Patent Technology Competitor Group Analysis Method Based on IPC

Yuan Fu
Information Technology Supporting Center, Institute of Scientific and Technical Information of China
No. 15 Fuxing Rd,.Haidian Distirct, Beijing 100038, P.R. China
+86 10 5888 2447
fuyuan2014@istic.ac.cn

Hongqi Han
Information Technology Supporting Center, Institute of Scientific and Technical Information of China
No. 15 Fuxing Rd,.Haidian Distirct, Beijing 100038, P.R. China
+86 10 5888 2447
bithhq@163.com

Lijun Zhu
Information Technology Supporting Center, Institute of Scientific and Technical Information of China
No. 15 Fuxing Rd,.Haidian Distirct, Beijing 100038, P.R. China
+86 10 5888 2447
zhulj@istic.ac.cn

ABSTRACT

It is crucial to understand the technical groups of intra-industry and to master the competition in the field of technology. In order to provide valuable information for industry participants and policymakers, a process model for mining technical competitor groups based on IPC classification number is put forward. Firstly, the patent numbers under each IPC are counted for building feature vectors for competitors. Then, technical similarities between each pairs of competitors are computed. Finally, the LinLog graph clustering algorithm is carried out to discover three levels of groups, i.e. institution, province and country. To obtain patent data for this research, an acquisition system for Chinese patent data is developed. Experiments on the field of fuel cell is conducted and the results show the technique is helpful and effective.

Categories and Subject Descriptors

Information extraction from patent documents

General Terms

Experimentation

Keywords

LinLog; IPC classification number; Technology competitor group

Published at Ceur-ws.org
Proceedings of the Second International Workshop on Patent Mining and its Applications (IPAMIN). May 27–28, 2015, Beijing, China.

1. INTRODUCTION

Competitiveness is a typical characteristic for industrial technology (Yoon, 2008) [1]. Practically, for almost every emerging industry, some kinds of technology will become leading and predominant after developing over a period of time. Agglomeration is common for an industry. When the industrial technology agglomerates to a certain extent so that it can meet the needs of product functions well, the industry will become mature, and the industrial technology system is established. On the other hand, the technology owner compete reciprocally into different technical groups. According to Porter's theory of competitive advantage, the real competitors inside an industry are companies similar to a company (Lee, 2006) [2]. These similar companies constitute a strategic group and become a sub-industry. A company has barriers to enter different strategy groups. Therefore companies which have very similar industrial technology are likely to be main competitors.

The clustering method of dividing data into several clusters can reflect relational schema of the data and the knowledge hidden in the data. The method of competitor group analysis of industrial technology is to use appropriate clustering algorithm to divide competitors into several groups, and thus identify similar competitors inside an industry competitions and their reciprocal influences. The level of technical competitor group analysis can be from different aspects such as countries, provinces, and institutions. The purpose of the analysis is to understand the technical groups inside an industry, and to master the competition in the field of technology from higher levels, and to provide valuable information for industry participants and policymakers.

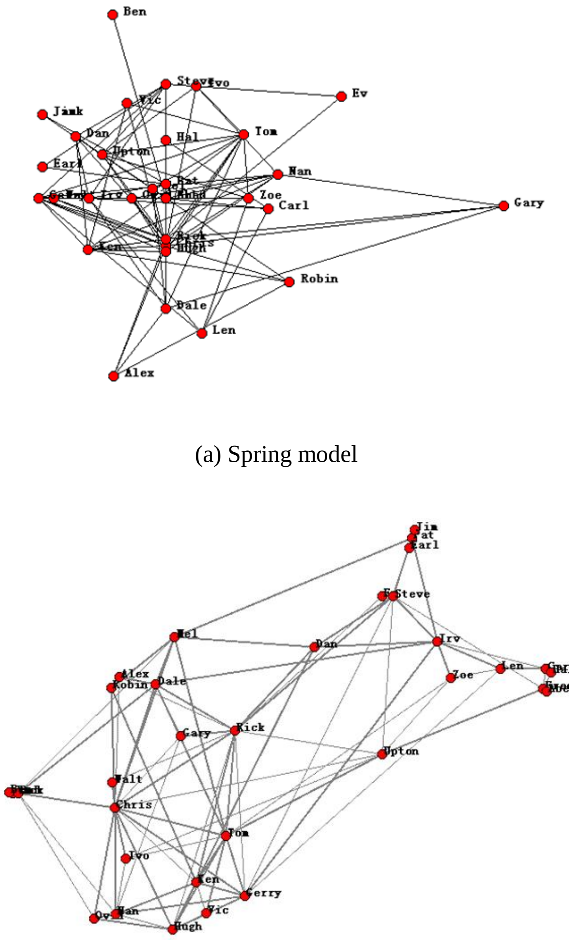
Some common clustering algorithms can be used to identify the competitor group of industrial technology, such as self-organizing mapping (SOM), K-means (Lee, 2009) [3], factor analysis, etc. In these models, each competitor is usually expressed as a feature vector which are measured by several technical characteristics. Similar objects will be clustered into one group by calculating distances between them. For example, (Pilkington, 2004)[4] used UPC number and IPC classification respectively as the technical features for competitors and used the factor analysis model to cluster 52 companies in the field of fuel cell into five groups.

