Mining Technical Topic Networks from Chinese Patents

Hongqi Han
Institute of Scientific and Technical Information of China
Fuxing road 15, haidian district(100038)
Beijing, China
bithhq@163.com

Shuo Xu
Institute of Scientific and Technical Information of China
Fuxing road 15, haidian district(100038)
Beijing, China
xush@istic.ac.cn

Lijun Zhu
Institute of Scientific and Technical Information of China
Fuxing road 15, haidian district(100038)
Beijing, China
zhulj@istic.ac.cn

Xiaodong Qiao
Institute of Scientific and Technical Information of China
Fuxing road 15, haidian district(100038)
Beijing, China
qiaox@istic.ac.cn

Jie Gui
Institute of Scientific and Technical Information of China
Fuxing road 15, haidian district(100038)
Beijing, China
guij@istic.ac.cn

Zhaofeng Zhang
School of Information Management, Nanjing University,
22 Hankou Road, Nanjing Jiangsu (210093)
Nanjing,China
zhangzf@istic.ac.cn

ABSTRACT

Patents are one of the most important innovative resources. It is a challenge and useful to discover technical topics and their relations from patents. A process framework is proposed to mine technical topics and construct their relation network from Chinese patents. The process consists of four stages. First, technical terms are extracted from patent texts and the equivalence index is selected to measure the link strength between them. Then, a clustering algorithm is used to group terms into topic clusters, in which terms are connected by internal links, and topic clusters are connected by external links. Afterwards, all topics are classified into three categories: isolated, principal and secondary. Finally, a technical topic network is created by using topic clusters as nodes, external links as edges and the number of external links as weights. Experimental results on Chinese fuel cell patents show the method is effective in mining technical topics and mapping their relations, and the constructed network is helpful for technology innovation.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Data Mining ; I.2 [**Computing methodologies**]: Artificial Intelligence

General Terms

Application

Keywords

Technical topic network, topic relation, co-word analysis, patent analysis, data mining

1. INTRODUCTION

Today, along with the rapid development of science and technology and integration of economic globalization process, innovation is becoming an important means to obtain technological advantage[8]. Patent documents are one of the major innovative data resources of technical and commercial knowledge, and thus patent analysis has long been considered helpful for R&D management and technoeconomic analysis[14]. By depicting technical topics and mapping their relations, researchers can acquire novel ideas for technology breakthrough, while enterprises can find technical routes for product planning and development, and policy makers can understand dynamic technology change for funding emerging and potential fields. Traditionally, a small number of experts are selected to undertake such work, yet the method has been widely criticized, such as weak representativeness, high cost, and low efficiency [5].

It is a challenge to detect technical topics and find the relations between them. On the one hand, rapid developing technology makes it difficult for researchers to grasp the latest topics, on the other hand the amount of patents is huge and increasing sharply, which also makes it difficult to mine technical topics hidden in the data. Yoon [15] and Lee [9] proposed approaches for identifying new technology opportunities using keyword-based morphology analysis and keyword-based patent maps respectively. Yoon [14] presented a network analysis for high technology trend forecast based on text mining technique, where nodes of the network are patents. However, these previous researches didn't explore technology topics and map their relations. Callon [2] presented co-word analysis techniques to map the relationship between concepts, ideas and problems in science. The following researches, for example, Coulter [4], Van [12] and Cobo [3], extended the technique. Now it is common to find scientific papers and reports that contain a science mapping analysis to show and uncover the hidden key elements [3], however most of these works were undertaken for academic purposes using bibliographic data, and few are for competitive animus using patent data.

In this article, we propose an approach based on co-word analysis technique for detecting technical topics and mapping their relations using patent data. The co-word analysis technique is

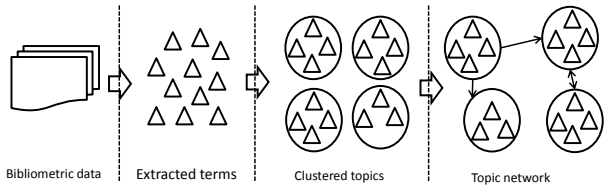

Figure 1: Framework

based on keywords and their co-occurrence, however most patent databases don't provide keywords, therefore one of the challenges is to extract technical terms from patents. To extract terms from patent text, we use a hybrid automatic term recognition technique integrating linguistic rules and statistics indexes. Another challenge comes from the way to detect topics in patents, because technical topics and their numbers are usually unknown. The presented approach is a process framework consisting of four steps. The first step is data collection and pre-processing, the second step is extracting technical terms, the third step is detecting technical topics, and the last step is constructing technical topic network.

