ATLANTIS PRESS
Advances in Engineering Research (AER), volume 131
3rd Annual International Conference on Electronics, Electrical Engineering and Information Science (EEEIS 2017)

The Method for Discovering Technology Competitor Groups Based on Graph Clustering

Yong-Sheng YU[1,a], Hong-Qi HAN[2,b,*], Zhong LI[3,c]

[1-3]Institute of Scientific and Technical Information of China, No. 15 Fuxing Rd,.Haidian Distirct, Beijing 100038, P.R. China

[a]yuys2015@istic.ac.cn, [b]bithhq@163.com, [c]lizhong2016@istic.ac.cn

*Hong-Qi HAN

Keywords: Term extraction, technology competitor groups, clustering analysis.

Abstract. For enterprise decision-makers, it is crucial to timely find technology competitors and analyze competitive situation of industrial technology. A method for discovering technology competitor groups based on graph clustering is put forward to improve the precision of clustering results. The method extracts the text terms of patents to build vector space model, counts the numbers of similar patents of technology competitors and uses LinLog graph clustering tool to mine technology competitor groups. The patent data of Chinese fuel cell was collected to carry out the experiment of this research and the experiment results showed that the presented clustering analysis method is effective.

Introduction

Understanding the development trends and competition situations of industrial technology is very important for enterprises. Traditional competition analysis methods of industrial technology mainly concentrate on the concept, sources and evaluation of regional industrial competitiveness, but it couldn't efficiently reveal competitive relationships as well as competitive intensity among enterprises [1]. According to life cycle theory, when industrial technology has developed for a period of time, some enterprises may gradually focus on specific aspects of industrial technology and the aggregation phenomenon will occur. These enterprises which aggregate to be a group usually have similar industrial technology and are most likely to be technology competitors with each other [2].

In order to discover valuable information for managers of enterprises to make decision, patent data is usually collected to find technology competitors and their groups by carrying out clustering analysis method. On the one hand, the patent text is the most effective carrier of the latest technical information; on the other hand, patent data is helpful to identify potential competitor and to master the industry competition situation on the whole [3].

Our method will use graph clustering method to divide technology competitor groups. Because graph clustering methods are becoming more popular and are used more widely in the researchers, and visualization methods can clearly display experiment results of abstract data and simply reveal the rules and logicality hidden in experiment data. For example, Beck [4] developed a more accessible visual analysis system, called SurVis, that is to disseminate a carefully surveyed literature collection. The graph clustering analysis method of technology competitor divides technology competitors into several strategy groups so that competitors in different groups have dissimilar technology, and thus identify competition relationships of technology competitors inside an industry.

Related Works

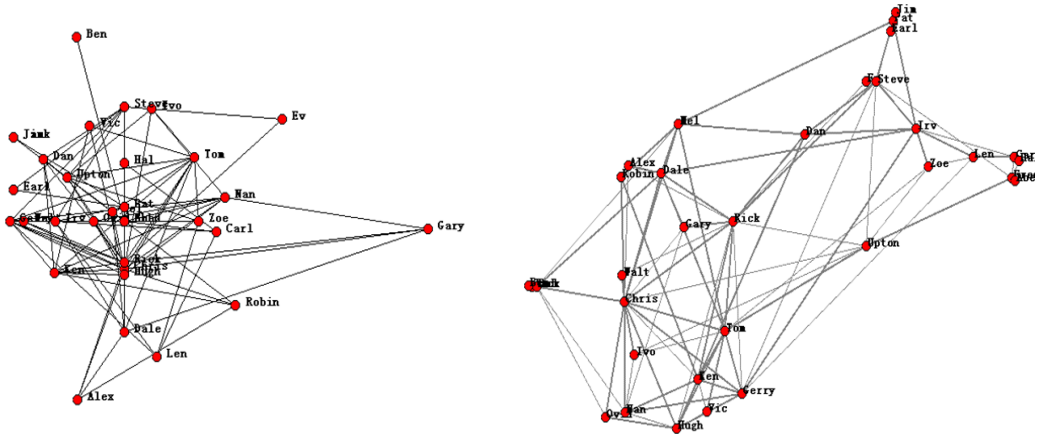
LinLog graph clustering method

Most traditional visualization clustering tools are based on physical force-directed model [5], such as Spring and Pajek. This type of models aim to draw a beautiful and readable visual graphs and are not designed for clustering. In the visualization of these models, the central nodes which usually have high degrees are put in the middle of graphs, while the nodes which have low degrees are put around

the central nodes. As a result, longer edges are cut down and nodes with dissimilar traits can't be assigned into different groups. Obviously, the wrong results might be produced if the traditional graph clustering tools are used to discover groups of technology competitors.

