PROCEEDINGS OF SCIENCE

Research of Knowledge Mapping Construction Method Based on Scientific Research Results

Yiqiong Zhang[1]
School of Software, Beihang university
Beijing, 100191, China
E-mail: zhangyq11@buaa.edu.cn

Guangyan Lin[23]
School of Software, Beihang university
Beijing, 100191, China
E-mail: lingy@buaa.edu.cn

Ji Li
School of Software, Beihang university
Beijing, 100191, China
E-mail: keeley@buaa.edu.cn

Based on the fact that domestic and overseas knowledge mapping researches commonly used literature data as sample, this paper puts forward a new approach of knowledge mapping construction to supplement the traditional way. This construction approach uses scientific research results such as research projects, teams and awards which may reflect technological development as sample data, and add data cleaning, knowledge cell selecting and information mining methods to manage the research results data in disorder and without unified collection. In this paper, the computer science discipline is chosen as an example to illustrate feasibility, efficacy and innovation of this approach. The empirical result shows that the knowledge mapping constructed by using this approach gives a more comprehensive analysis of discipline development. This knowledge mapping construction approach provides a new idea of analyzing, supervising and evaluating discipline development from the perspective of multi-dimensional data analysis, also reveals more discipline development situation than the approach based on traditional literature data.

CENet2017
22-23 July 2017
Shanghai, China

[1] Speaker
[2] Corresponding Author
[3]The study is supported by Scientific research funds of Beihang University (YWF-16-RJXY-001)

PoS(CENet2017)069

1.Introduction

With the improvement of scientific researches in our country, the scientific research projects and achievements in various fields have witnessed an explosive growth. The research results of various disciplines have been flourishing and the research hotspot has shown the characteristics of omni-directional and cross. It is necessary to review, summarize, evaluate and rethink the achievements of scientific researches and learn the current situation of scientific researches, research hotspots and development trend.

The knowledge mapping is a kind of measurement method which uses scientific knowledge as the object of study, applies data mining, information processing, knowledge measurement and data visualization methods to show the development process and structure of scientific knowledge, reveal the regularity, show the knowledge structure relationship and the law of evolution [1]. There has been a commonly used approach which selects data from literature database and inputs the literature data into software such as CiteSpace to extract knowledge cell and visualize data automatically. Some researchers in library and information science use literature data from WOS and CNKI to analyze the development of specific fields including economics [2] and evidence-based medicine etc.[3]; however, this software approach still features some limitations because there are many other scientific results such as project and awards not included in literature, the knowledge mapping as constructed by using literature data is incomprehensive and unconvincing.

To solve the above problems, this paper provides a solution of knowledge mapping constructing which uses various kinds of scientific results including research projects, awards and teams as the sample data to supplement the literature knowledge mapping.

2.Methods

Scientific results data such as projects, awards and teams don't have unified collection and complete attributes as the literature data do. Our new knowledge mapping constructing approach adds the process of data pre-processing and information extracting for the semi-structured and unstructured data, and consider the traditional data analysis process as the eample. The overall flow chart was shown in Figure 1. In this chapter, we mainly discuss the process of data cleaning and information extraction, and briefly introduce the data analysis and visualization methods.

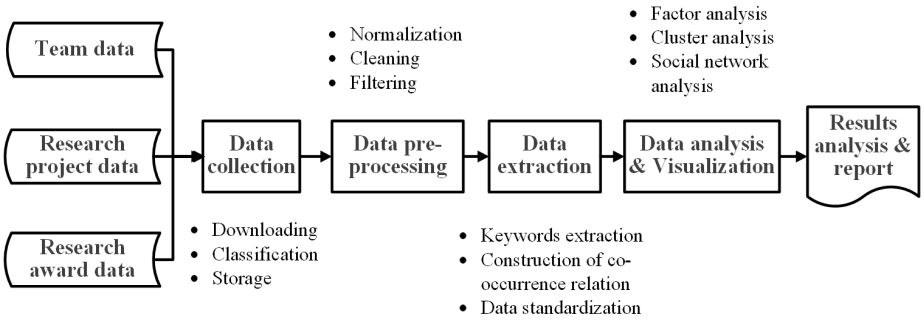

Figure 1 : Flow Chart of Knowledge Mapping Construction

2.1Data Collection

After comparing various scientific results from websites retrieval, we choose LetPub, NAST and NOSTA as our data sources where the data are more complete and well classified [4,5,6]. Upon classification retrieval, we can get the scientific results of a specific discipline. These data collected from various resources feature various formats and attributes and one result may also be found more than one time. We retain the unified attributes including *Type*, *Name*, *Year*, *Serial Number* and *Owner*, and save the data in our self-built database.

