Research on patent classification by extracting the content features of patent documents

Jie Gui, Wen Zeng, Hongqi Han, Zhaofeng Zhang

Institute of Scientific and Technical Information of China, 100038 Beijing, China
Email:guij@istic.ac.cn

Abstract: Since the public patent classification systems can't satisfy the patent analysis, a new research on patent classification has been developed in technological domain according to technological features of specific domains and content features of patent documents in this paper. The patent datum of new energy automotive domain was selected for empirical research.

Keywords: patent classification in technological domain; technology and effect; patent analysis

1. Introduction

With the economic development, Intellectual Property (IP) management has been taken more attention by the enterprises. The quantitative patent analysis at macro level, as an important tool for IP management, can't satisfy the needs of R&D management for the enterprises again. Patent analysis by mining text information has been developed increasingly in R&D management research field.

Now, many researches focus on the patent content analysis, such as technology-effect matrix, patent maps and tech-mining. To realizing above analysis methods, a key problem must be considered that how to identify and classify useful technical information by an effective way. However, International Patent Classification (IPC) system is based on the function and application classification and can't disclose the key technologies or effects. In the view of the enterprises, they need technological categories from multiple perspectives that are different from the classification at patent application status.

In this paper, we aim to develop a patent classification method by extracting technology and effect according to the content features of patent documents which would be expected to support text-mining and analysis of patent documents.

2. Research framework of patent classification by extracting the content features

2.1. Research framework

In this research, we would develop the method of patent classification from the content features of patent documents. A major problem is how to determine the classification standard which also affects the content extraction standard.

Considering of the patent full-text features, a patent document would be divided into four parts commonly, respectively for the technical topic, effect, application field and invention type, and these four parts can reflect the technological, economic and legal value of patent information [1][2][3]. So, in our research, we will develop the patent classification at technology-effect dimension, and the figure 1 shows the research framework.

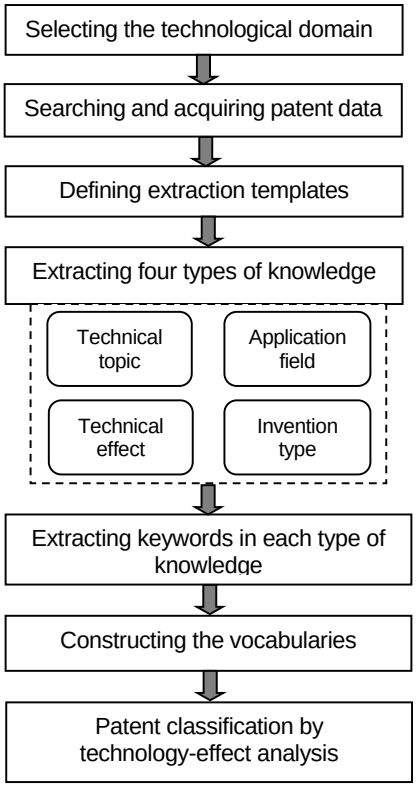

Figure 1. Research framework

2.2. Content feature analysis of patent documents

Patent documents contain rich information. Some information can be acquired from patent front pages easily, while the others are hidden in the full-text literatures. For instance, novelty of a patent may be distributed in the title, abstract, the claims, background, summary, drawings and detailed description, the emphasis of the above information may be also very different.

Considering of data acquiring, full-text patents include comprehensive information, but most of them are PDF and image files which are difficult to be converted into word or text format which couldn't be suit for data extracting and analysis by software systems. So, in the research, we select the patent title, abstract and the first claim as the analysis objects. These text items can be acquired with digital format and include the content of technical topic, application field, technical effect and invention type. We can extract the keywords according to technological content features.

3. The empirical research of patent classification in the new energy automotive domain

3.1. Data preparing

We select 100 Chinese patent documents of new energy automotive domain as raw datum of empirical research. The first step, we extract the key sentences from title, abstract and the first claim for each patent that include the content features of technical topic, application field, technical effect and invention type. The second step, the keywords would be extracted from the sentence collection and be classified into four content features again. Each sentence and keyword will be correlated to the original patent in the database. The result of extraction is shown in table I.

Table 1. The result of keyword extraction in four content features

Content feature	Num of keyword to be extracted
Technical topic	77
Application field	108
Technical effect	44
Invention type	12

3.2. Patent classification by technology and effect

By extracting content features and keywords, the technological emphasis in the patent documents can be revealed on distinctly. By this method, patent would be classified by key technologies, application filed and effect, and we can analyze patent value from multidimensional views to supporting R&D management [4][5].

Table 2. Technology-effect matrix analysis

Technical Effect / Technical Topic	Low cost	Multi work modes	Reasonable structure	Benefit of component integration	Fast start speed	High efficiency	Low noise
Transmission system	6	3	1			2	1
Power coupling device			2				
Hybrid vehicle	1				3	1	1
Control device				1	1		
Power drive			2			3	
Water cooling system	1			4			

We make the empirical research by the result of keyword extraction in four content features. We select two content features, technical topic and technical effect, as classification dimensions. Table 2 shows an example of technology-effect matrix analysis. In table 2, we choose the several typical technological topics and corresponding effect to construct a matrix that helps to identify key technologies at micro level. When we make the classification by technology-effect matrix, we can find out the technological knowledge quickly and accurately. Similarly, technical topic and application field can be as the two-dimensional indicators for patent classification.

4. Conclusion

The paper has explored the patent classification by technological dimension. The current patent classification system, such as IPC, USPC, ECLA and FI-FT can't meet the demand of practical technology classification at all. On the other, artificial technology classification depend on personnel knowledge and experience strongly. In this research, we construct classification by patent content features which help to mine technological knowledge from the view of technology and effect. The empirical research in the green energy vehicle domain identifies that our method can improve the patent classification and support further patent mining and analysis.

Acknowledgement:

The project of this paper is supported by the National Social Science Fund Project (Grant No.14BTQ038): Research on Information Analysis Method and Integrated Platform Based on Fact-type Scientific and Technical Big Data.

5. References

[1] Zhang Zhaofeng, Gui Jie, Zhang Yunliang, and Liu Xiwen. 2013. “Research of Patent Indexing and Application Based on Chinese Scientific and Technical Vocabulary System,” Digital Library Forum, (11):9-14.

[2] Li Peng, Gui Jie, Qiao Xiaodong and Zhang Zhaofeng. 2010. “Information Extraction of Patent Summary Based on Integration of CRFs and Rule,” Digital Library Forum, (9):2-6.

[3] Chen Ying, and Zhang Xiaolin. 2011. “Study on the differentiating method of technical and effect words in patent,” New Technology of Library and Information Service, (12): 24-30.

[4] Huo Cuiting, JiangvYongqing, Ling Feng and Liu Huijing. 2013. “Application study on patent technology/function matrix based on Japanese classification system (FI/F-term),” Journal of intelligence, (11):140-144.

[5] Chen Ying, and Zhang Xiaolin. 2011. “Research progress on construction of patent technology-effect matrix,” New Technology of Library and Information Service, (11): 1-8.