

Analysis of Hotspots in the Field of Domestic Knowledge Discovery Based on Co-Word Analysis Method

X. D. Wang¹, J. J. Liu¹, F. S. Sheng²

¹School of Economics and Management, Beijing Jiaotong University, China

²Tianjin Campus of Naval Engineering University, China

Emails: xdwang@bjtu.edu.cn 13120626g@bjtu.edu.cn sfs@163.com

Abstract: In this paper, choosing highly frequent keywords from core journals in the field of 1992-2013 national knowledge discovery in CNKI database, counting the number of two frequent keywords co-occurrences in the same journal, then constructing the highly frequent keywords matrix, and transforming the highly frequent keywords matrix into a correlation matrix and a dissimilarity matrix, we analyze the dissimilarity matrix based on the use of factor analysis, cluster analysis. After discussing the results of the analysis, we found that the current hotspots in the field of domestic knowledge discovery have focused on the following six aspects, knowledge discovery based on data research, knowledge discovery algorithm optimization research, the model of knowledge discovery and references research, knowledge management based on domain ontology, expert system construction research, and applied research of the knowledge discovery. Finally, we summarized the research hotspots in the field of international knowledge discovery in the same way and suggested the domestic scholars to extend some directions of the research in the field of knowledge discovery.

Keywords: Knowledge discovery, co-words analysis, cluster analysis, factor analysis.

1. Introduction

In the last ten years, with the rapid development of modern information technology, such as computer and Internet, knowledge discovery is always a hot topic of the academic research and industrial application. Fayyad et al. [1] defined knowledge discovery as a nontrivial process of distinguishing valid, novel,

potentially useful and ultimately understandable patterns from the data set. Nowadays the field of knowledge discovery has basically two branches – Knowledge Discovery in a Database (KDD) and knowledge discovery based on literature. The difference in the two branches is that KDD is for structured data, while knowledge discovery based on literature is mainly for unstructured data [2].

2. Research method

We used co-word analysis method which is a content analysis method and an extension of the word frequency analysis method. Firstly we must count the frequency of the keywords which can express the core content of the literature, and then select the highly frequency keywords. At last, we should discuss the relationship between the high frequency keywords in this field by analyzing the highly frequency keywords which appear together. Therefore, the co-word analysis method can present a research hotspot in the field [3].

There is great difference between the co-word analysis and the co-citation analysis. Since the paper is cited with a certain time lag from the time of its publication, the co-citation analysis is more suitable for a mature and specified field, while the co-word analysis is more suitable for un-methodical and current field. Because there are a lot of participants in the emerging field of research, the focus of the paper is more dispersed, and the cited situation is not stable, as a result, a relatively fixed academic community has not been formed. Under these circumstances, the subject headings and keywords can better represent the hotspots and directions in the new field [4]. Therefore, the co-word analysis is more suitable for discussing the research hotspots of “knowledge discovery”.

The application of the co-word analysis follows four steps: firstly, filtering the highly frequency keywords in the field of knowledge discovery. Secondly, constructing a co-word matrix and a correlation matrix, if there are too much zeroes in the correlation matrix, it is needed to transform the correlation matrix into a dissimilarity matrix. Thirdly, choose multivariate statistical methods to analyze the correlation matrix or dissimilarity matrix. Fourthly analyze the statistical results by multivariate statistical methods [5].

3. Data acquiring and processing

3.1. Data acquiring

We retrieved the paper in CNKI which satisfies three conditions:

- 1) the name of the paper equals to “knowledge discovery”;
- 2) deadline equals to 2013;
- 3) the source journals equal to SCI source journals, EI source journals, core journals and CSSI (these four periodical indices have a certain influence, so the papers retrieved from them can reflect the research focus in the field of domestic discovery to a great extent).

By retrieving, we can get 372 papers whose “title” includes “knowledge

discovery”. After an excluding meeting notice, repeated papers and a paper without keywords, we can get keywords with the help of “CNKI literature derived”. Some keywords like “knowledge discovery in database” and “KDD” may cause a statistical error due to different expressions, thus we need to unify the expressions in order to eliminate this kind of influence. Counting the frequency of the keywords by processing the keywords with the help of EXCEL PivotTable, then selecting the keywords which frequency are not less than 5, the keywords-“knowledge discovery” must be removed because of their too high frequency, too thick lines and less practical significance in research. In this way we can finalize 24 high frequency keywords in Table 1.