2.2Data Cleaning

Although the data have been unified in terms of the collection process, the duplicated data may still reduce the accuracy of information extraction. As a result, we add the data cleaning process which removes the duplicated data to improve the quality of sample data.

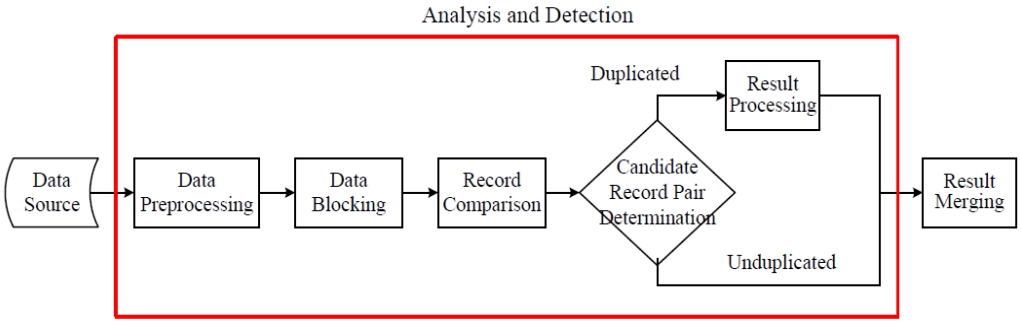

Figure 2 : Flow Chart of Data Deduplication

In Figure 2, we chose the attribute Name of data as the unique identifier and design a short text deduplication method. The whole flow includes four steps. Firstly, we removed spaces, punctuation and other special characters to eliminate interference; secondly, we divided the data records into blocks according to other attributes such as *Type* or *Year,* which can narrow down the scope and reduce the time complexity of similarity comparison. After partitioning, we compared all the records in the same block in pairs and calculated the similarity of each records pair. Finally, the records pair whose similarity value was higher than the empirical value we measured would be marked as the duplicated. We will merge the duplicated results and remove the redundant records.

In this deduplication method, the key point is the similarity comparison. In this step, we made word segmentation of name and then used the TF-IDF formula to calculate the representative value of record. The formula is shown in Equation (2.1).

$$Seg = \sum tf \cdot idf = \sum_{i=1}^{k} m_{term} \cdot \log \frac{n}{m_{term}} \tag{2.1}$$

In this Equation, the parameter *term* is the segmentation of a record, *n* is the total number of records, *mterm* is the total number of times the word *term* appeared, *k* is the number of words a record disparted, and *Seg* is the sum TF-IDF value of all words one record segmented into and is regarded as the representative value of this record. After calculating the result *Seg1*, *Seg2* and the measurement of intersection in the record pair *Seginter*, we choose the bigger one in *Seg1* and *Seg2* as the denominator *Segtarge*. The similarity of a record pair is

$$\frac{Seg_{inter}}{Seg_{large}} \in [0,1] \tag{2.2}$$

2.3Information Extraction

The kknowledge cell is the minimum unit in knowledge mapping, based on which the relationship of records and analysis is established As the scientific data don't have keywords or quotations as literature data do, we choose the name field as the text material to extract keywords and their co-occurrence relationship in text. The steps are described as follows.

1) Add specialized vocabulary dictionary of the discipline we analyzed and the stop-words dictionary to improve the granularity of segementation and the quality of words.
2) Use NLPIR segementation interface to segment the name field of records.
3) Calculate the information entropy of each word, select the top 50 -100, then merge the words that have same meaning and thus get the final keywords.
4) Summaarize the co-occurrence time of keywords in pairs and get the keywords co-occurrence matrix which represents the relationship of hotspots.

In this Section, we mainly introduce the keyword selecting method. According to the Zipf's Law, the most frequent words are not the most important. The selection method of high-frequency words is not suitable for short text. We take both the word frequency and universality into account to calculate the information entropy of each word. The information entropy of word *i* is

$$H_i = TF \cdot IDF = m_i \cdot \log \frac{n}{n_i + 1} \tag{2.3}$$

In this Equation, m_i is the number that the word *i* appears, and *n* is the total essay number of NLPIR corpus, and the n_i is the number of essay in corpus which contains the word *i*. And $\log \frac{n}{n_i + 1}$ represents the universality of a word. The word with higer value may feature higer frequency and is less universal.