The remainder of the article is organized as follows. In section 2, the process framework is illustrated to introduce the basic idea to create topic network. In section 3, the main techniques used in the article are introduced. In section 4, an experimental results on Chinese fuel cell patents are described and discussed. Finally, the conclusion are made.

2. METHODS

The process framework of the presented method is first introduced. Then three core techniques in the framework are introduced, including technical terms extraction, topic detection, and topic network construction.

2.1 Framework

The process framework contains four continuous stages (Fig.1): patent data collection and pre-processing, technical terms extraction, topic detection and topics network construction. In the first stage, patent data are collected and stored into database after preprocessing operations. In the second stage, technical terms are extracted from patent text. In the third stage, terms are clustered into technical topics. In the last stage, topic clusters are used to create network based on their link relations.

2.2 Technical Terms Extraction

Fig.2 depicts the overall process of extracting technical terms from patents. The term extraction process integrates linguistic rules and statistics index. Because there is no delimiter between Chinese words like the space character in English text, word segmentation is necessary at first. Then POS tagger is used for identifying part-of-speech of words, e.g. nouns, verbs or adjectives, etc.. Afterwards, we collocated words into phrases. The collated phrases are filtered by linguistic rules and stop words for generating candidate terms. The linguistic rules used in the article are shown in Table 1. The letters in the second column of Table 1 are codes of part-of-speech. These codes are defined in the Chinese segmentation tools developed by Hailiang Technology Company. The meanings of these letters are shown in Table 2.

Then statistics index are used to compute the termhood and unithood of the candidate terms. Termhood refers to the degree that a linguistic unit is related to domain-specific concepts, while

Table 1: Linguistic Rules for Extracting Chinese Terms

Length of Term	Rule
1	n,v,l
2	nv+nv,a+nv,b+n, m+n
3	nv+nv+nv, a+nv+n,d+v+n,b+v+n
4	nv+nv+nv+n
5	nv+nv+nv+nv+n,a+n+v+n+n, b+v+n+v+n
6	nv+nv+c+vn+nv+n, nv+nv+nv+c+nv+nv, n+n+u+b+vn+n, n+vn+u+n+vn+n

Table 2: Meaning of Letters in Table 1

Code	Part of speech
a	adjective
b	determiners
c	conjunction
d	adverb
l	idiom words
m	quantifier
n	noun or product terms
v	verb
u	auxiliary words
nv	noun or adverb

unithood refers to the degree of strength or stability of syntagmatic collocations[7]. Afterwards, the candidate terms are sorted according to the statistics index and evaluated by domain experts, and finally, selected technical terms are stored for co-word analysis.

2.3 Topics Detection

After technical terms are extracted, keyword network can be constructed based on their co-occurrence relation, however such network contains so many nodes and complex relations that it can't be better understood [2]. Therefore, Callon [2] presented a method to cluster keywords into topics, in which several keywords are closely connected. Each topic represents an interested problem of researchers, and so it will be more easily understood than single keywords. Moreover, the number of topics is far less than that of keywords, which makes it clearer to map the relations of concepts.

To measure the link strength is an important process for detecting topics. Many metrics have been proposed for computing the link relations between keywords. The common indexes include association strength [4, 12], Equivalence Index [2], Inclusion Index[5], Jaccard Index [10], and Salton' s cosine [10]. Among these indexes, the Equivalence index shows the probability that two keywords co-occur when given the frequency of two keywords appearing in documents. It provides an intuitive measure of link strength between keywords, rather than imposing conceptual inclusion property like other metrics. Moreover, the metric is easier to be understood and utilized in the production and interpretation of keyword association maps than other metrics [4]. It also allows associations of both major and minor keywords and is symmetrical in their relationships [2]. Let C_i be the number of times keyword i is used in the corpus, and let C_{ij} be the number of co-occurrences of keyword i and j. The link strength E_{ij} between keyword i and j is given by Eq. 1:

$$E_{ij} = (C_{ij}/C_i) \times (C_{ij}/C_j) = C_{ij}^2/(C_i \times C_j) \qquad (1)$$

