustering algorithms can be divided into the clustering algorithms respectively based on partitioning, hierarchy, density and grid. Important recent clustering algorithms include: the clustering algorithm based on computational intelligence [34][35], the semi-supervised clustering algorithm [36][37], the spectral clustering algorithm [38] and so on.

(2) Ensemble clustering

Both the cross- integration of multi-mode data and matrix fusion of multi-relational data eventually form a comprehensive matrix. When making clustering analysis on the comprehensive matrix, different clustering algorithms may lead to different clustering results. How to choose or integrate these various results so as to obtain more reasonable cluster needs the use of ensemble clustering. Ensemble clustering is to cluster different relational data respectively, and then merge different clustering results into one clustering result by using a fusion function. Before ensemble clustering, we need to assess these clustering results, and select appropriate clustering methods for ensemble clustering.

Ensemble clustering is an active research area of machine learning. The retrieved literatures have not shown a relevant research on ensemble clustering algorithms used in scientometrics methods to analysis scientific structure. Clustering fusion is to use different parameters to get a large number of cluster memberships, and then fuse the cluster memberships and obtain the final clustering results by using a fusion function [39].

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

(3) Clustering fusion function

The fusion function, also known as consensus function, is the key to ensemble clustering researches. Common fusion functions include Co-association matrix method[40] [41], Hypergraph-Partition method[39] [42] [43], and the methods respectively based on information theory[44], hybrid model[45], voting[46], accumulation of evidence[47] and neural network[48].

The differences among cluster memberships can significantly affect the final result of fusion. Therefore, the selection strategy based on clustering differentiation is usually to select the cluster memberships with large differentiation as the members participating in fusion, referred as fusion members. The key to this method is to measure the degree of difference between the cluster members. The common measurement methods include normalized mutualinformation (NMI) [39][46], rand index(RI) [49], Jaccard index(JI) [50], adjusted randindex(ARI) [51], clustering error(CE) [52],

variation of information(VI) [53] and so on. The methods to measure the quality of cluster memberships include F-measure, RI, JI, Cophenetic correlation coefficient and CDbw (composing density between and within clusters) index, etc. [54]

4 future method of MSDF in scientometrics

The improvement of MSDF methods requires a solid mathematical foundation. The development or breakthrough of MSDF may occur in sensor, automation or other fields in the future. As the research methodology being application-oriented, scientometrics should learn the existing methods of MSDF from mathematical field, and form its own methodology of MSDF by taking into account its own characteristics.

This paper assumes a future research model of MSDF in scientometrics. As shown in Fig.1，the method can realize the fusion among different kinds of information, data relations and clustering results.

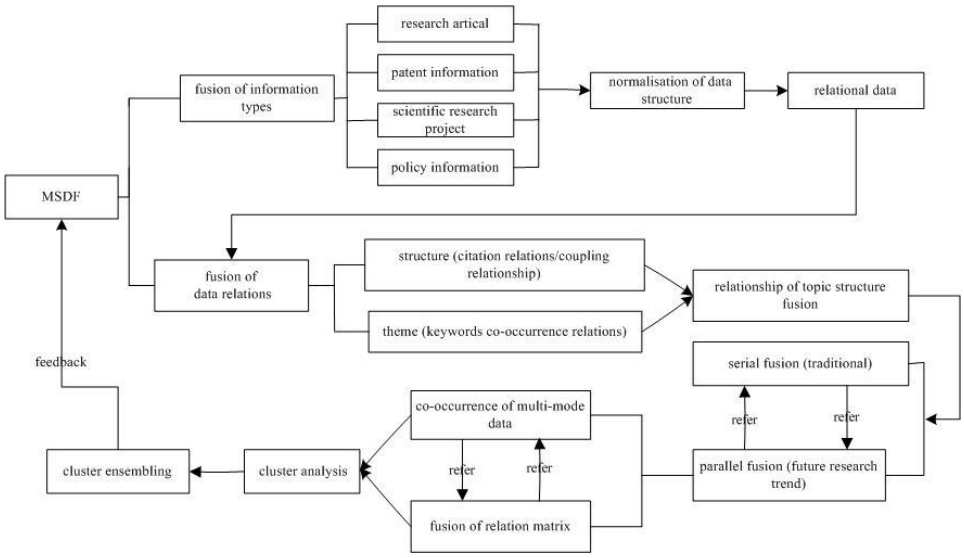

Fig.1 Model of MSDF in topic discovery

Firstly, collect a variety of data sources, such as journal articles, conference information, dissertations, patent information, project information, book information, and even industrial and economic data should be included in the scope of scientometrics analysis. Different data types reflect different contents of technology. For example, scientific papers focus on basic scientific research outputs, patent information focuses on technological innovations, and data of industrial economics is to grasp the technology market information. Therefore, only taking multi-source information into consideration comprehensively, can we get a more objective result from scientometrics analysis.

Secondly, obtain various data relations and fuse them effectively. Whether cross-integration of multi-mode data or matrix fusion of multi-relational data has its corresponding fusion methods, which complement each other, based on the data features.