2.4Data Analysis and Visualization

As the keywords and co-occurrence matrix can only reflect the hotspots of research, we need to mine the relationship between keywords, and further explain the relevance of hotspots and the aggregation of scientific research by data analysis and visualization. We usually use factor analysis, cluster analysis and social network analysis methods. By factor analysis and cluster analysis, we can get the scree plot and cluster dendrogram which can tell us how many fields in this discipline and which field the keywords belong to [7]. By social network analysis, we can get the keywords co-occurrence network graph and their centrality which can show the degree of relevance and the edge or centrality of keywords [8]. The analysis methods and relative software have been widely used and won't be explain in this paper.

3.Results and Analysis

To illustrate the methods and verify the feasibility of our approach, we take computer science discipline as a example to construct the knowledge mapping by using scientific results

as sample data and briefly analyze the construction results.

Firstly, we select the scientific research project, team and award data of computer science during the period of 2011-2016 from web and after data compilation we get 24446 relevant records.

In the data cleaning process, we set the standard similarity value as 0.7, and calculate the similarity of each records pair. After de-duplication of the high similarity records, we get 23709 valid and unique records.

For the information extraction, we selected more than 20000 computer science vocabularies as the user dictionary and set 20 stop-words including Based on and Methods. By using the NLPIR segmentation interface and self-made program to calculate the information entropy of each words, we selected 55 words with high entropy. The results are shown in Table 1. And we summarize the co-occurrence matrix.

No	Keywords	Frequency	No	Keywords	Frequency	No	Keywords	Frequency
1	sensor network	885	20	super-resolution	75	39	context awareness	45
2	cloud computing	561	21	manifold	108	40	dimensionality reduction	48
3	isomerism	450	22	MT	102	41	visual attention	51
4	system structure	303	23	deep learning	99	42	software defined network	39
5	internet of things	300	24	secret key	90	43	fractional order	48
6	privacy protection	243	25	remote sensing image	75	44	smart grid	39
7	modality	210	26	transfer learning	72	45	partial differential equation	36
8	the nuclear	141	27	medical image	69	46	pervasive computing	33
9	embedded system	180	28	unexpected events	63	47	human-computer interaction	33
10	social networks	180	29	supply chain	57	48	reinforcement learning	33
11	social network	168	30	access control	60	49	congestion control	30
12	semi supervised	111	31	data collection	57	50	intrusion detection	30
13	machine learning	150	32	workflow	60	51	augmented reality	30
14	robust	147	33	rough set	54	52	bayes	54
15	neural network	132	34	support vector machines	48	53	information hiding	24
16	feature extraction	123	35	electronic commerce	48	54	peer-to-peer network	30
17	power waste	120	36	registration	51	55	deficiency and excess fusion	18
18	target tracking	117	37	protein-protein interaction	48			
19	compressed sensing	117	38	public key cryptography	45			

Table 1: The Keywords and Frequency

In Table1, the words *sensor network*, *cloud computing* and *heterogeneous network* etc. with high-frequency reflect the focus of several research fields in the computer science; the words *semi supervised*, *machine learning* and *neural network* are in the middle of the table, present the research fields which are rising gradually in recent years; the words *software defined network, augmented reality* and *virtuality and reality combination* in the end of the table represent the newest research field in computer science. The results show that the keywords extraction mostly cover the representative research field in computer science.

In the data analysis and visualization process, we used the cosine coefficient based correlation matrix and the principal component method of factor analysis to get the factors scree plot and cumulative variance contribution rate. In Figure 3, by using 9 factors, more than 90% of the information of all keywords will be described, indicating that it is appropriate if we classify the computer science research hotspots into 9 fields [9]. Based on the correlation matrix, we use Q cluster analysis method which is more suitable for sample cluster to cluster the keywords. The clustering result is shown in Figure 4.

According to the factor and the cluster analysis results, all the keywords can be classified into nine groups such as video analysis including *target tracking* and *compressed sensing*, computer architecture including *flushbonading* and *many-core system* etc.