Table 1. The high frequency keywords list

No	Keyword	Word frequency
1	Data mining	82
2	Rough set	42
3	KDD	28
4	Model	28
5	Database	25
6	Non-interactive literature	23
7	Association rules	19
8	Clustering	17
9	Ontology	11
10	Decision tree	11
11	Expert system	11
12	Information system	11
13	Machine learning	10
14	Knowledge discovery system	10
15	Knowledge management	9
16	Double-base cooperating mechanism	8
17	Knowledge base	8
18	Digital library	7
19	Data warehouse	7
20	Attribute reduction	6
21	Genetic algorithm	6
22	Domain knowledge	6
23	CRM	5
24	Knowledge mining	5

Although the keywords above given can reflect the hotspots in the field of knowledge discovery to some degree, almost all the papers have more than one keyword. So it is unsuitable to reflect the research hotspots in the field of knowledge discovery only by one keyword, however we can have a further study about the relationship between these high frequency keywords by constructing a co-word matrix and a correlation matrix (if there are too many zeroes in the correlation matrix, we need to convert it to a dissimilarity matrix) of the high frequency words.

3.2. Constructing of the matrix

3.2.1. Constructing a co-word matrix

The numbers along the main diagonal of the co-word matrix are the frequencies of the keywords, the numbers along the non-main diagonal shows the frequencies of the two different keywords appearing together. We counted the frequency of the two different keywords appearing together with the help of EXCEL; in this way a 24×24 co-word matrix can be constructed (Table 2).

Table 2. A partial co-word matrix

$i \backslash j$	1	2	3	4	5	6	7	8	9	10
1	82	5	9	8	7	0	5	3	0	4
2	5	42	4	3	2	0	0	2	0	2
3	9	4	28	3	4	0	3	1	0	1
4	8	3	3	28	0	2	0	0	2	2
5	7	2	4	0	25	0	1	2	0	1
6	0	0	0	2	0	23	0	0	0	0
7	5	0	3	0	1	0	19	3	0	0
8	3	2	1	0	2	0	3	17	0	0
9	0	0	0	2	0	0	0	0	11	0
10	4	2	1	2	1	0	0	0	0	11

3.2.2. Constructing a correlation matrix

In order to meet the needs of cluster analysis and factor analysis, data processing must be based on a co-word matrix of the keywords, so we constructed a correlation matrix using the correlation coefficient ochiai [6].

Ochiai correlation coefficient is

$$E_{ij} = \frac{C_{ij}}{C_i C_j^{\frac{1}{2}}}, \quad i, j = 1, 2, \dots, 24,$$

where: the letters i and j represent the number of the keywords; E_{ij} is between 0 and 1, which represents the value of the correlation coefficient; C_{ij} represents the co-occurrence frequency of the keywords i and j ; C_i represents the number of appearing of the keyword i ; C_j represents the number of appearing of the keyword j .

In the correlation matrix, the value of the correlation coefficient represents the degree of correlation between two keywords. The larger the value is, the stronger the degree of correlation between the two keywords is; while the smaller the value, the weaker the degree of correlation between the two keywords is.

Table 3. A partial correlation matrix

$i \backslash j$	1	2	3	4	5	6	7	8	9	10
1	1	0.0852	0.187826	0.166957	0.154604	0	0.126674	0.080351	0	0.133185
2	0.0852	1	0.116642	0.087482	0.061721	0	0	0.074848	0	0.093048
3	0.187826	0.116642	1	0.107143	0.151186	0	0.130066	0.045835	0	0.05698
4	0.166957	0.087482	0.107143	1	0	0.078811	0	0	0.113961	0.113961
5	0.154604	0.061721	0.151186	0	1	0	0.045883	0.097014	0	0.060302
6	0	0	0	0.078811	0	1	0	0	0	0
7	0.126674	0	0.130066	0	0.045883	0	1	0.166924	0	0
8	0.080351	0.074848	0.045835	0	0.097014	0	0.166924	1	0	0
9	0	0	0	0.113961	0	0	0	0	1	0
10	0.133185	0.093048	0.05698	0.113961	0.060302	0	0	0	0	1