Literature studies found that many researchers have used visualization methods. The traditional clustering algorithm is based on the unsupervised learning so people often doubt the effectiveness of the analysis results. The visualization method can display abstract data using graph or picture because it combines the computer technology and human cognitive ability effectively. Therefore, the visualization method enhances the user’s confidence for the analysis results, so it has been widely accepted in recent years. Considering the advantages of visualization, the proposed method will use graph clustering method to find technical competitor groups.

2. RELATED WORK

2.1 LinLog graph clustering methods

LinLog algorithm was first put forward by (Noack, 2007) [5]. The aim of the algorithm is to produce ideal and visual clustering graphs. Figure 1 shows an example mentioned in Noack's paper (Noack, 2005) [6]. In the example, Spring and LinLog algorithm were employed respectively for graph clustering using the same data. Comparatively, LinLog algorithm clearly divided data into two large clusters which are connected by two nods, Dan and Upton, while Spring algorithm positioned nodes with high degree in the center and nodes with low degree near the borders.



(a) Spring model

(b) LinLog model

Figure 1 Comparison of Spring and Linlog method

The LinLog model does not conform to the traditional aesthetic standard, it aims to group nodes of closely connected and separate nodes of partially connected. There are two kinds of LinLog models: node-repulsion model and edge-repulsion model(Coscia, 2009) [7]. The two models are based on two famous clustering standards respectively (Li, 2008) [8], namely density of cut and normalized cut. Normalized cut and edge-repulsive model can produce unbiased results, therefore it is especially suitable for normally distributed data. In this paper, LinLog algorithm of Barnes and Hut hierarchy algorithms is used to draw clustered graphs (Stegmann, 2003) [9]. After the algorithm draw graphics, it also divide nodes into several clusters.

2.2 IPC

IPC means the international patent classification. IPC is an international standard which is used by the patent offices of all countries or regions in the world. Although some countries or regions make its own patent classification system, such as CPC system of USPTO, ECLA system of EPO, they provide the IPC classification number. Chinese patent classification system also use IPC system. A patent has at least one IPC number, but is not limited to one IPC classification number. In other words, some patents are endowed with two or more IPC classification numbers. The first classification number is called the main classification number when there are multiple patent classification numbers.

According to the characteristics of technical topics of the invention, the technology fields in IPC system are divided into 8 sections. Each section represents a kind of technology, designated by one of the capital letters A through H as shown in Table1.

Table 1   section of technology in IPC system

Section	Section Title
A	HUMAN NECESSITIES
B	PERFORMING OPERATIONS; TRANSPORTING
C	CHEMISTRY; METALLURGY
D	TEXTILES; PAPER
E	FIXED CONSTRUCTIONS
F	MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS;BLASTING
G	PHYSICS
H	ELECTRICITY

The structure of IPC classification system is hierarchical. Sections are the highest level of hierarchy in the system. Each section is subdivided into classes which are the second hierarchical level. Each class comprises one or more subclasses which are the third hierarchical level. Each subclass is broken down into subdivisions referred to as “groups”, which are either main groups (the fourth hierarchical level) or subgroups (lower hierarchical levels dependent upon the main group level). A complete classification symbol comprises the combined symbols representing the section, class, subclass and main group or subgroup, as shown in Figure 2. Currently, there are approximately 70,000 subdivisions in the classification system. Figure 3 is a sample of the hierarchical structure.

