Expert Systems with Applications 42 (2015) 4348–4360

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Keyword selection and processing strategy for applying text mining to patent analysis

Heeyong Noh, Yeongran Jo, Sungjoo Lee *

Department of Industrial Engineering, Ajou University, San 5, Woncheon-dong, Yeongtong-gu, Suwon-si, Gyeonggi-do 443-749, South Korea

ARTICLE INFO

Article history:
Available online 2 February 2015

Keywords:
Patent analysis
Text-mining
Keyword selection
Keyword processing
Document clustering

ABSTRACT

Previous studies have applied various methodologies to analyze patent data for technology management, given the advances in data analysis techniques available. In particular, efforts have recently been made to use text-mining (i.e. extracting keywords from patent documents) for patent analysis purposes. The results of these studies may be affected by the keywords selected from the relevant documents – but, despite its importance, the existing literature has seldom explored strategies for selecting and processing keywords from patent documents.

The purpose of this research is to fill this research gap by focusing on keyword strategies for applying text-mining to patent data. Specifically, four factors are addressed; (1) which element of the patent documents to adopt for keyword selection, (2) what keyword selection methods to use, (3) how many keywords to select, and (4) how to transform the keyword selection results into an analyzable data format. An experiment based on an orthogonal array of the four factors was designed in order to identify the best strategy, in which the four factors were evaluated and compared through k-means clustering and entropy values. The research findings are expected to offer useful guidelines for how to select and process keywords for patent analysis, and so further increase the reliability and validity of research using text-mining for patent analysis.

1. Introduction

Patent documents include bibliographical information such as application date, filing date, assignees and inventors, as well as descriptions of the novelty of the invention and its application areas as covered by the corresponding patent (Yoon & Park, 2004). They have widely regarded as an important source for evaluating technological strength and weakness and/or corporate R&D efforts and performance (Li, Wang, & Hong, 2009), and the bibliographic information in patent documents have been widely used for technology analysis and management – e.g. identifying technology trends, predicting emerging technologies (Basberg, 1987), and assessing technological capabilities at individual, firm, sector and national levels (Ernst, 2003).

Technological information extracted from patent data – the descriptive element of patent documents – has also recently been utilized in various advanced data analysis techniques and in developing text-mining tools (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998; Murphy et al., 2014; Trippe, 2003): in particular, the automatic extraction of major keywords from patent documents has been applied in technology management contexts (Dou, Leveillé, Manullang, & Dou, 2005). Whereas some researchers are still skeptical about the effectiveness of patent analysis based on this keyword-based approach (Krier & Zacca, 2002), others have emphasized its value and potential. For example, Fattori, Pedrazzi, and Turra's study proved that patent classification using text-mining could be effective, and could also overcome the limitations of conventional patent classifications (Fattori, Pedrazzi, & Turra, 2003): other researchers have used text-mining to conduct patent analyses, and shown that the approach is valuable for creating new technology and identifying technology opportunities.

In keyword-based studies, researchers have commonly tried to achieve their study goals through analyses using sets of keywords extracted from patent documents (Yoon, Lee, & Lee, 2010). Analysis results will thus depend on the keyword set that is selected – if it does not represent the characteristics of the entire document well, the reliability and accuracy of the subsequent analysis may be affected, which in turn will make it difficult to draw reliable insights from the results. Thus selecting and processing keywords that represents the patent's key technological concepts accurately is critical but challenging in patent analysis as modeling bibliographic data is significant but challenging in bibliometric analysis

* Corresponding author. Tel.: +82 31 219 2419; fax: +82 31 219 1610.
E-mail addresses: nhy6692@ajou.ac.kr (H. Noh), mistylake2357@ajou.ac.kr (Y. Jo), sungjoo@ajou.ac.kr (S. Lee).

http://dx.doi.org/10.1016/j.eswa.2015.01.050
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

(Ferrara & Salini, 2012). The importance of keyword selection and processing has been recognized not only in the field of patent analysis research, but also in text-mining application (Cheong, Chiu, Shu, Stone, & McAdams, 2011; Clifton, Cooley, & Rennie, 2004; Li et al., 2009) – but despite its importance, few previous studies have dealt with the factors that affect effective keyword selection and processing for patent analysis. Most have assumed that the keywords used in their studies have been extracted well, and have not examined the keyword selection processes carefully: so a systematic investigation of keyword selection and processing strategy for patent analysis is badly needed.

This research, therefore, focuses on the keyword selection and processing strategy for applying text-mining to patent analysis, and proposes some relevant guidelines. The strategies commonly used in the existing literature are reviewed and a method developed to evaluate their performance. Based on this, the performance is evaluated and the best suggested. The research findings are expected to help in the effective strategic use of keywords for patent analysis and thus further increase the reliability and validity of future research applying text-mining to this end.

The overall structure of this paper is as follows. Section 2 describes the basic trends of text-mining based patent analysis, and Section 3 discusses four significant factors regarding keyword selection and processing strategies for patent analysis. Section 4 explains the overall research framework and the detailed research methods, and the research results are described in Section 5. Finally, Sections 6 and 7 present the implications and limitations of our research, together with some concluding remarks.

