International Journal of Hybrid Information Technology
Vol.8, No. 5 (2015), pp. 207-214
http://dx.doi.org/10.14257/ijhit.2015.8.5.23

Research on Patent Information Analyzing and Predicting System Based on Data Mining

Yang Liu

Harbin University of Science and Technology, Harbin, China
liuyang1576@126.com

Abstract

Patents, as technological innovation achievements and sign, are potential and intangible assets of enterprises and can enhance their core competitiveness. How to obtain useful information and knowledge from a large number of patents data, have become an urgent problem. In order to satisfy the actual needs of enterprises for patent information, the paper proposed a frame of patent information analyzing and predicting system based on data mining. We designed the function structure of the system which consists of patent data preprocessing, patent data mining and patent mining results visualization. We also introduced the implementation method of each part in the system. By using the designed patent information mining and predicting system, enterprises can automatically gained the required patent information and improve their core competitiveness.

Keywords: *patent analysis, patent trend predicting, data mining, text mining, patent maps*

1. Introduction

Today, the world is undergoing a transition to a knowledge-based economy from an industrial economy, and knowledge as a high value of intangible assets is becoming the dominant society and changes the pattern of the world. Patents, as technological innovation achievements and sign, are both the commanding heights of business proprietary technology and potential and intangible assets which enhance their core competitiveness. A report from World Intellectual Property Organization (WIPO) [1] shows that: patent information is one of the world's largest open technology sources, which contains technical information on the world of 90% to 95%, and the public than other technical information carriers as early as 1 to 2 years. According to another report from WIPO, the patent is the most powerful tool in all research and development activities and the economic value of patents is over 90 % in average output of research and development in the world, comparison with other activities. Therefore, especially in the knowledge economy, patent information has a pivotal role for the country and businesses, and to maximize the development and use of patent information is a national and an important competitive advantage of enterprises guarantee.

Internationally, Europe, America and other developed countries look the patent data as an important source of information and have established their own patent database and related patents Citation Index database. For example, UK Derwent Company established a Patent Citation Index database in 1995, which is the first patent information database [2]. In 1998, Thomson Science And Technology Co., Ltd developed Derwent INN0VAT10NS INDEX Database which included 9 million basic invention and 18 million patents from more than 40 patent authorities worldwide dating back to 1963. In America, MICROPATENT Company developed CAPS database which collected all US patent citation information since 1975 [3]. European Patent Office (EPO) developed EPOQUE system which also included patent citation database REFI [4].

ISSN: 1738-9968 IJHIT

With the rapid growth of patent information data amount, the traditional patent information retrieval and analysis tools could not meet the actual needs of patent information analysis. The rapid development of computer science and technology promote the transform of the patent information analysis method from the traditional previous text analysis, simple statistical analysis to data mining, artificial intelligence, and information visualization techniques [5]. Deeply mining implied rules in patent literature and patent information can provide a reliable basis for intelligence support and decision making for e technology innovation management.

2. Related Works

With the rapid development and extensive application of computer technology and network technology, a large number of data, even vast amounts of data, have been accumulated from every fields, and a "data explosion, lack of knowledge" phenomenon has appeared. In order to effectively solve the contradictory of the data rich and poor knowledge, data mining technology emerged. Data Mining [6] is a technique which extracts and mines potential and useful information and knowledge that implies in a lot of, incomplete, noisy, fuzzy data from practical applications, and these information and knowledge is generally that people do not know in advance. At present, there are many data mining methods for different applications, and the commonly used methods include association analysis, classification and prediction, clustering analysis, Outlier analysis, trend analysis and evolution, *etc.*

Patent information mining is the process that data mining or machine learning technique in computer field are used in patent analysis to mining patent information from the patent literature databases and predict patent trends. Before making research and development project and carrying out business activities, enterprises firstly need retrieve intellectual property information and process this information in deep-level to mining the potential of information. So that enterprises can form intellectual property rights and improve the abilities of transform innovation into intellectual through original innovation, integrated innovation and the introduction of absorption and innovation. So, whether it is from the perspective of technological innovation itself, or from the perspective of technical and trade risk aversion, there is a pressing need to support patented technology. The development of patent information analyzing and predicting platform based on data mining will help enterprises better use intellectual property technology to enhance the capability of independent innovation. Through a comprehensive and rigorous analysis of intellectual property information, the patent platform can also help enterprises fully learn and utilize the existing achievements, reasonably circumvent patent rights of others, share market, save development time and money.