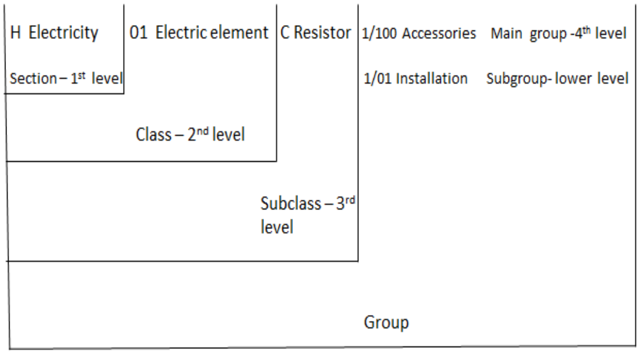


Figure 2 Hierarchical structure of the IPC classification system

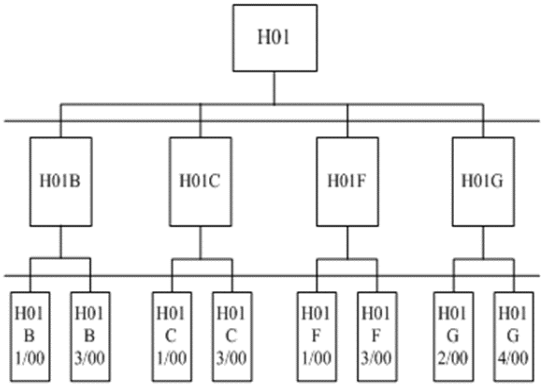


Figure 3 A sample of IPC hierarchical structure

3. METHOD

An industrial technology field can be divided into several subfields, and each subfield may have smaller technology subfields. Technology competitors often have different research background, bases, objectives and priorities. Competitors with similar technology may be competitors or partners on the market, and they are likely to interact with each other. IPC classification codes are designated by patent examiner with professional knowledge. Therefore IPC provide an effective way to know industrial hot points, and research and development directions of technology competitors. A technology competitor tend to invest research in several technical subfields, so it is difficult to determine whether two competitors have similar research technology only from the IPC count statistics. Therefore, a graph clustering method based on main IPC number is put forward to identify technology competitor groups within an industrial technology field. Figure 4 shows the process model of this method.

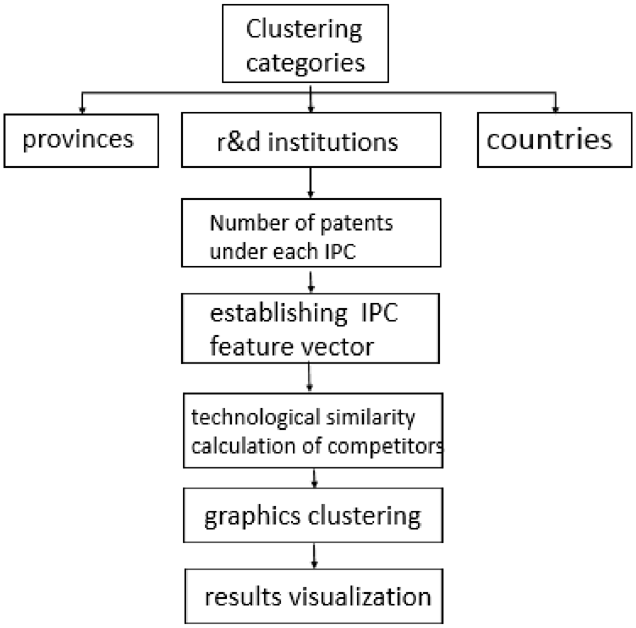

Clustering categories
provinces
r&d institutions
countries
Number of patents under each IPC
establishing IPC feature vector
technological similarity calculation of competitors
graphics clustering
results visualization

Figure 4 The process model of the graph clustering method

Firstly, selecting a clustering level from three categories: institutions, provinces and countries. Then, counting the patent number under each main IPC classification number for each technology competitor. Then the association matrix is established between technology competitors and the main IPC classification number (Dibattista, 1994)[10]. Each technology competitor is expressed as a feature vector whose attributes are IPC classification numbers. The value of each attribute item is the number of patents under the main IPC classification number. Finally, calculating the similarity between each pair of technological competitors by using cosine formula(Fruchterman, 1991)[11]. Let $|IPC|$ as the number of the IPC main classification number covered by industrial technology, and the patent number of competitor *i* under *k*-th IPC classification number is IPC_{ki} . The equation (1) shows how to compute the technological similarity between competitor *i* and *j*.