LinLog algorithm was first put forward by Noack in 2007 [6]. It is based on force-repulsion model which is designed to produce ideal graphs of visualization clustering. This type of models will group the nodes connected tightly and separate the nodes connected sparsely.

Fig. 1 is the compare results of two graph clustering models, Spring model and LinLog model, which is detailed described in the paper of Noack.



(a) Spring model (b) LinLog model

Fig. 1. Comparison of Spring and Linlog methods

Fig. 1 shows that LinLog model can clearly separate the nodes into two groups by the intermediary points, Dan and Upton, while Spring model can't do that. In this research, LinLog graph clustering method based on Barnes and Hut hierarchy algorithms is employed to divide competitors into several clusters and draw the clustering graphs.

Term extraction method

The earliest research of term extraction was conducted by British scientist Firth, who proposed the context theory in 1957 which emphasized the importance of the context [7]. Up to date, researchers have done many researches and proposed many kinds of automatic term extraction methods. These methods generally can be classified into three categories: 1) Linguistic rule method; 2) Statistical method; 3) Hybrid method.

Linguistic rule method

The methods utilized the information of lexicality and syntax to identify terms by analyzing the special syntax structure of terms context [8]. Frantzi [9] presented a terms extraction method which used the nouns, verbs and adjectives of terms context to improve the accuracy of terms extraction.

Linguistic rules provide a simple method to recognize the term, however they mainly depend on the prior knowledge of human and usually it is difficult to find rules. Especially for open corpora, the styles of word formation are very flexible, so the linguistic rules may not work well because the linguistic rules need to be change very fast to satisfy the situation.

Statistical method

The linguistic rule methods highly depend on the corpora, which bring restrictions to use discovered rules into other corpora and to improve the accuracy of term extraction to a higher level. Therefore, researchers begin to seek some new methods. Statistical methods were presented and used in 1980s, Tseng put forward a method to extract keywords and phrases [10].

These methods need less manual intervene and have better applicability and adaptivity, such as mutual information, the Log-likelihood, Chi-squared and Z-score method. They are independent of the corpus and dictionary and can be used into patent analysis, but these models are usually complex.

Hybrid method

The hybrid methods were proposed integrate the advantages of above two methods by later researchers. For example, Frantzi [11] proposed the C-value and NC-value method which aimed to extract the term more efficiently and accurately, the experiment results proved the good performance of the hybrid extraction method.

Method

Generally, two organizations or countries are mostly likely to be technology competitors if they have a relatively high number of similar patents [12]. Based on this point, a clustering analysis method is proposed to identify groups of technology competitor by calculating similarities of their patents. The process of discovering technology competitor groups is described as follows:

Firstly, an appropriate clustering level is selected from R&D institutions, Provinces or Countries. This is determined by the analysis purposes. Secondly, technology terms are extracted from patent texts and each patent is represented by a feature vector based on vector space model. Thirdly, the similarity matrix of patents is established by calculating the similarity between each pair of patents. Fifthly, LinLog graph clustering algorithm is invoked to discover competitor groups from the created network. The clustering results will be displayed in visualization graphs.

Each patent, say d in document set (d_1, d_2, .., d_n), is represented as a vector with terms as features. Denote d_i=(t_{1i}, t_{2i}, ..., t_{mi}) and d_j=(t_{1j}, t_{2j}, …., t_{mj}), where t_{k*} is the feature value of term t_k. TF-IDF metric is selected to measure the feature value. The similarity of patent di and dj, $sim(d_i, d_j)$, is defined with cosine similarity (Equation 1).

$$sim(d_i, d_j) = \cos\theta = \frac{\sum_{k=1}^{m}(t_{mi} \times t_{mj})}{\sqrt{\sum_{k=1}^{m} t_{mi}^{\ 2}} \times \sqrt{\sum_{k=1}^{m} t_{mj}^{\ 2}}} \quad (1)$$

Definition 3.1 Similar Patent: given a patent similarity threshold ε, d_i and d_j are called similar patent if their similarity $sim(d_i, d_j) \geq \varepsilon$.

Definition 3.2 The number of similar patents of each pair of competitors: Denote P_A and P_B are patent set of competitor A and B respectively. Without loss of generality, let $|P_A| \leq |P_B|$. Given the similarity threshold ε, the numbers of similar patents of P_A and P_B is the number of similar patents of competitor A and B.

Experiments

The Chinese patent data of fuel cell was collected for experiment from the official website of the State Intellectual Property Office of China. In order to obtain the patent data quickly, a patent data acquisition system [13] was employed which could download the description information of patent automatically and storage them into the local database. Totally, 6346 patents were collected via this system.