Currently, many scholars began to try to apply data mining techniques to patent information analysis and forecast. Meng [7] *et. al.* use data mining method to research patent trend to explore business competitive intelligence in depth. Yuan [8] *et. al.* explained the advantages of using data mining technology for patent intelligence analysis, and confirmed the feasibility and effectiveness of this method by using it for intelligence analysis in Zhongguancun Science Park in China. Their task explores a new direction for patent intelligence analysis. Ma [9] *et. al.* Use data mining techniques to the patent information analysis in order to discover the knowledge of interest to the user, and make it into effective competitive intelligence. Trippe proposed Patinformatics (the discipline analysis of patent data to uncover relationships and trends) concept by Bioinformatics inspired, and this make patent information research into a new phase [10]. Mogee *et. al.* Studied patent analysis and application in different fields [11]. Breitzman studied specific application of patent citations analysis in the enterprise evaluation and merger plan selection [12].

There is also some succeeded patent mining software being developed. For example, Derwent patent analysis software [2], developed by the world-renowned Derwent company based Derwent patent map theory, utilizes data mining and visualization tools to implement automated analysis of patent information and management. "Intelligent Miner for Text" (IM4T) software [13], developed by IBM, provides us with the characteristics retrieval, clustering, citation analysis of patent information and other functions, which is one of the leadership patent analyzing software. China's State Intellectual Property Office developed an intensive management system platform which can implement patent information collection, information processing, information retrieval, information analysis, information applications. Through the complete value chain system, the patent information service platform can transform the patent literature into the valuable patent information [14]. Beijing East Linden introduces and optimizes its intelligent semantic analysis retrieval system, which has a semantic analysis, concept retrieval, automatic classification, automatic indexing and automatic translation function, and it greatly improved the efficiency of information retrieval and retrieval of massive data. It also implements conceptual retrieval functions, in addition to search by entering a keyword, you can take similarity retrieval by any length lengthy statement even literature (including patents, scientific literature, reports, news) [15].

In summary, patent information mining gained more and more attention, whether qualitative or quantitative research, many scholars have achieved certain results. However, most studies about patent information utilities currently stuck in the quantity characteristic statistics of patent data, qualitative analysis and computer management of patent information, few software could automatic mining patent law and patent trends in knowledge content. In this paper, we designed a patent information analyzing and predicting system based on data mining technology, which is a comprehensive patent information analysis and forecasting application platform. The proposed patent system is capable to identify an effective, innovative, potentially useful and ultimately understandable knowledge. So, By means of this platform, enterprises and researchers can more easily reveal interrelated relationship from a large of patent literatures, in order to make technology research and development innovation, patent evaluation and assessment, patent asset management and patent protection (including patent infringement warning, patent litigation and patent border protection, *etc.*)

3. System Design and System Structure

3.1. System Function Analysis

According to the actual needs of patent analysis, the main function of patent analyzing and predicting system based on data mining technology should include the following aspects.

3.1.1. Strategic Management Functions

Through the depth mining of patent data, the system can help users understand the entire state as well as technical changes in specific technical fields recognizes technology events, trends and research developments, and make the strategic decision to develop technology plans and programs and macro management.

3.1.2. Patent Early Warning Functions

Patent warning mechanism refers to the rapid response to emergencies patent disputes and likely happened patent disputes. The basis of patent early warning is patent information retrieval and analysis. By gathering and analyzing patent information in related techniques field, the system can forecasting and publishing the patent dispute and the proposed measures to deal with it.