$$sim(i,j)=\frac{\sum_{k=1}^{|IPC|} IPC_{ki}\times IPC_{kj}}{\sqrt{\sum_{k=1}^{|IPC|} {IPC_{ki}}^2}\times\sqrt{\sum_{k=1}^{|IPC|} {IPC_{kj}}^2}} \qquad (1)$$

In order to obtain good visual graphics, a minimum similarity threshold (Noack, 2004)[12] should be set. Generally, the threshold is set to the mean of similarity, yet it can also be determined by experiments. There will be a connect between two technology competitors when the similarity between them is higher than the set threshold. Using technology competitors as nodes, the connections between them as edges, and the weight of the edges are the technological similarity values between them, LinLog graph clustering algorithm will generate visual map. The map will show the clusters for identifying competitor groups.

4. DATA

4.1 Data acquisition

Nowadays, almost all patent offices of major countries and regions provide patent databases on their official web sites. People can connect these websites any time and everywhere via the Internet to obtain the patent data freely. In order to get patent data quickly, a patent data acquisition system (Laura, 2008) [13] is developed. The model of the system model is shown in Figure 5. The acquisition system can fetch HTML web pages which contains the patent description information from the official website of the state intellectual property office of China (http://www.sipo.gov.cn/). After the patent information is collected, it can automatically obtain the items of description and legal status of patents through the content analysis of web pages and save them into the local databases.

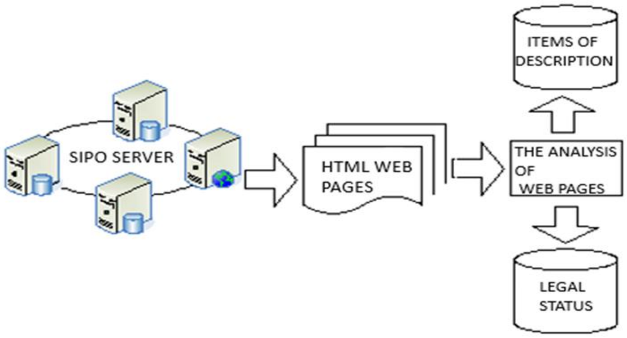


Figure 5 The data acquisition system model

In order to test the effectiveness of proposed method, the patent acquisition system is run to download patent data in the field of fuel cell technology. 6346 patents are collected totally. The following preprocessing steps and the empirical analysis will employ the downloaded patent data.

4.2 Data preprocess

The collected data often have some problems, and it must be preprocessed before the formal analysis. In the experiment, the patent data will be preprocessed to meet the analysis requirements, including identifying the patent categories, countries and provinces of applicants, and categories of applicants, etc.

If the first applicants are Chinese individuals or organizations, the addresses of the applicants often contain the information of its province (Kayal, 1999) [14]. Generally, the first 6 digits of the address description is the applicant's postcode, so the province information can be obtained according to the postcode. If the first applicants are foreign individuals or organizations, the priority item and the international publication item in patent descriptions contain the state information. For example, the priority item of a patent is "1999.8.27 JP 242132/1999", where JP means that the applicant is a Japanese.

For the purpose of the research, applicants are divided into 5 categories: company, university, research institute, personal and the other. The categories are identified by the keywords in the applicant names. The corresponding relation of keywords and categories are shown in Table 2. If there are more than one applicants in a patent description, only the first applicant is considered. For example, there are two applicants of the patent No. 00112136.7: Nanjing Normal University and Changchun Institute of Applied Chemistry Chinese Academy of Sciences, the system will designate "university" category to the patent.

Table 2 keywords for identifying application category

category	Key words
company	company, partnership
university	university, college
institute	research institution,
others	committee, association, foundation
personal	

5. EXPERIMENTAL RESULTS

5.1 Research and development institutions

In order to have clear visual map, we choose top 20 research and development institutions for graph clustering algorithm. The result is shown in Figure 6. In the map, the size of nodes represents the number of granted invention patents, and the color of nodes shows the group they belong to (Reinhard, 2007) [15].