Based on link strength, research topics in a domain corpus can be detected using keywords clustering algorithm. The main

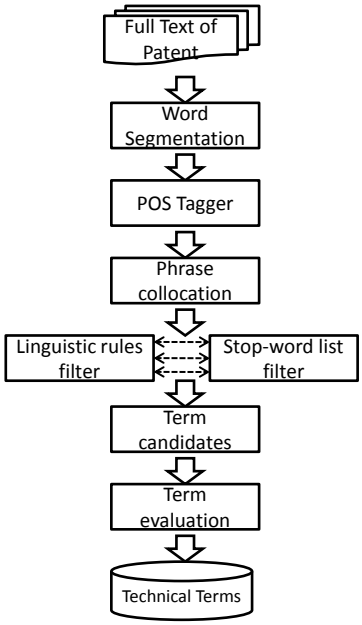

Figure 2: Term extraction

effective clustering algorithms include Callon's method [2], Coulter's method [4], Multidimensional Scaling (MDS) [13], Latent Dirichlet Allocation (LDA) [1] and others.

This study uses the two passes algorithm proposed by Coulter [4]. Pass-1 builds keywords clusters that can identify areas of strong focus as research topics. The nodes with big circular shape in Fig.3 show such topics. The internal nodes with triangle shape in a topic node represent strongly connected keywords. The links between keywords in a topic are called internal links. Pass-2 identifies keywords that associate in more than one topics, and thereby generates links between Pass-1 nodes across topics and indicate pervasive issues. The links between keywords in different topics are called external links (Fig.3).

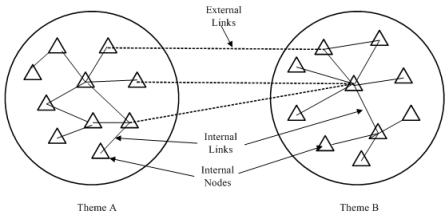

Figure 3: Topic links

2.4 Topics Classification

Denote the set of detected topics as $T = \{t_1, t_2, ..., t_m\}$, where m is the number of topics. Then for a topic, t_l, where $l \in [1, m]$, denote the equivalence index of internal link between keywords k_i and k_j as E_{ij}, where $i, j \in [1, n]$, and n is the number of keywords in topic t_l. Learning from Callon [2] and Coulter [4], we use Eq.2 and Eq.3 to define two indexes: ceiling and saturation. For each topic, ceiling measures the maximum link strength (Eq.2), and saturation measures the minimum link strength (Eq.3).

$$ceiling(t_l) = max(E_{ij}) \tag{2}$$

$$saturation(t_l) = min(E_{ij}) \tag{3}$$

Next, considering the external links and their association values, all the topic clusters can be classified into three categories: isolated, secondary, principal.

- isolated topics: which have no external links with other topics, or the numbers of external links between which and other topics are below threshold, so the only question regarding them is their internal homogeneity;
- secondary topics: the strength values of external links between which and other clusters are above the ceiling threshold, and so it is naturally considered that they are the extension of one of these;
- principal topics: whose saturation values are greater than links associated to one or more other (secondary) clusters.

According to such classification, principal topics seem to be basic technologies for some other ones, and secondary topics seem to be dependant technologies on basic ones, while isolated topics seem to be independent technologies.

2.5 Topics Network Construction

Based on the classification of topic clusters, using detected topics as nodes, the external links as edges, and the numbers of external links as weights of edges, the topics network is constructed to illustrate the relations between topics. We don't use multiple edges to represent the relations between two topics. That is to say, if two topic clusters have external links, even when the number of external links are greater than 1, there is a single edge between them. In practice, a threshold of minimum number of external links is used to remove weaker connected edges for getting better results. In order to illustrate the relation between two topics, the classification information is used to decide the direction of edges. The direction of edges between principal and secondary topics are unidirectional, from the former to the latter, while the edges between two principal topics are bidirectional.