3.2.3. Constructing a dissimilarity matrix

As we can see in the correlation matrix in Table 3, there are too many zeroes in it, which may cause an error when we analyze. Thus we substituted $1 - E_{ij}$ for E_{ij} , then we can get a dissimilarity matrix in chart 4. Contrary to the correlation matrix, the larger the value is, the weaker the degree of correlation between the two keywords is, while the smaller the value, the stronger the degree of correlation between the two keywords in the dissimilarity matrix is.

Table 4. A partial dissimilarity matrix

$i \backslash j$	1	2	3	4	5	6	7	8	9	10
1	0	0.9148	0.812174	0.833043	0.845396	1	0.873326	0.919649	1	0.866815
2	0.9148	0	0.883358	0.912518	0.938279	1	1	0.925152	1	0.906952
3	0.812174	0.883358	0	0.892857	0.848814	1	0.869934	0.954165	1	0.94302
4	0.833043	0.912518	0.892857	0	1	0.921189	1	1	0.886039	0.886039
5	0.845396	0.938279	0.848814	1	0	1	0.954117	0.902986	1	0.939698
6	1	1	1	0.921189	1	0	1	1	1	1
7	0.873326	1	0.869934	1	0.954117	1	0	0.833076	1	1
8	0.919649	0.925152	0.954165	1	0.902986	1	0.833076	0	1	1
9	1	1	1	0.886039	1	1	1	1	0	1
10	0.866815	0.906952	0.94302	0.886039	0.939698	1	1	1	1	0

3.3. Cluster analysis

Cluster analysis or clustering is the task of grouping a set of objects in such a way that the objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). We can associate highly frequency keywords which are closely related together by cluster analysis, in order to form a various class, and then show the structure of the hotspots in the field of knowledge discovery.

By importing the dissimilarity matrix into SPSS19.0 to cluster analysis and choosing a hierarchical cluster, a square sum of deviations and Euclidean square distance, we finally can get a tree diagram. Based on the tree diagram in Fig. 1, we can initially determine the degree of association among the highly frequency keywords.

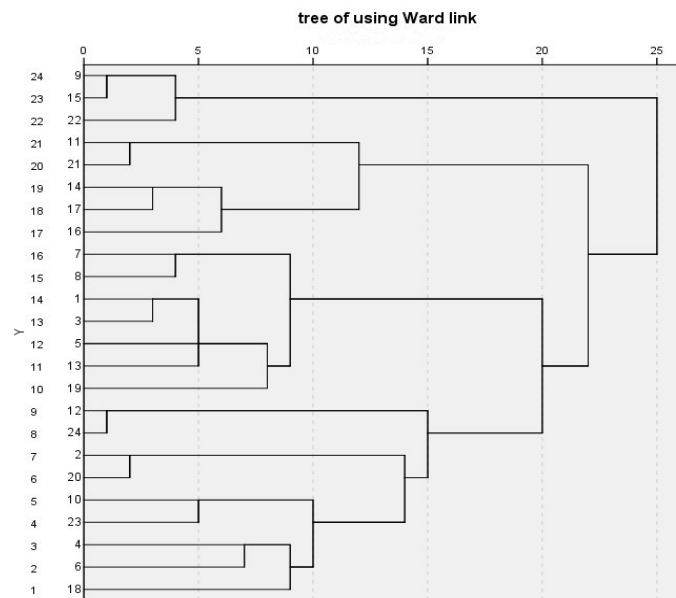


Fig. 1. Cluster analysis-tree

As we can see from Fig. 1, the clustering result is not very good, but it is obvious that the following seven groups of highly frequency keywords have stronger correlations:

- 1) ontology, knowledge management, domain knowledge;
- 2) expert system, genetic algorithm;
- 3) knowledge discovery system, knowledge base;
- 4) association rules, clustering;
- 5) data mining, KDD;
- 6) information system, knowledge mining;
- 7) rough set, attribute reduction.

