A Semi-Automatic Annotation Method of Effect Clue Words for Chinese Patents Based on Co-Training

Na Deng, Hubei University of Technology, Wuhan, China
Chunzhi Wang, Hubei University of Technology, Wuhan, China
Mingwu Zhang, Hubei University of Technology, Wuhan, China
Zhiwei Ye, Hubei University of Technology, Wuhan, China
Liang Xiao, Hubei University of Technology, Wuhan, China
Jingbai Tian, Hubei University of Technology, Wuhan, China
Desheng Li, Hubei University of Technology, Wuhan, China
Xu Chen, Zhongnan University of Economics and Law, Wuhan, China

ABSTRACT

In the era of big data, the latest and most advanced technologies are usually revealed to the world in the form of patents. Patents include abundant technical, economic and legal information. A deep analysis and mining of patents can provide important support for enterprises. Patent effect annotation is an important step in patent analysis and mining, and the extraction of patent effect clue words can greatly improve the accuracy and recall rate of annotation. This article summarizes the classification and characteristics of effect clue words, and proposes a co-training-based method of extracting effect clue words from Chinese patents suitable for various fields. Through a strategy called self-filtering, this method can gradually enrich effect clue words thesaurus by iterations, not relying on any other third-party filters. The experiments give the detailed steps, comparisons and boosting of the method.

KEYWORDS
Automatic, Data Mining, Information Extraction, Iteration, Machine Learning, Self-Filtering, Technology Effect Matrix

INTRODUCTION

In the era of big data, all walks of life carry out business through the network, resulting in the accumulation of large amounts of data in the network (Bouramoul, 2016; Mary & Malarvizhi, 2014; Pereira & Pereira, 2015; Qumsiyeh & Ng, 2016; Shen, Liu, Shen, Liu, & Sun, 2017; Shen, Shen, Chen, Huang, & Susilo, 2016; Tsai, 2011; Tsou, 2010). In healthcare field, a large number of patient information, drug information, and diagnosis and treatment information are stored. (Barbantan, Porumb, Lemnaru, & Potolea, 2016; Wang X, 2015). In education, there are a lot of information

DOI: 10.4018/IJDWM.2018100101

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
about students, teachers and specialties. In telecommunications industry, massive traffic data and communication data are generated every day (Trasarti, Giannotti, Nanni, Pedreschi, & Renso, 2011). The analysis and effective use of data in various areas can help each industry arrange resources reasonably, increase productivity and discover opportunities. Mining the hidden information in these data can help managers to make decisions to improve the quality and efficiency of production and life (Daly & Taniar, 2004; Silvestri, Corazza, Benerecetti, & Alicante, 2016; Taniar, Rahayu, Vincent, & Daly, 2008). However, regardless of which field, the latest and most advanced technologies and methods are usually revealed to the world in the form of patents, in order to grab technology heights as early as possible.

Patent is a kind of special text in the Internet, with strict format requirements and writing habits, which would bring in conflict and contradiction. On the one hand, it is necessary to express the techniques and inventions clearly; and on the other hand, the expressions should be as obscure as possible to prevent the invention from being imitated or infringed. As a carrier of human wisdom and innovation, patents contain rich technical, economic and legal information. In recent years, patent has become a competing object of analysis and mining. The effective use of patent information can provide important support for enterprises on technological innovations, avoiding risks, purchasing patents, safeguarding their interests and so on (Mandl, 2017; Tseng, Lin, & Lin, 2007; Zhang, Li, & Li, 2015).

Patent annotation (Agatonovic et al., 2008; Carvalho, Franca, & Lima, 2014) is a key step in patent online retrieval, analysis and mining. Patent annotation extracts important information from patents, such as techniques, functions, keywords and so on, which can help to realize online retrieval, patent analysis and mining on semantic level, reflecting a certain degree of intelligence. In patent online retrieval, you can improve the recall rate by extending the search terms with similar techniques or functions. In patent analysis, patent technology effect matrix can be constructed by enumerating the technologies and effects of multiple patents in tabular form to help patent applicants discover patent minefield and patent blank area (Chen, 2011; Zhang, 2017). In patent mining, patent annotation is the key step of patent classification, clustering and recommendation.

Patent abstract is an important component of patent text. It describes and summarizes patent background, purpose, method and function in brief space, and usually does not include any professional and complicated legal information, while retains most of important information in patent. Thus, patent abstract is a very good data source for patent annotation, analysis and mining.

Due to the complexity and particularity of patents, patent annotation is usually accomplished by manual work. The staff mark out function or effect statements from patent texts after reading. This method is relatively accurate, but time-consuming. Therefore, people begin to seek automatic or semi-automatic method of patent annotation. However, the text of a patent abstract is very short, usually with only 50-250 words. It is not easy to mark the effect statement from the short text. The authors have to think about the reason that the accuracy of artificial annotation is high. Artificial annotation is accurate because manual work can determine whether a statement is a valid effect statement in semantics. Through a lot of observation, the authors found that, in most cases, effect statements often contain some special words, which play a role of tips, instructions or implication. They help people understand sentences in semantics. Based on this discovery, this paper tries to find a way to collect these special words as many as possible. In this paper, they are called as effect clue words. The extraction of effect clue words can help improve the accuracy and recall rate of patent annotation and provide support for patent analysis and retrieval.

In this paper, the processed objects of annotation are Chinese patent abstract texts. In the effect annotation of patent abstracts, whatever manual or automatic, determining whether a statement contains effect clue words is necessary. In order to make the annotation more accurate, the collection of effect clue words should be as accurate and complete as possible. Using the idea of co-training, this paper proposes a semi-automatic method to annotate effect clue words. In this method, the recognition of
Using TreeNet to Cross-Sell Home Loans to Credit Card Holders
www.igi-global.com/chapter/using-treenet-cross-sell-home/40413?camid=4v1a

A Parallel Implementation Scheme of Relational Tables Based on Multidimensional Extendible Array
www.igi-global.com/article/parallel-implementation-scheme-relational-tables/1775?camid=4v1a

A Subspace-Based Analysis Method for Anomaly Detection in Large and High-Dimensional Network Connection Data Streams
www.igi-global.com/chapter/subspace-based-analysis-method-anomaly/73456?camid=4v1a