3. EXPERIMENTS

3.1 Data

The experimental data is provided by SIPO (http://www.sipo.gov.cn). Chinese patents in the domain of fuel cell are collected using the retrieval strategy combining keywords and IPC codes. All collected patents are pre-precessed. Finally, we get 6,346 patents. Because the full text of patents are not provided, we just use title and abstract to extract terms in the experiment.

3.2 Technical terms

First, Chinese word segmentation tools are run to split sentences in patent title and abstract into words. Let the threshold of term frequency be 2, we get 28,113 candidate terms. All single words are eliminated, because single words alone are often too general in meanings or ambiguous to represent a concept in patent analysis, while multi-word phrases can be more specific and desirable[11]. Then the termhood and unithood of all candidates are computed

Table 3: Parameters Used to Generate Network

Parameter	Value
Minimum concurrence Number	2
Maximum Node Number in a Topic	20
Maximum Link Number	24

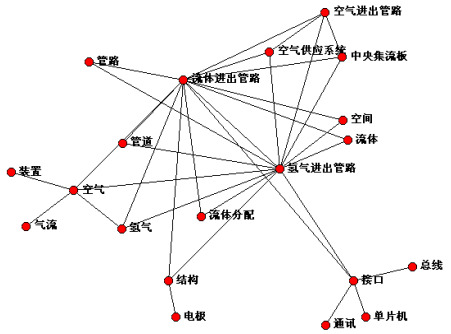

Figure 4: The internal structure of the first topic

using the methods in [6]. Afterwards, let the threshold of termhood and unithood be their mean value, and the threshold of document frequency be 5, we get 1,669 technical terms. Finally, 1,123 terms are selected after the evaluation process of domain experts for detecting topics and creating network.

3.3 Technical Topics

With the extracted 1,123 technical terms, we use the method in [4] to detect topic clusters. The parameters used to generate topic clusters are shown in Table 3.

We get 62 topics totally. All the topics are numbered, ranging from 1 to 62 according to the generation sequence. The first generated topic is assigned number 1 and the last one is assigned number 62. The name of each topic is the internal terms with high degrees. Fig. 4 shows the internal structure of first detected topic cluster, the name of which is "氢气进出管路-流体进出管路" (Hydrogen outlet pipeline-liquid outlet pipeline).

3.4 Network

Detected topics are used to construct networks (Fig. 5). In the network, nodes are detected topics, edges are external links between them. Isolated topics are not shown. With the information provided by the network, we can not only understand the relation between topics but also find out the structure of domain technology.

In Fig. 5, the value of parameter Minimum External Links is set 4, i.e. only when the number of external links between any two topics are greater than 4, there is an edge between them. Under such condition, there are 10 sub-domain technology. Each sub-domain technology is composed of several connected topics. In a sub-domain technology, the importance of each topic is different. For example, in the sub-domain technology which contains topic 2, topic 24 is the joint of topic 2, 22, 24, 30, 39 and 48, so it may play an essential role in the transformation of the network. Such topics are called crossroads clusters. By identifying them, we can find the important technology in the domain.

In Fig. 5, the arrow direction of an edge shows the extension relation between two topics. The topic nodes in the heads of arrows are secondary clusters, while the topic nodes in the tails are principal clusters. As stated before, the secondary clusters are

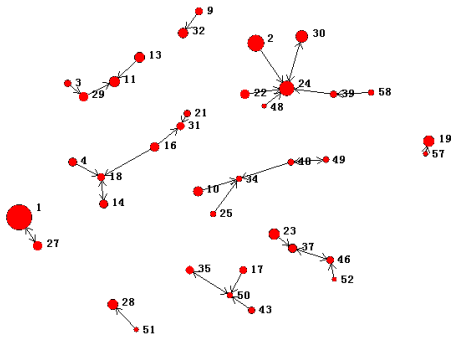

Figure 5: Technical topic network of fuel cell

extensions from principal clusters. In the figure, the sizes of nodes represent the patent numbers related to a topic. If a term in a topic occurs in a patent, the patent is related to the topic. Therefore, from the figure, we know topic 1 is the most preferred developed technology in the domain of fuel cell. This gives useful information to find the popular technologies in the domain.