3.3. Factor analysis

Factor analysis is used to explore the covariance structure among a set of observed random variables. The technique constructs a reduced number of random variables called common factors. Its basic idea is grouping the highly frequency keywords according to their correlation, which makes the high correlation in the same group and a low correlation in a different group. Each group represents a basic structure which is called a common factor, by means of the factor analysis method we can transform several keywords into a keywords group, and present the result of cluster analysis [7].

Then by importing the dissimilarity matrix into SPSS19.0, analyze the factor by means of the principal components. In Table 5, 11 common factors were extracted from 24 keywords, they can explain 70.679% of the keywords. Besides, according to the theory of data mining the extracted common factor must generally be more than 60% of the information, as we can see in Table 5, 9 common factors extracted generally 61.563% of the keywords, which shows that 24 keywords can be subordinate to 9 or 11 categories.

Table 5. Factor analysis, explaining the total variance of the table

Component	Initial eigenvalues			Extraction of sum of squares loaded		
	Total	The variance (%)	Accumulation (%)	Total	The variance (%)	Accumulation (%)
1	2.435	10.144	10.144	2.435	10.144	10.144
2	2.184	9.102	19.246	2.184	9.102	19.246
3	1.811	7.547	26.793	1.811	7.547	26.793
4	1.579	6.581	33.374	1.579	6.581	33.374
5	1.511	6.297	39.671	1.511	6.297	39.671
6	1.433	5.969	45.639	1.433	5.969	45.639
7	1.385	5.772	51.412	1.385	5.772	51.412
8	1.256	5.234	56.646	1.256	5.234	56.646
9	1.180	4.918	61.563	1.180	4.918	61.563
10	1.115	4.646	66.209	1.115	4.646	66.209
11	1.073	4.470	70.679	1.073	4.470	70.679
12	0.974	4.058	74.737			
13	0.840	3.501	78.238			
14	0.734	3.059	81.297			
15	0.719	2.997	84.294			
16	0.681	2.836	87.131			
17	0.613	2.555	89.686			
18	0.530	2.208	91.894			
19	0.481	2.006	93.899			
20	0.441	1.838	95.737			
21	0.400	1.668	97.405			
22	0.337	1.405	98.810			
23	0.286	1.190	100.000			
24	1.086×10^{-16}	4.527×10^{-16}	100.000			

The values in the component matrix reflect the relevance of each keyword in a different common factor. We can judge the relevance of the keywords by the value of the correlation coefficient. This can provide a more detailed basis for clustering of the keywords. The larger the critical value is, the simpler the classification structure is. According to the results of the cluster analysis, we set the load critical value at 0.5. As we can see in Table 6, 11 keywords factor load values are less than 0.5, such as data mining, rough set, KDD, model, database, association rules, clustering, decision tree, machine learning, knowledge discovery system, and attribute reduction. The keywords with a load value less than 0.5, have a low relevance, on one hand, some of the keywords have a high frequency, but they are not distinctive during the analysis of the correlation factors because they represent a large range of research topics, for example, data mining and a rough set. On the other hand, some of the keywords represent a more novel topic [8], like knowledge

discovery system, these 11 keywords can be determined by the branch of their hotspots according to the result of the cluster analysis.

Table 6. Factor analysis, component matrix (the loading factor value is greater than 0.50)

No	Component										
	1	2	3	4	5	6	7	8	9	10	11
1											
2											
3											
4											
5											
6										-0.629	
7											
8											
9	-0.679										
10											
11		0.665									
12			-0.586								
13											
14											
15	-0.634										
16		0.512									-0.585
17		0.688									
18								-0.633			
19									-0.722		
20											
21				-0.506							
22	-0.577										
23						-0.737					
24			-0.563								

As we can see in Table 6, 13 keywords factor loadings with values greater than 0.5 can be divided into the following 8 groups according to their correlation:

- 1) ontology, knowledge management, domain knowledge;
- 2) expert system, double-base cooperating mechanism, knowledge base;
- 3) information system, knowledge mining;
- 4) genetic algorithm;
- 5) CRM;
- 6) digital library;
- 7) database;
- 8) non-interactive literature.