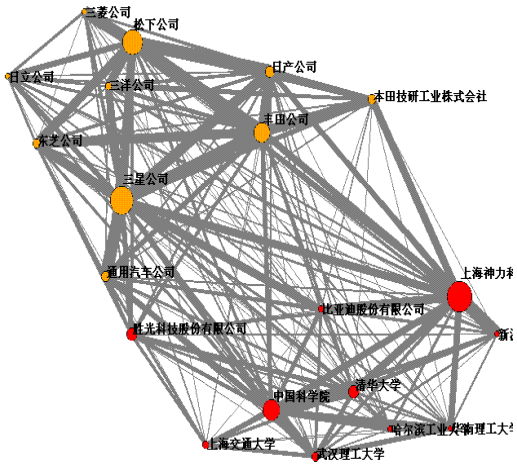
In order to count the number of similar patents, the patent similarities of each pair of competitors need to be calculated firstly under the selected levels. According to the test results of training data set, the similarity threshold was set as 0.6. Moreover, LinLog graph clustering tools was used to cluster and visualize the groups of technology competitors. In the visual graphs, the node size denotes the number of valid granted patents of technology competitors, while the node color denotes different groups of technology competitors, and the edge width denotes the number of similarity patents between two technology competitors.

The experimental results of three clustering levels are shown as below.

R&D Institution Level

By counting the numbers of valid granted patents, top 20 patent assignees (the first patent assignee) are chosen for graph clustering analysis in R&D institutions level. The results are shown in Fig. 2. Table 1. shows corresponding English names of Chinese Names in Fig. 2.

These patent assignees were clustered into two technology competitor groups. Red nodes denote the first group and Orange nodes denote the second group. It is funny to note that the assignees of the first group come from China and the assignees of the second group all come from abroad.



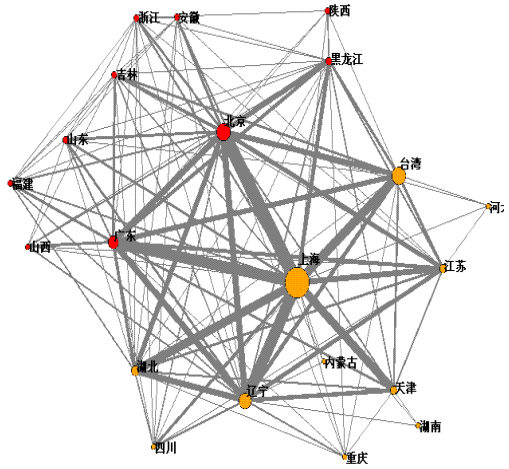


Fig. 2. Clustering results of R&D Institutions level Fig. 3. Clustering results of Provinces level

Table 1. Corresponding English Names Of Chinese Names Of R&D Institusions In Fig. 2

Chinese Name	上海神力科技有限公司	中国科学院	胜光科技股份有限公司	清华大学	武汉理工大学	上海交通大学	比亚迪股份有限公司
English Name	Shanghai Shen-Li High Tech	Chinese Academy of Sciences	Antiq	Tsinghua University	Wuhan University of Technology	Shanghai Jiaotong University	BYD
Chinese Name	新源动力股份有限公司	哈尔滨工业大学	华南理工大学	三星公司	松下公司	丰田公司	日产公司
English Name	Sunrise Power	Harbin Institute of Technology	South China University of Technology	Samsung	Panasonic	Toyota	Nissan
Chinese Name	通用汽车公司	本田技研工业株式会社	东芝公司	三洋公司	三菱公司	日立公司	
English Name	General Motors	Honda	Toshiba	Sanyo	Mitsubishi	Hitachi	

Table 2. Corresponding English Names Of Chinese Names Of Provinces In Fig. 3

Chinese Name	北京	广东	黑龙江	山东	安徽	吉林	浙江	陕西	福建	山西	上海
English Name	Beijing	Guangdong	Heilongjiang	Shandong	Anhui	Jilin	Zhejiang	Shanxi1	Fujian	Shanxi2	Shanghai
Chinese Name	台湾	辽宁	湖北	江苏	天津	重庆	四川	湖南	河北	内蒙古	云南
English Name	Taiwan	Liaoning	Hubei	Jiangsu	Tianjin	Chongqing	Sichuan	Hunan	Hebei	Neimenggu	Yunnan

Province Level

All 22 provinces were chosen for graph clustering analysis in provinces level. The clustering results are shown in Fig. 3. Table 2. shows corresponding English names of Chinese Names in Fig. 3.

Two groups of technology competitors in province level are identified. Red nodes denote the first group and orange nodes denote the second group. Yunnan province don't appear on the graph in that it hasn't similar patent with any other provinces under the given similarity threshold.

Country Level

All 22 countries in developing fuel cell technology are chosen for graph clustering analysis in country level. The clustering results are shown in Fig. 4. Table 3. shows corresponding English names of Chinese country names in Fig. 4.