3.1.3. Dynamic Tracking Functions

The system can help users monitor, track and analyze patent information of main competitor, insight into competitor's strategic intent, understand its technical characteristics and development. It also helps people analyze patent information of major export products country to understand patent protection of the products and related technology to help companies make strategic decisions.

3.1.4. Trends Predicting Function

On the one hand, the system can track, collect, and analyze the situation and the impact of competitors' patent technology in the enterprise market; On the other hand, it can track and collect licensed, legal status and product development implementation of patented technology of competitors. Further, it can help analysis technology trends of competitors and propose countermeasures.

3.2. System Structure Design

This project mainly uses data mining methods, including cluster analysis and association analysis, to analysis a large number of patent information in-depth and to find that users interested and valuable patent knowledge, and then develop the patent search and analysis systems with own intellectual property rights for scientific research , business and other technological innovations required. Aimed to the actual needs of the system for patent information processing and analysis, the system is designed consist of three major parts: data preprocessing, data mining and patent information visualization. The system frame of the patent information analyzing and predicting system is designed as Figure 1.

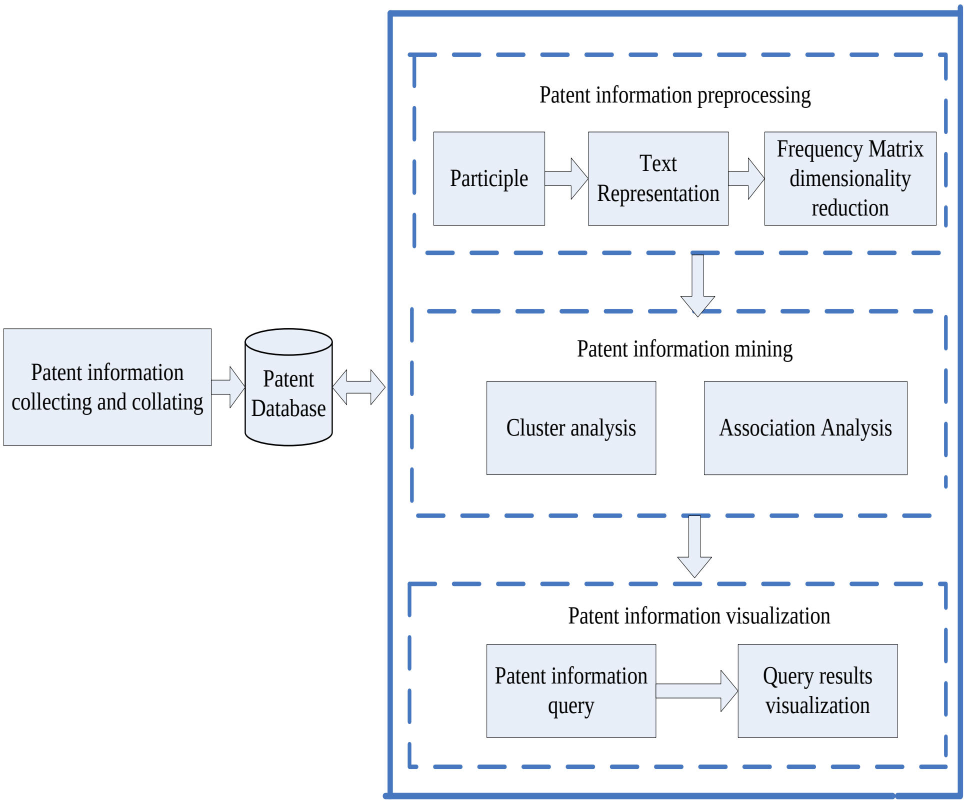
Patent information preprocessing
Participle
Text Representation
Frequency Matrix dimensionality reduction
Patent information collecting and collating
Patent Database
Patent information mining
Cluster analysis
Association Analysis
Patent information visualization
Patent information query
Query results visualization

Figure 1. Patent Information Processing Frame Based on Data Mining Technology

4. System Implementation

4.1. Patent Information Preprocessing

The type of patent information is generally textual data. It is different with traditional structured data, text data is semi-structured data or none-structured data. Even if there is some structure in patent data, it is still focused on the format, not the content of the text. So, the first step of patent information mining is the pretreatment and standardized of text data, and this means that we need convert text to a computer recognizable data format. The preprocessing of patent information consists of three steps.