3.4. Results analysis

According to the results of cluster analysis and factor analysis, the highly frequency keywords in the field of domestic knowledge discovery can be divided into the following nine groups:

- 1) ontology, knowledge management, domain knowledge;
- 2) expert system, genetic algorithm, knowledge discovery system, knowledge base, double-base cooperating mechanism;
- 3) association rules, clustering;
- 4) data mining, KDD, database, machine learning, data warehouse;
- 5) information system, knowledge mining;
- 6) rough set, attribute reduction;
- 7) decision tree, CRM;
- 8) model, non-interactive literature;
- 9) digital library.

Due to the grouping results above and with the help of the interactive literature, we can summarize that in the field of domestic knowledge discovery, the researches are mainly focused on the following six aspects.

1. Knowledge discovery based on the data research (keywords in groups 4 and 5)

How to get useful knowledge (data mining, KDD, machine learning and knowledge mining are the processes of extraction or discovery knowledge from a database) from different and substantial knowledge sources (database, data warehouse and information systems are all stored data collection) has always been a hotspot in the field of knowledge discovery research.

2. Knowledge discovery algorithm optimization research (keywords in groups 3 and 6)

Knowledge discovery algorithm optimization research can improve the efficiency of knowledge discovery, therefore, the research about the algorithm optimization in knowledge discovery has become a hot topic in domestic knowledge discovery. In algorithm optimization research, the rough set knowledge classification method based on the attribute reduction and association rules based on the clustering technology can reduce the amount of information to be processed and improve the efficiency of knowledge discovery dramatically.

3. The model of knowledge discovery and research of literature study (keywords in group 8)

By analyzing the interactive literature with the help of non-interactive literature and model research, finding the scientific value implicit associations that were never found, thus, the scientific suppositions can be established, making the research more direct. Thus the research on knowledge discovery in non-interactive literature grows stronger day by day.

4. Based on domain ontology knowledge management (keywords in group 1)

Finding and understanding knowledge of related fields are crucial in knowledge management. As a formal specification of shared conceptualization, ontology offers a clear definition of the concept of the relationship between the various concepts in order to achieve the user's knowledge about the field of mutual understanding and a consensus. It was widely used in the field of knowledge management recently. For this reason, research on building domain ontology to discover and understand the relevant field of knowledge has become a research hotspot in the field of knowledge discovery.

5. Expert system construction research (keywords in group 2)

Whether to deal with the simulation lost when solving the highly complex data in databases by using a genetic algorithm (Zhu Rui, 2000), or to expand the primary knowledge base which direct source is the experience of the experts or journals knowledge with the help of the double-base cooperating mechanism Chen Jinhai (2003), both are designed to improve the expert system and the efficiency of decisions. Their improvement can lead to an improvement of the enterprises to some degree. For this reason the expert system construction research has become a research focus in the field of knowledge discovery.

6. Applied research of knowledge discovery (keywords in groups 7 and 9)

The applied research of knowledge discovery can strengthen the core competence of the enterprises which have a high research value. Current applications of knowledge discovery mainly focus on the management of the customer relationship and the application of the digital libraries. Individualized information service has become one of the main tendencies in the future development of digital libraries. Moreover, knowledge discovery technology can be applied to acquire individualized information service, thus the satisfaction of the users can be improved. By applying the knowledge discovery technology to customer relationship management, the customer relationship management could go to a more intelligent direction, which is good for the enterprises to forecast their business trends, and helps them to mining the potential customers and keep the current customers. In this way the enterprises can be in a more favorable competitive situation.

4. Hotspots in the field of international knowledge discovery

Analyzing the research hotspot in the field of domestic knowledge discovery can make the domestic scholars have a clearer understanding on the research status in the field of domestic knowledge discovery. However, in order to better understand the research status, we also need to know the current hotspots in the field of international knowledge discovery, to find the limitations of domestic research on knowledge discovery. We selected the foreign language database in CNKI as the source of foreign literature, which includes more than 40 famous international publishers to citations of periodical literature data. The database covering SCI, EI, 90% of SSCI journals, makes the retrieval of “knowledge discovery” journal literature to a large extent reflect the international research hotspot in the field of knowledge discovery.