Two groups of technology competitors are discovered. Red nodes denote the first group and orange nodes denote the second group.

Fig. 4 shows that China, Japan, America, Korea and German have much more similar patents than other pairs, indicating that these 5 countries are the core technology competitors of fuel cell field in China.

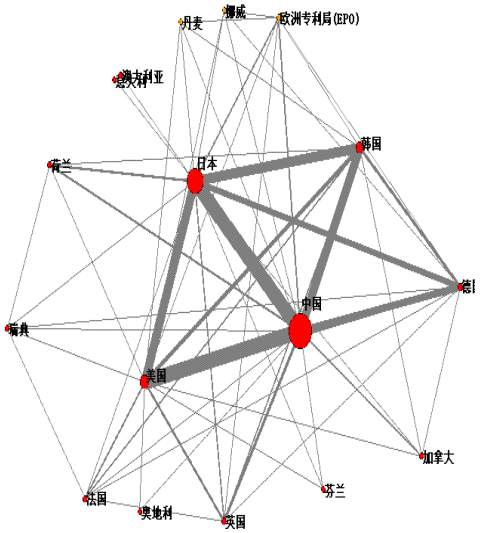


Fig. 4. Clustering results of Countries level

Table 3. Corresponding English Names Of Chinese Names Of Countries In Fig. 4

The Chinese Name	*The English Name*
中国	China
日本	Japan
美国	America
韩国	Korea
德国	Germany
英国	Britain
法国	France
瑞典	Sweden
荷兰	Netherlands
澳大利亚	Australia
加拿大	Canada
芬兰	Finland
意大利	Italy
奥地利	Austria
欧洲专利局	EPO
挪威	Norway
丹麦	Denmark

Conclusion

In this paper, a graph clustering method of technology competitor based on patent text terms is proposed. Three levels can be selected depending on the analysis purpose. The Chinese patent data were collected by a patent data acquisition system to obtain research dataset. The LinLog graph clustering tool is used to cluster technology competitors into different strategy groups and to display the experiment results in visualization mode. The experiment results on fuel cell domain testify the effectiveness of the proposed method in clustering technology competitors.

Acknowledgements

This work is mainly supported by National Natural Science Foundation of China (Project 71473237), and partially supported by Innovation Foundation of the Institute of Scientific and Technical Information of China (Project MS201703), and The Program of the China Knowledge Centre for Engineering Science and Technology (CKCEST-2017-1-12). Authors are grateful to National Natural Science Foundation of China, Ministry of Science and Technology of China and Chinese Academy of Engineering for financial support to carry out this work.

Reference

[1] C. Huang, Z. Zhang, and B. Lu, "An analytical frame of industry competition from the viewpoint of input-output," Science & Technology Management Research, 2015.

[2] C. K. Lee and R. Ong, "An analysis of the liquid crystal cell patents of lg and samsung filed at the uspto," 2006.

[3] G. Liu, J. Wu, H. Wang, and D. O. Management, "On the industry competition pattern analysis method based on patent technology association," Journal of Intelligence, 2014.

[4] F. Beck, S. Koch, and D. Weiskopf, "Visual analysis and dissemination of scientific literature collections with survis," IEEE Transactions on Visualization & Computer Graphics, vol. 22, no. 1, pp. 1–1, 2016.

[5] W. Li, P. Eades, and N. Nikolov, "Using spring algorithms to remove node overlapping," pp. 131–140, 2005.

[6] A. Noack, "Energy models for graph clustering.," Journal of Graph Algorithms & Applications, vol. 11, no. 2, pp. 453–480, 2007.

[7] H. Han, D. Zhu, and X. Wang, "Technical term extraction method for patent document," 2011.

[8] H. E. Yan, Z. F. Sui, H. M. Duan, and Y. U. Shi-Wen, "Term mining combining term component bank," Computer Engineering & Applications, vol. 42, no. 33, pp. 4–7, 2006.

[9] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms:. the c-value/nc-value method," International Journal on Digital Libraries, vol. 3, no. 2, pp. 115–130, 2000.

[10] Y. H. Tseng, C. J. Lin, and Y. I. Lin, "Text mining techniques for patent analysis," Information Processing & Management, vol. 43, no. 5, pp. 1216–1247, 2007.

[11] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms:. the c-value/nc-value method," International Journal on Digital Libraries, vol. 3, no. 2, pp. 115–130, 2000.

[12] S. Lee, B. Yoon, C. Lee, and J. Park, "Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping," Technological Forecasting & Social Change, vol. 76, no. 6, pp. 769–786, 2009.

[13] L. Ruotsalainen, "Data mining tools for technology and competitive intelligence," VTT Tiedotteita - Valtion Teknillinen Tutkimuskeskus, no. 2451, 2008.