There are basically two modes for multi-source data fusion, one is to obtain a variety of associated data types respectively, and fuse relation matrixes of different data types by mapping; the other is to directly identify the community topics of multi-mode data. Both methods can enhance the data relationship strength by acquiring complementary information.

a. The relation matrix integration includes obtaining various types of data relationship, which relate to the analysis target, projecting the variety of relational data to the target relational and generating the target analysis matrix through matrix fusion operation (Fig.2).

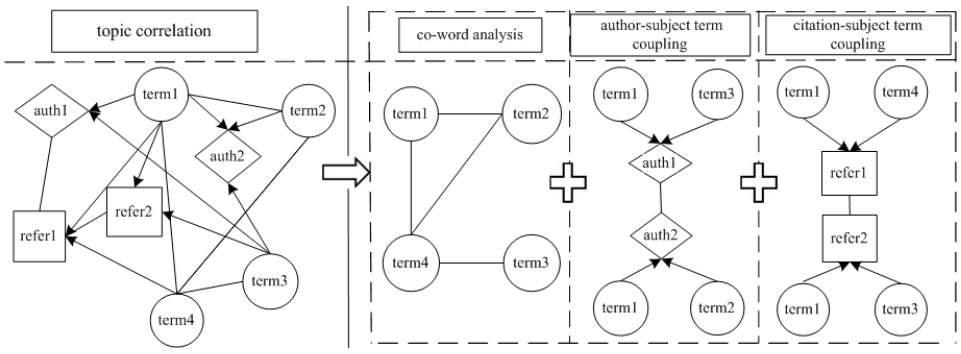

Fig.2 Various types of data relation matrix fusion

b. Cross-integration of multi-mode data can directly identify the relation communities of multi-mode network can be performed, which can save more relationships of variables and avoid data distortion during data projection (Fig.3).

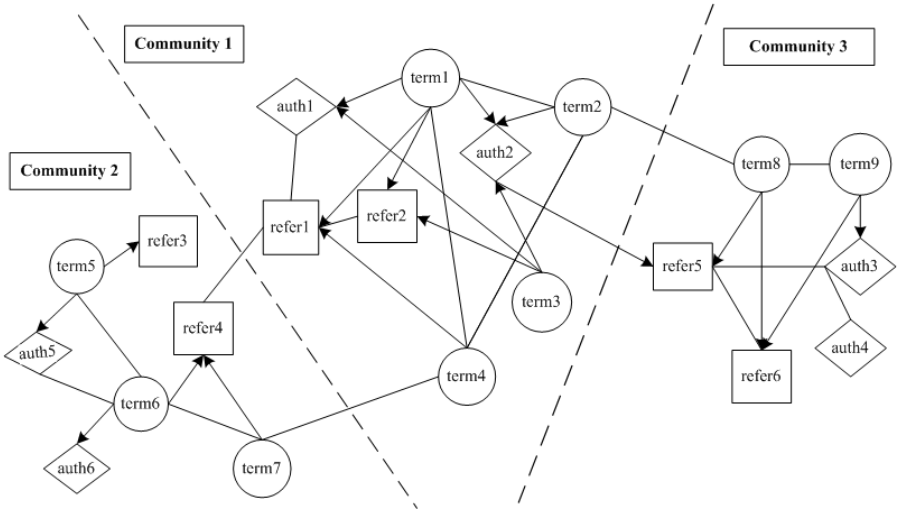

Fig.3 Community identification of the cross-integration multi-mode data

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-ocurrence of multi-mode data still has much room for improvement, for example, visualizing information and association of more dimensions, which is important for knowledge discovery.

Matrix fusion of multi-relational data can bring in existing fusion methods from the field of sensor，automation and so on. According to the objects and characteristics of scientometrics, we can improve these methods and eventually formed fusion methods applied to scientometrics. The matrix fusion has different ways of fusion, which derive from different rationales, strengths and weaknesses, so they can complement each other to enhance the effectiveness of matrix fusions. For example, the method of evidence theory is simple and effective, but typical D-S theory is sensitive to highly conflict evidences. Neural network algorithm is relatively complex, but has a strong fault tolerance, hierarchy, self-learning, self-adaptation and parallel processing capabilities. So, naturally, we can combine

evidence theory with neural network to achieve the fusion of relational data according to the data characteristics and the complexity of the analysis problem.

Finally, with the development of outstanding clustering analysis methods, it is crucial to choose the most effective clustering algorithm and consensus function. The algorithm of clustering ensemble can be introduced to merge a variety of clustering results into a final desired one.

5 Conclusion

By investigating the main thoughts of MSDF, this review proposes that MSDF can be divided into the fusion of data type and fusion of data relation in Scientometrics. The fusion of data type can be subdivided into the cross-integration of multi-mode data and matrix fusion of multi-relational data. The clustering results of relation matrix can be optimized by the clustering ensemble. In addition, we assume that the improvement of MSDF methods requires a strong mathematical foundation, and breakthrough of MSDF in future may come from advanced fields in data fusion research and their applications, such as sensors, the automation, etc. Scientometrics should draw on the experience of these advanced fields and build its own analytic methodology of MSDF in the future.

Acknowledgments

This work was supported by National Social Science Fund of China (Grant No. 14CTQ033), West Light Fund of Chinese Academy of Science (Grant No: Y4C0091001), and Youth Innovation Fund of Promotion Association, CAS.