In the case, the LinLog algorithm identified two technology competitor groups (shown in Figure 6). The group with red node color is the first group, including 10. They are: Samsung (177), Chinese Academy of Sciences(128), Antiq(74), General Motors(56), Honda(52), Wuhan University of Technology(49), Shanghai Jiaotong University(38), Sanyo(37), BYD(32), and Harbin Institute of Technology(26); The group with orange node color is the second group, including 10 other institutions. They are: Shanghai Shen-Li High Tech(194), Panasonic(154), Toyota(120), Tsinghua university(72), Nissan(62), Toshiba(48), Sunrise Power(26), Hitachi (24), LG(20) and UTC (19). The numbers in parentheses after company names means the numbers of their granted invetion patents. Table 3 shows corresponding English names of Chinese Names in Figure 6.

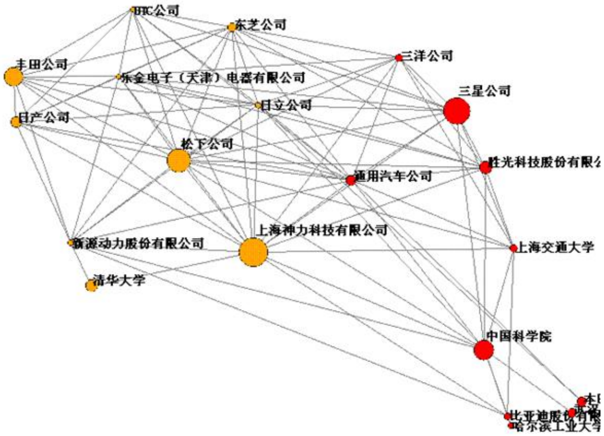


Figure 6 Clustering result of R&D institutions

Table 3 Corresponding English names of Chinese names of R&D institutions in Figure 6

Chinese name	English name
清华大学	Tsinghua University
新源动力股份有限公司	Sunrise Power
上海神力科技有限公司	Shanghai Shen-Li High Tech
松下公司	Panasonic

日产公司	Nissan
丰田公司	Toyota
日立公司	Hitachi
东芝公司	Toshiba
BTC 公司	BTC
乐金电子电器有限公司	LG
上海交通大学	Shanghai Jiaotong University
中国科学院	Chinese Academy of Sciences
三星公司	Samsung
三洋公司	Sanyo
通用汽车公司	General Motors
胜光科技股份有限公司	Antiq
哈尔滨工业大学	Harbin Institute of Technology
比亚迪股份有限公司	BYD
本田株式会社	Honda
武汉大学	Wuhan University of Technology

5.2 Provinces

In the case, totally 22 provinces are extracted in all fuel cell patents. The graph clustering result is shown in figure 7. The biggest node in the picture is Shanghai, which means the research strength of Shanghai province is the strongest one in China. While the smallest node is Hebei, which means Hebei province is the weakest one on the research of fuel cell in these provinces.

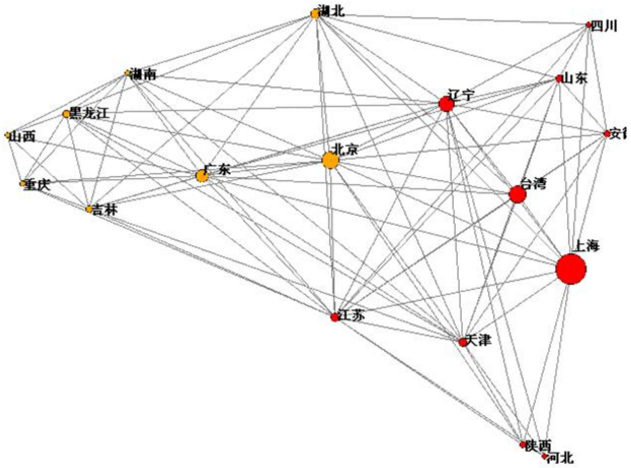


Figure 7 The clustering results of provinces

Table 4 Corresponding English names of Chinese names of provinces in Figure 7

The Chinese Name	The English Name
上海	Shanghai
台湾	Taiwan
辽宁	Liaoning
江苏	Jiangsu
天津	Tianjin
山东	Shandong
安徽	Anhui
陕西	Shaanxi
四川	Sichuan
河北	Hebei
北京	Beijing
广东	Guangdong
湖北	Hubei
黑龙江	Heilongjiang
吉林	Jilin
重庆	Chongqing
湖南	Hunan
山西	Shanxi

