Study of Multi-source Data Fusion in Topic Discovery

Hai Yun Xu, Chao Wang, Li Jie Ru, Zeng Hui Yue, Ling Wei and Shu Fang

Abstract This review provides an introduction to MSDF in topic discovery, and discusses the status quo of the methods and applications of MSDF. This review has investigated the main thoughts of MSDF 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.

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

1 Introduction

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

H.Y. Xu · C. Wang · L.J. Ru · L. Wei · S. Fang
Chengdu Library of Chinese Academy of Sciences,
Chengdu, Sichuan 610041, People's Republic of China

H.Y. Xu
Institute of Scientific and Technical Information of China,
Beijing 100038, People's Republic of China

C. Wang (✉) · L.J. Ru · L. Wei
University of Chinese Academy of Science, Beijing 100190,
People's Republic of China
e-mail: wngchao2015@mail.las.ac.cn

Z.H. Yue
Information Engineering, Faculty of Jining Medical University,
Rizhao, Shandong 276826, People's Republic of China

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 393,
DOI 10.1007/978-981-10-1536-6_95

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 topic discovery, we learn from the methods of MSDF in sensors or some other fields, and innovate these methods to be suitable for topic discovery. 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 topic discovery analysis methods.

2 Overview of MSDF in Topic Discovery

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 [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 topic discovery 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. [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 [3] discovered the clustering network built by direct citation, comparing with co-citation, it had more similar content.

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. Similarly, the network of relationships among three collections is called a 3-mode crossing 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.

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.

2.2.2 Matrix Fusion of Multi-Relational Data

Braam et al. [4] 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. [5] 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 [6] 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. [7] 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 [8] identified the direct and indirect connection relationship between literatures. Janssens et al. [9] 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.

2. Fusion in the triple relations

Wang and Kitsuregawa [10] 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. [11] 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 [12] 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 topic discovery, 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 topic discovery analysis

3 Research and Application on Relational Fusion

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 [13] 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. [14] 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.

4 Future Method of MSDF in Topic Discovery

This paper assumes a future research model of MSDF in topic discovery. The method can realize the fusion among different kinds of information, data relations and clustering results.

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 topic discovery 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.

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 topic discovery 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 topic discovery, we can improve these methods and eventually formed fusion methods applied to topic discovery. 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.

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 topic discovery. 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. Topic Discovery 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.

References

1. Hua BL (2013) Research on the methods of multi-source fusion. Inf Stud Theory Appl 36 (11):16–19
2. Shibata N, Kajikawa Y, Takeda Y, Matsushima K (2009) Comparative study on methods of detecting research fronts using different types of citation. J Am Soc Inf Sci Technol 60(3):571–580

3. Klavans R, Boyack KW (2006) Quantitative evaluation of large maps of science. Scientometrics 68(3):475–499
4. Braam RR, Moed HF, Van Raan AF (1991) Mapping of science by combined co-citation and word analysis. I Struct Aspects JASIS 42(4):233–251
5. Calero-Medina C, Noyons EC (2008) Combining mapping and citation network analysis for a better understanding of the scientific development: the case of the absorptive capacity field. J Informetrics 2(4):272–279
6. Zhang H, Wang XY, Cui L (2007) Trend: Co—word analysis method combined with literature CI research thematic areas. Inf Stud Theory Appl 30(3):378–380
7. Weiss R, Vélez B, Sheldon MA (1996) HyPursuit: a hierarchical network search engine that exploits content-link hypertext clustering. In: 7th ACM conference on hypertext, pp 180–193
8. 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
9. Janssens F, Glänzel W, De Moor B (2008) A hybrid mapping of information science. Scientometrics 75(3):607–631
10. Wang Y, Kitsuregawa M (2002) Evaluating contents-link coupled web page clustering for web search results. In: 11th international conference on Information and knowledge management, pp 499–506
11. He X, Zha H, Ding CH, Simon HD (2002) Web document clustering using hyperlink structures. Comput Stat Data Anal 41(1):19–45
12. Calado P et al (2006) Link-based similarity measures for the classification of Web documents. J Am Soc Inf Sci Technol 57(2):208–221
13. Latapy M, Magnien C, Del Vecchio N (2008) Basic notions for the analysis of large two-mode networks. Soc Netw 30(1):31–48
14. Guimerà R, Sales-Pardo M, Amaral LAN (2007) Module identification in bipartite and directed networks. Phys Rev E 76(3):036102