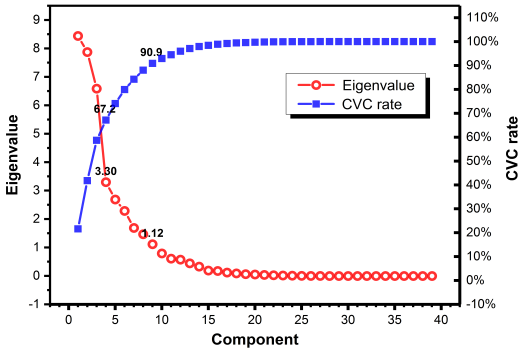

Figure 3: Scree Plot and CVC Rate

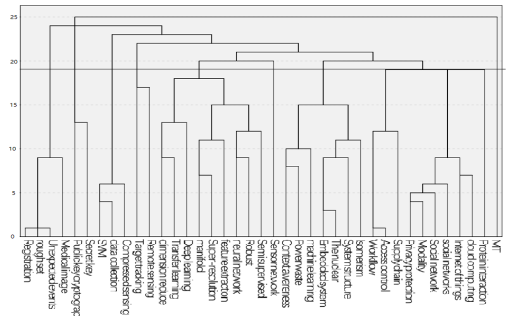

Figure 4: The Cluster Dendrogram

Figure 5 presents the co-occurrence network based on the original keyword co-occurrence matrix by using the software Gephi. As we can see in the network, the words *sensor network*, *cloud computing* and *heterogeneous* are in the core position, which reflects that these words are the main technologies and research fields. The words *e-commerce*, *information hiding* and *congestion control* are at the edge of network, indicating that they are not closely contacted with other keywords and can be judged as independent application areas or new research problems. The words *modal*, *machine learning* and *robust* which, in the border region of clusters, are the link of other research fields and may exert greater influence on the development of the whole network.

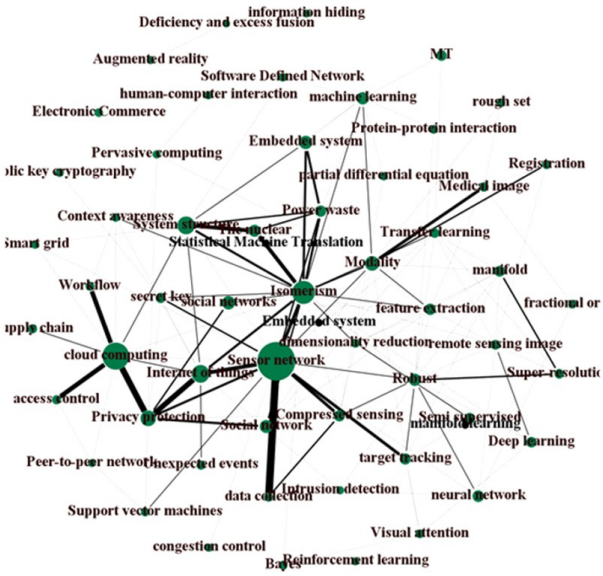

Figure 5: The Keywords Co-occurrence Network

4.Conclusion

Although the knowledge mapping is a new concept, it is a practical technique based on the research methods of many fields. Our solution breaks the limit of sample data in the traditional knowledge mapping construction caused by the software and provides a data processing approach in combination of the latest technology including natural language processing, data mining and data visualizing. It will greatly supplement the completeness and veracity of discipline knowledge mapping.

The application of knowledge mapping in the field of scientific research has shown its charm and prospect. With the improvement of science and technology, the scientific knowledge

mapping--the method of data measurement and information mining, will be widely adopted in more fields.

References

[1] Liang Xiujuan. *Review of Mapping knowledge Domains*[J]. Library Journal, 2009(6):58-62.

[2] Jiang Chunlin, Du Weibin, Li Jiangbo. *Economy Papers Map of Co-occurrence Analysis Based on CSSCI*[J]. Journal of Information, 2008, 27(9):78-80.

[3] Shen Jiantong, Yao Leye. *A Case Study of Applying Multiple Analysis and Social Network in Mapping Knowledge Domain*[J]. Journal of Intelligence, 2009, 28(8):33-36.

[4] The LetPub Website of ACCDON[DB/OL]. http://www.letpub.com.cn/index.php?page=grant

[5] The National Achievements System of Science and Technology[DB/OL]. http://www.tech110.net/portal.php

[6] National Office for Science & Technology Awards[DB/OL]. http://www.nosta.gov.cn

[7] Zhou Lei, Yang Wei, Zhang Yufeng. *Issues and Re-consideration on Cluster Analysis in Co-occurrence Matrix*[J]. Journal of Intelligence, 2014(6):32-36.

[8] Börner K, Chen C, Boyack K W. *Visualizing knowledge domains*[J]. Annual Review of Information Science & Technology, 2005, 37(1):179-255.

[9] Feng Guohe, Liang Xiaoting. *Analysis on Knowledge Mapping of Academic Research on Recommender Engine in China*[J].Information Science, 2012(1):146-150+162.