In retrieval conditions of (1) the name of the paper equals to “knowledge discovery”; (2) the deadline equals to 2013. By retrieving, 985 papers were discovered, the data set constructed in this paper, the number of keywords given by the authors is 1911 totally, after removing the duplicate keywords, the number of keywords is 1012. Table 7 is the list of high frequency keywords.

Table 7. The high frequency keywords list of foreign literature

No	Keyword	Word frequency
1	Knowledge discovery	82
2	Humans	42
3	Data mining	28
4	Artificial intelligence	28
5	Algorithms	25
6	Databases, factual	23
7	Information storage and retrieval	19
8	Computational biology	17
9	Data interpretation, statistical	11
10	Software	11
11	Machine learning	11
12	Database management systems	11
13	Female	10
14	Knowledge discovery in databases	10
15	Classification	9
16	Knowledge	8
17	Knowledge bases	8
18	Natural language processing	7
19	Animals	7
20	Internet	6
21	Text mining	6
22	Medical informatics	6
23	Cluster analysis	5
24	Male	5

Table 8. A partial co-word matrix of foreign literature

$j \backslash i$	1	2	3	4	5	6	7	8	9	10
1	93	0	40	1	10	0	2	0	1	3
2	0	72	7	29	22	28	14	14	13	8
3	40	7	85	3	7	1	1	3	2	1
4	1	29	3	48	19	16	12	5	4	7
5	10	22	7	19	50	12	14	8	5	9
6	0	28	1	16	12	34	12	7	7	6
7	2	14	1	12	14	12	31	7	4	9
8	0	14	3	5	8	7	7	20	3	4
9	1	13	2	4	5	7	4	3	15	3
10	3	8	1	7	9	6	9	4	3	20

In the same way we constructed a correlation matrix by using the correlation coefficient ρ , and transformed it into a dissimilarity matrix. Then, based on the results of cluster analysis and factor analysis, which were shown in Fig. 2 and Table 9, and combining the related literatures, we can divide the high frequency keywords in the field of international knowledge discovery into 6 groups as follows:

1. Knowledge discovery; data mining; classification; machine learning; knowledge discovery in databases.
2. Information storage and retrieval; database management systems; natural language processing; Internet; knowledge bases; cluster analysis.
3. Humans; artificial intelligence; female; male; databases, factual; knowledge; animals; computational biology; data interpretation, statistical; software.
4. Text mining.
5. Algorithms.
6. Medical informatics.

We found that data mining and knowledge discovery are inseparable according to the above grouping, and data mining is the core and basis in the field of international knowledge discovery. In addition, no matter in the field of domestic or international knowledge discovery, KDD and the algorithm research of knowledge discovery are the research emphasis. However, most of the researches in the field of international knowledge discovery are focused on biology and medicine, which is quite different from the field of domestic knowledge discovery.