2. Text-mining based patent analysis

Text-mining and its applications have received a lot of attention as a method to acquire useful information from unstructured corpora. Text-mining applications can be utilized in various domains; i.e. not only to help novel thinking (Gentner & Markman, 1997; Segers & De Vries, 2003), but also to create artificial intelligence (Falkenhainer, Forbus, & Gentner, 1986; Salton & Waldstein, 1978). In addition, as more reliable tools are being developed for text analysis, it has become possible to capture useful text information for an analysis that was unavailable within conventional approaches (Fujii, Iwayama, & Kando, 2007; Mukherjea, Bamba, & Kankar, 2005; Trippe, 2003). Especially in recent days, text-mining approach is utilized actively in technology management fields. A specific application includes text-mining based patent analysis, where patents are analyzed to investigate technology characteristics. Patent documents are considered as a valuable database for understanding technology trends and establishing innovation strategy because of the four reasons. First, patent documents are fully opened to the public, being accumulated for each year and each technological field. They contain information about almost all relevant technological fields, and (although there are a few exceptions) the great majority of novel inventions are patented. Hence, if text-mining tools can extract technological contents effectively, patent databases can provide a valuable source for in-depth technology analysis. Thus we can use patent documents to investigate technological trends, assess technological capabilities, and analyze the commercial value of technologies (Choi & Hwang, 2014). Secondly, patent databases are easily accessible – the advancement of IT and patent database systems has made it easier to obtain patent documents by downloading them through the internet (Schwander, 2000). Thirdly, the descriptive parts of patent documents are written in natural language, but in the same formats with consistent headings. Patent data are semi-structured, rather than unstructured, and technological contents are relatively easy to extract using text-mining tools (Kang, Na, Kim, & Lee, 2007). Finally, patent database can be a way to resolve a chronic limitation of the text-mining approach. The limitation of keyword-based approaches is that keywords can have various meanings, so keyword-based analysis results may misrepresent facts. However, most of terms used in patent documents are technical in nature, making it more likely that keywords have only single meanings, so the problems associated with text-mining approaches are expected to be relatively less severe in patent data than other applications (Lee, Yoon, & Park, 2009). Cheong et al. (2011) argued that engineering (or technological) keywords are not always useful for representing documents' contents but this is not true for patent documents (Kang et al., 2007). That is, engineering (or technological) terms can be used as representative keywords of patent documents because of the unique characteristics of patents, as was mentioned above. In addition, a number of studies showed that a text-mining based patent analysis with WordNet or latent semantic analysis enables to construct word ontologies systematically by identifying synonyms or hypernyms–hyponyms of a set of keywords (Fu, Cagan, Kotovsky, & Wood, 2013; Fu, Chan, Cagan, Kotovsky, Schunn, & Wood, 2013; Mukherjea et al., 2005; Murphy et al., 2014; Verhaegen, D'hondt, Vandevenne, Dewulf, & Duflou, 2011). These four characteristics of the patent database make text-mining – whose main merits are comprehensiveness, standardization and general applicability – has become more widely utilized for patent analyses.

Previous studies applying text-mining to patents can be divided into three categories. First, there are patent-map related studies, which have suggested methods to map the technological characteristics of patent documents visually, so as to identify new technology opportunities (Kim, Suh, & Park, 2008; Kim et al., 2014; Li et al., 2009; Son, Suh, Jeon, & Park, 2012; Tseng, 2005) or even management opportunities such as M&A (Park, Yoon, & Kim, 2013). Other studies have addressed the relationships between patents by conducting network analyses based on keywords (Yoon & Park, 2004). Text-mining has been combined with other analysis methods – such as conjoint analysis and data envelopment analysis – to obtain more meaningful findings for analyzing technology trends and identifying new technologies (Daim, Rueda, Martin, & Gerdsri, 2006; Lee, Lee, & Yoon, 2011; Seol, Lee, & Kim, 2011; Yoon & Park, 2007). A second research stream has emphasized text-mining methods' ability to reduce the huge amounts of resources and efforts necessary for the technology classification of patent documents, not simply advancing patent classification techniques, but also proposing automatic classification systems for patent documents (Chakrabarti, Dom, Agrawal, & Raghavan, 1998; Fall, Törcsvári, Benzineb, & Karetka, 2003; Lamirel, Al Shehabi, Hoffmann, & François, 2003; Lee et al., 2009; Liang, Tan, & Ma, 2008; Trappey, Trappey, Hsu, & Hsiao, 2009; Tseng, Lin, & Lin, 2007), whose effectiveness has already been verified by many research institutes. Finally, there are a set of previous studies concerning how to extract meaningful keywords when a text-mining approach is applied to patent documents. These studies are again grouped into two types. Most of them have focused on meaningful keywords extraction as tools to solve a certain problem. For example, researchers tried to solve a TRIZ problem by constructing meaningful keyword ontology (Liang et al., 2008; Souili & Cavallucci, 2013; Souili, Cavallucci, Rousselot, & Zanni, 2011). On the other hand, a few others, though not many, have concentrated on the type of text-mining approaches that are appropriate for patent analysis. For example, researchers tried to compare several keyword selection criteria including keyword frequencies in documents, variances of keyword frequencies across documents, and TF–IDF values (Lee et al., 2009; Li et al., 2009), while others have sought to identify the most appropriate parts of patent documents from which to extract keywords, such as titles, abstracts, claims and descriptions (Xie & Miyazaki, 2013).