4. CONCLUSION

A method based on co-word analysis technique is presented to detect domain research topics and their link relations from Chinese patents. Because keywords are not provided in patent data, the method extracts terms from free text data in title and abstract. The term extraction technique integrates linguistic rules and statistics indexes. Extracted terms are clustered into topic clusters based on equivalence index. Internal links and external links are defined to classified all topic clusters into three categories, viz. isolated, secondary and principal clusters. Using topic clusters as nodes, external links as edges, the number of external links as weights, the technical topic graph is created. Experimental results on fuel cell patents show that it can map the relation of topics, and find important research topics.

Although the method is designed for Chinese patents, it is also applicable for other patent data, like USPTO and EPO. However, the discovered topics in the method are based on links, and we limit the number of keywords in clustered topics. In addition, threshold values, such as document frequency and maximum external link number in the experimental part is too naive. These human factors will affect the clustering result, and maybe the topic clusters can not cover relative technical terms. In the future, we will try more specific methods to detect research topics for generating network, such as topic models based on statistics technology. In the theory of co-word analysis, it is difficult to evaluate the accuracy of topic selection and the effectiveness of topic network. Although we believe researchers will be inspired with the topic network in technology innovation, the reliability of the method should be considered in the future.

Acknowledgments

The authors are grateful to Hailiang Technology Company for providing the Chinese word segmentation software for this research. This research was funded partially by "The study on the disconnected problem of scientific collaboration network" which is sponsored by ISTIC Pre-research Foundation under grant number YY–201418 , the Key Technologies R&D Program of Chinese 12th Five-Year Plan (2011-2015): Key Technologies Research on Data

Mining from the Multiple Electric Vehicle Information Sources under grant number 2013BAG06B01, and Key Technologies Research on Mining and Discovery from Patent Resources under grant number 2013BAH21B02. Authors are grateful to the Ministry of Science and Technology of China for financial support to carry out this work.

5. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] M. Callon, J. P. Courtial, and F. Laville. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientmetrics*, (22):155–205, 1991.

[3] M. Cobo, A. L ó pez-Herrera, E. Herrera-Viedma, and F. Herrera. Scimat: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, 63(8):1609 – 1630, 2012.

[4] N. Coulter, M. Ira, and K. Suresh. Software engineering as seen through its research literature: a study in co-word analysis. *Journal of the American Society for Information Science*, 49(13):1206–1223, 1998.

[5] Q. H. Knowledge discovery through co-word analysis library trends. *Library trends*, 48(1):133–159, 1999.

[6] H. Han and X. An. Chinese scientific and technical term extraction using c-value and unithood measure. *Library and Information Service*, 56 (19):85–89, 2012.

[7] K. Kageura and B. Umino. Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289, 1996.

[8] H. Lee and D. H. Technological innovation of high-tech industry and patent policy-agent based simulation with double loop learning c,intelligent agents:specification,modeling and applications. In *Proceedings of 4th Pacific Rim International Workshop on Multi-agents,PRIMA*, 2001.

[9] S. Lee, ByungunYoon, and Y. Park. An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, 29:481–497, 2009.

[10] H. Peters and A. F. van Raan. Co-word-based science maps of chemical engineering. part i: Representations by direct multidimensional scaling. *Research Policy*, 22(1):23–45, 1993.

[11] Y.-H. Tseng, C.-J. Lin, and Y.-I. Lin. Text mining techniques for patent analysis. *Information Processing & Management*, 43(5):1216–1247, 2007.

[12] N. J. Van Eck and L. Waltman. Bibliometric mapping of the computational intelligence field. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 15(05):625–645, 2007.

[13] D. Ying, C. G. G., and F. Schubert. Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management*, 37(6):817–842, 2001.

[14] B. Yoon and Y. Park. A text-mining-based patent network:analytic tool for high-technology trend. *The Journal of High Technology Management Research*, 15(1):37–50, 2004.

[15] B. Yoon and Y. Park. A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. *Technological Forecasting & Social Change*, 72:145–160, 2005.