In the province level, two technology competitor groups are identified. The group with red nodes is the first group, including 10 provinces: Shanghai (311), Taiwan (152), Liaoning (127), Jiangsu (41), Tianjin(40), Shandong(23), Shaanxi(13), Anhui (19),Sichuan (4) and Hebei(2), The group with orange node color represents the second group, including 8 provinces: Beijing(150), Guangdong(93), Hubei(58), Heilongjiang(29), Jilin(18), Chongqing (5), Hunan(4) and Shanxi Province (4). Because the technology similarity value of Zhejiang (16), Fujian (12), Yunnan (1) and Inner Mongolia (1) is lower than the set threshold, the clustering result do not include these provinces. Similarly, the number in parentheses is the number of granted patents of provinces.

5.3 Countries

In the case, totally 17 countries or regions are extracted in all fuel cell patents. The graph clustering result is shown in Figure 8. Obviously, the biggest node in the graph is China, the granted patent number of which is 1123. While the smallest nodes are Denmark and Finland. The granted patent number of both country are 3.

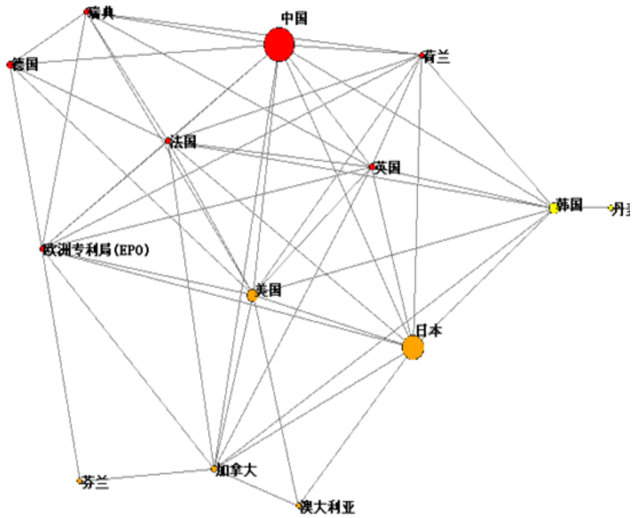


Figure 8 The clustering result of countries

Table 5 Corresponding English names of Chinese names of R&D institutions in Figure 8

The Chinese Name	The English Name
中国	China
德国	Germany
英国	Britain
法国	France
欧洲专利局	EPO
瑞典	Sweden
荷兰	Netherlands
日本	Japan
美国	the United States
加拿大	Canada
澳大利亚	Australia
芬兰	Finland

In the country level, four technology competitor groups are identified, containing 16 countries and regional organizations. The group with red node color represents the first group, including seven countries and regional organizations: China (1123), Germany (58), Britain (28), France (16), EPO (10), Sweden (6), and Netherlands (5). The group of orange node color represents the second group, including 5 countries: Japan (740), the United States (292), Canada (33), Australia (4) and Finland (3). The third group consists of Korea (202) and Denmark (3) two countries. The fourth group includes Norway (3) and Italy (1). There is an edge between Norway and Italy, but there are no edges with other nodes (Figure 9), however Figure 8 can't show them because LinLog algorithm has problems to generate clusters with unconnected graphs. The technology similarity of Austria (1) with other countries is lower than the threshold, so the clustering figure does not include Austria (1).

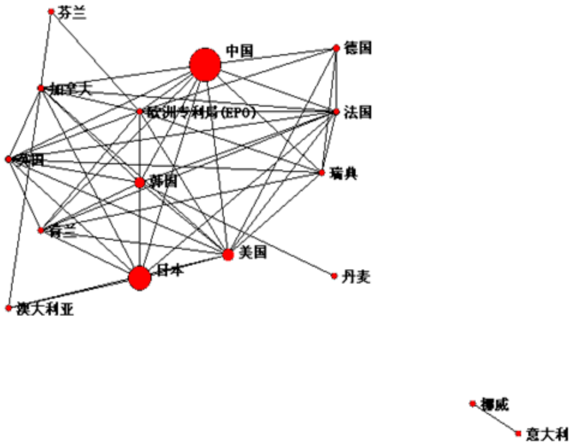


Figure 9 The clustering figure of unconnected states

Table 6 Corresponding English names of Chinese names of R&D institutions in Figure 9