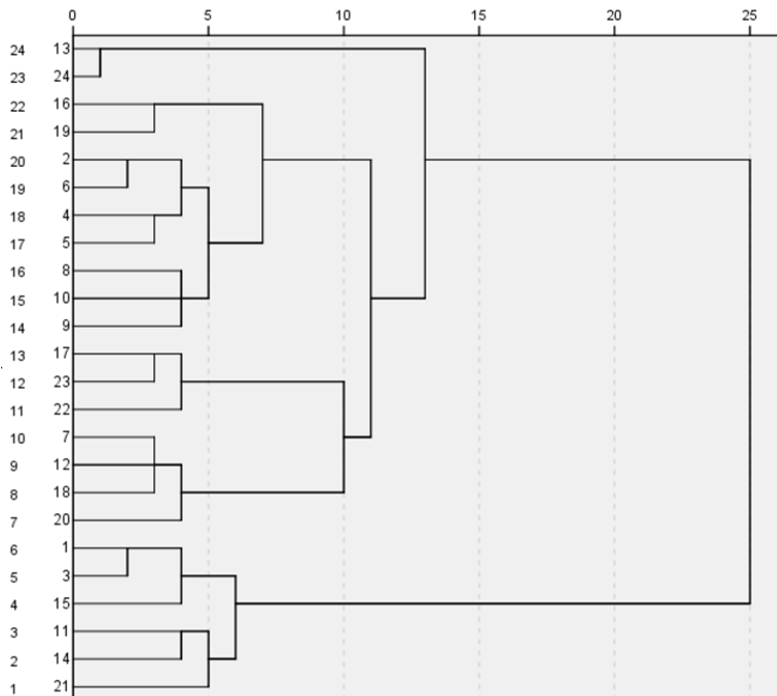


Fig. 2. Cluster analysis-tree of foreign literature

Table 9. Factor analysis, component matrix of foreign literature
(loading factor value is greater than 0.50)

No	Component							
	1	2	3	4	5	6	7	8
1	-0.562							
2	0.824							
3	-0.570							
4	0.669							
5	0.503			0.616				
6	0.794							
7	0.713							
8	0.604							
9								
10								
11	-0.570							
12		0.681						
13		-0.690						
14								
15								
16			-0.557					
17			0.657					
18		0.665						
19			-0.659					
20		0.504						
21								-0.501
22					-0.662			
23			0.598					
24		-0.690						

5. Conclusion

Using SPSS software, based on the co-word analysis method, we analyzed the domestic core journals before 2014, and finally got a conclusion of 6 hotspots in the field of domestic knowledge discovery, including the knowledge discovery based on data research, knowledge discovery algorithm optimization research, the model of knowledge discovery and research of literature study, knowledge management based on domain ontology, expert system construction research, and applied research of knowledge discovery. Finally, we summarized the research hotspots in the field of international knowledge discovery in the same way, by comparing the differences of the two, it is concluded that domestic scholars can broaden the research of knowledge discovery in the field of biology and medicine.

Acknowledgements: This research was supported by the Fundamental Research Funds B11JB00500 for the Central Universities.

References

1. Fayyad, U., G. Piatetsky-Shapiro, P. Smyth. From Data Mining to Knowledge Discovery in Databases. – *AI Magazine*, Vol. **17**, 1996, No 3.
2. Hua, Bolin. Data Mining and Knowledge Discovery Relationship Analysis. – *Information Studies: Theory & Application*, Vol. **31**, 2008, No 4, 507-510 (in Chinese).
3. Ma, Feicheng, Zhang Qin. Comparative Analysis of Knowledge Management Literature between China and Overseas: A Bibliometric Analysis. – *Journal of the China Society for Scientific and Technical Information*, Vol. **25**, 2006, No 2, 163-171 (in Chinese).
4. Zhang, Qin, Ma Feicheng. An Approach to the Structure Map of Knowledge Management Research in China: A Bibliometric Analysis. – *Journal of the China Society for Scientific and Technical Information*, Vol. **27**, 2008, No 1, 93-101 (in Chinese).
5. Sun, Xiaoning, Chu Jiewang. On Hotspots of Master and Ph. D. Degree's Dissertations in the Field of Knowledge Management during the Last Decade: A Co-Words Analysis. – *Journal of Intelligence*, Vol. **6**, 2012, 85-90 (in Chinese).
6. Zhang, Qin, Xu Xusong. On Discovering the Structure Map of Knowledge Management Research Abroad – Integration of a Bibliometric Analysis and Visualization Analysis. – *Journal of Industrial Engineering and Engineering Management*, Vol. **4**, 2008, 30-35, 50 (in Chinese).
7. Zhang, Qin, Ma Feicheng. On Paradigm of Research Knowledge Management: A Bibliometric Analysis. – *Journal of Management Sciences in China*, Vol. **6**, 2007, 65-75 (in Chinese).
8. Dong, Wei. On Hotspots in the Field of Digital Library during the Last Decade: A Co-Words Analysis. – *Document, Information & Knowledge*, Vol. **5**, 2009, 58-63 (in Chinese).