4.1.1. Participle

Firstly, we need word segmentation for patent abstract samples or patent documents. At present, the most widely used participle method is sub-lexical dictionary which requires the use of a specialized dictionary, and the dictionary contains technical terms in the field of the sample to be analyzed. When indexing text information to match the dictionary, the reverse long way match reverse [16] can be used to establish inverted file.

4.1.2. Text Representation

The text can be seen as a keyword collection appears in the text and these keywords are called feature items. Data preprocessing must first remove empty words in the text, such as the "the”, “that" and so on. Then the concept of mapping and concept disambiguation are needed, because some words are different in the form but the concept is the same. In the field of information processing, the representation of text usually adopt the vector space model (referred to as VSM), in which the text are represented by the vector, *i.e.*, the document- to-item matrix [17].

4.1.3. Frequency Matrix Dimensionality Reduction

When the amounts of text are too large, it means that the text will have a significant number of features of vocabulary, making document feature vector is an ultra-high- dimensional sparse matrix. Thus, not only the amount of computation is increased, but also the information processing efficiency is reduced. At the same time, it increases the difficult to find the relationship between words. Therefore, to reduce the dimension of the text vector space model is necessary. Usually the advanced latent semantic indexing method can be used for this purpose, which adopts singular value decomposition technology in matrix theory to translate the high-dimensional matrix into low-dimensional matrix.

4.2. Patent Information Mining

Patent analysis research the number relationship between each other by using a variety of quantitative and qualitative analysis methods and technology. The proposed system process and analysis patent information based on data mining technology, which can extract the appropriate patent information from patent documents including the large number of messy, isolated, lengthy, technical and legal terms. By mining its contents in-depth and researching interconnectedness between patents information, the system can discover the status, development and distribution of patented technology, and a lot of implicated information, patterns and knowledge to predict the development trend of specific technologies or technical fields. It can help users track competitors and result in the guidance for production and business decisions of national, industry, and enterprises. Patent information mining method mainly involves cluster analysis and association analysis.

4.2.1. Cluster Analysis

Clustering is a type of data mining technique which divided a data set into several groups or cluster, and such that data objects within the same group have a high degree of similarity, and different groups of data objects is dissimilar, namely to achieve so-called "feather flock together" [18]. Cluster analysis of patent information is a process that automatically classifies patent documents based on the text feature similarity. Because the neural network has the advantages of high endurance and low noise data error, the Self-Organizing Feature Map (SOM) neural network clustering algorithm is wildly used in text clustering. SOM is an unsupervised self-organization self- learning neural network consisting of full connections neuronal array, which can determine the category of input samples based on the location of the winning neuron. When SOM is used for patent information clustering analysis, samples vector consist of 0 or 1 produced from patent information preprocessing are input into SOM to the training model. By following the training, samples containing similar words patent were assigned to the same category, which is the patent with similar subject was in the same category, and this can help users determine the core technology in the technical field based on the distance.

4.2.2. Association Analysis

Association analysis is one of the most active data mining technology, namely by association rule mining to find interesting links between items dataset. Mining association rules consists of two steps: ①to find all frequent item sets; ②generate strong association rules from frequent item sets. Most classical association analysis algorithm is Apriori [19] algorithm, the properties of which are: all non- empty subsets of frequent item set must also be frequent. In this paper, the association analysis is applied to the analysis of patent information; the aim is to find the hidden interesting things in patent information, which described closeness between groups of patents. Apriori algorithm is used for further analyzing invention person obtained from cluster analysis of in the core technology field and mining frequent item set by the company as a key. The relationship between enterprises in the same frequent items focused more closely in, and this can help users determine major competitors in the technical field.