The Chinese Name	The English Name
中国	China
德国	Germany
英国	Britain
法国	France
欧洲专利局	EPO
瑞典	Sweden
荷兰	Netherlands
日本	Japan
美国	the United States
加拿大	Canada
澳大利亚	Australia
芬兰	Finland
挪威	Norway
意大利	Italy
韩国	Korea

6. CONCLUSION

In the paper, a graph clustering algorithm is used to obtain technology competitor group analysis based on IPC. The proposed method consists of four stages. First, the clustering level is determined. There are three levels for selected, i.e. institute, province and country. Second, the numbers of patents are counted under each IPC for each object (competitor) in the selected level. Third, each object is expressed with a vector, the attributes of which are IPC classification codes, and the value of each attribute is corresponding patent count. Fourth, technology similarities are computed between each pair of competitors. Finally, Linlog algorithm is used to cluster competitors into groups and display them in a graph to improve the confidence of analysis results.

Experimental results on fuel cell demonstrate the effectiveness of the proposed method.

7. ACKNOWLEDGMENTS

This work is partially supported by National Natural Science Foundation of China (Project 71473237), and partially supported by the Key Work Project of Institute of Scientific and Technical Information of China (ISTIC) (ZD2014-7-1). Authors are grateful to the National Natural Science Foundation of China and the Ministry of Science and Technology of China for financial support to carry out this work.

8. REFERENCES

[1] Yoon, B. and Lee, S. 2008. Patent analysis for technology forecasting: sector-specific applications. *C. 2008 IEEE International Engineering Management Conference*.

[2] Lee, C. K. and Ong, R. 2006. An analysis of the liquid crystal cell patents of LG and Samsung filed at the USPTO. *C. 2006 IEEE International Conference on Management of Innovation and Technology*.

[3] Lee, S., Yoon, B., and Park, Y. 2009. An approach to discovering new technology opportunities: Keyword-based patent map approach. *J. Technovation.* 29, 6, 481-497.

[4] Pilkington, A. 2004. Technology portfolio alignment commercialisation: an investigation of fuel cell patenting. *J. Technovation.* 24, 10, 761-771.

[5] Noack, A. 2007. Energy models for graph clustering. *J. Journal of Graph Algorithms and Applications.* 11, 2, 453-480.

[6] Noack, A. Energy-based clustering. *C.13th International Symposium on Graph Drawing*. 2005.

[7] Coscia, M., Giannotti, F., and Pensa, R. 2009. *Social network analysis as knowledge discovery process: a case study on digital bibliography*. *C. International Conference on Advances in Social Network Analysis and Mining (ASONAM)*.

[8] Li, Wanchun., Eades, P., and Nikolov, N. 2008. Using spring algorithms to remove node overlapping. *C. Proceedings of the 2005 Asia-Pacific symposium on Information visualization*.

[9] Stegmann, J. and Grohmann, G. 2003. Hypothesis generation guided by co-word clustering. *J. Scientometrics*. 56, 1, 111-135.

[10] Dibattista, G., Eades, P., Tamassia, R., and Tollis, I. G. 1994 Algorithms for drawing graphs: an annotated bibliography. *J. Computational Geometry: Theory and Applications.* 4, 5, 235-282.

[11] Fruchterman, T. M. J. and Reingold, E. M. 1991. Graph drawing by force-directed placement. *J. Software-Practice and Experience.* 21, 11, 1129-1164.

[12] Noack, A. 2004. An energy model for visual graph clustering. *C. 11th International Symposium on Graph Drawing*.

[13] Laura, R. 2008. Data mining tools for technology and competitive intelligence. *R. VTT TIEDOTTEITA Research notes 2451*.

[14] Kayal, A. A. and Waters, R. C. 1999. An empirical evaluation of the technology cycle time indicator as a measure of the pace of technological progress in superconductor technology. *J. IEEE Transactions on Engineering Management.* 46, 2, 127-131.

[15] Reinhard, H., Martin, K., and Marcus, K. 2007. Patent indicators for the technology life cycle development. *J. Research Policy*. 36, 3, 387-398.