4.3. Patent Information Visualization

Visualization view generation and management is an important component of patent information processing and analysis system. Topic Maps [20] is a new digital knowledge organization and knowledge navigation mode, which described the relationship between abstraction and links and between topics and specific resources. The use of topic maps manner for patent intelligence analysis and visualization, can directly reflect the knowledge structure of patent information, and can help find and identify correlate between patent information, in order to achieve effective organization and quickly and accurately navigate of patent information. In this paper, we proposed a Topic Maps visualization method suitable for text clustering analysis results, which take full advantage of the characteristics of the text clustering tree structure to construct topic maps and abstract layout resources document in thematic maps.

5. Conclusions

Aimed to the actual needs of enterprises for patent information retrieval and analysis, the paper proposed a frame of patent information analyzing and predicting system based on data mining. The designed system integrated data preprocessing, data mining and patent mapping technologies, which can provide patent information retrieval, analysis, early warning and other services for enterprises to enhance the core competitiveness of enterprises. The paper introduced the function structure and implementation method of each parts of the system.

References

[1] http://www.wipo.int
[2] http://www.derwent.com
[3] http://www.micropat.com
[4] http://www.european-office.org
[5] Y. Ming, “Research on Application of Data Mining in Analyzing and Forecasting the Patent Information”, Master's Degree Thesis of Wuhan University of Technology, **(2006)**.
[6] J. Han and M. Kamber, “Data Mining”, Concepts and Techniques, Elsevier, Holland, **(2011)**.
[7] M. Shih, D. Liu and M. Hsu, “Discovering competitive intelligence by mining changes in patent trends”, Expert Syst. Appl. vol. 37, no. **(2010)**, pp. 2882-2890.
[8] Y. Bing, Z. Donghua and R. Zhijun, “Analysis of Patent Intelligence Based on Data Mining Technology”, Journal of Information, vol. 12, **(2006)**.
[9] F. Ma and X. Wang, “Analysis of Patent Intelligence Based on Data Mining”, Qing Bao Ke Xue, vol. 26, no. 11, **(2008)**.
[10] A. J. Trippe, “Pat informatics: Identifying Haystacks from Space”, Searcher, vol. 10, no. 9, **(2002)**.
[11] M. Mogee and A. Breitzman, “The many applications of patent analysis”, informant Sci. vol. 28, no. 3, **(2002)**.
[12] A. Breitzman and P. Thomas, “Using patent citation analysis to target value M&A candidates”, Res Techno Mgrnt., vol. 45, no. 4, **(2002)**.
[13] www.delphion.com
[14] http://www.sipo.gov.cn/
[15] http://www.eastlinden.com/ch/index.aspx
[16] D. Zhenguo, Z. Zhuo and L. Jing, “Improvement on reverse directional maximum matching method based on hash structure for Chinese word segmentation”, Computer Engineering and Design, vol. 29, no. 12, **(2008)**.
[17] S. Guoju and Z. Jie, “An Evaluation of Feature Selection Methods for Text Categorization”, Journal of Harbin University Science and Technology, vol. 10, no. 1, **(2005)**.
[18] W. Jishang and W. Rousseeu, “Visual Cluster Analysis of Temporal Sequences”, PH.D Theses of University of California, Davis, **(2014)**.
[19] Z. Song and S. Liquan, “Improvement of Apriori Algorithm”, Journal of Harbin University Science and Technology, vol. 12, no. 5, **(2007)**.
[20] B. Wtodarczyk, “Topic Map as a Method for the Development of Subject Headings Vocabulary”, An Introduction to the Project of the National Library of Poland, Cataloging & Classification Quarterly, vol. 51, no. 7, **(2013)**.