Engineering Applications of Artificial Intelligence 24 (2011) 1510–1520

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A Semantic-based Intellectual Property Management System (SIPMS) for supporting patent analysis

W.M. Wang*, C.F. Cheung

Knowledge Management Research Centre, Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 20 June 2010
Received in revised form
6 April 2011
Accepted 14 May 2011

Keywords:
Knowledge management
Knowledge-based system
Semantic analysis
Technology management
Concept extraction
Patent analysis

ABSTRACT

Patent databases provide valuable information for technology management. However, the rapid growth of patent documents, the lengthy text and the rich of content in technical terminology, and the complicated relationships among the patents, make it taking a lot of human effort for conducting analyses. As a result, an automated system for assisting the inventors in patent analysis as well as providing support in technological innovation is in great demand. In this paper, a Semantic-based Intellectual Property Management System (SIPMS) has been developed for supporting the management of intellectual properties (IP). It incorporates semantic analysis and text mining techniques for processing and analyzing the patent documents. The method differentiates itself from the traditional technological management tools in its knowledge base. Instead of eliciting knowledge from domain experts, the proposed method adopts global patent databases as sources of knowledge. The system enables users to search for existing patent documents or relevant IP documents which are related to a potential new invention and to support invention by providing the relationships and patterns among a group of IP documents. The method has been evaluated by benchmarking with the performance against traditional text mining technique and has successfully been implemented at a selected reference site.

1. Introduction

In knowledge-based economy, the management of intellectual properties (IP) has become more and more important for an industry. It also plays an important role in technology management. For instance, a patent is a contract between an inventor and the government, whereby in return for full public disclosure of an invention, the government grants the inventor the right to exclude others for a limited time from making, using and selling the invention (Hufker and Alpert, 1994). Many companies hold their patents not only as an invisible asset, but also as a strategy for its development and competition against its competitors in the market (Jung, 2003). However, inventing a new patent is not an easy task. Some "novel" ideas might impinge on some claimed rights protected by other's patents. Therefore, before inventing a new patent, it should ensure that it does not infringe other's patents by consulting with some global patent databases. In fact, the World Intellectual Property Organization revealed that from 90% to 95% of world's inventions are found in patented documents (Brockhoff et al., 1999). The European Patent Office also disclosed that more than 80% of man's technical knowledge is described in patent literature (European Patent Office, 2010). By carefully analyzing the patent documents, it cannot only increase the efficiency and effectiveness of making new invention substantially but also reduce the risk of infringing the patent rights of others (Soo et al., 2005).

Patent documents analysis has often been employed to generate economic indicators that gage the linkage between technology development and economic growth (Campbell, 1983; Grandstrand, 1999; Grilliches, 1990; Holl et al., 2000), estimate technological knowledge flows and their impact on productivity (Evenson and Puttnam, 1988; Scherer, 1982), compare innovative performance in international context (Paci et al., 1997), evaluate the competitiveness of firms (Narin and Noma, 1987), develop technology plans (Mogee, 1991), prioritize R&D investment (Hirschey and Richardson, 2001), or monitor technological change in firms (Archibugi and Pianta, 1996; Basberg, 1987). However, patent analysis requires considerable effort and expertise. It requires the analysts to have a certain degree of expertise in information retrieval, domain-specific technologies, and business intelligence. In addition, patent documents are often lengthy and rich in technical and legal terminology. It consumes a lot of time to read and analyze them even for experts. As a result, it is necessary to have an automated method for supporting the analysts in processing and analysis of massive patent documents as well as supporting technology innovation.

* Corresponding author.
E-mail address: mfming@inet.polyu.edu.hk (W.M. Wang).

0952-1976/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2011.05.009

The paper is outlined as follows. Section 2 reviews the related work of the study which includes patent analysis, text mining, and keyword extraction. Section 3 describes the architecture and the components of the proposed system. Experiments and results for evaluating the system are provided in Section 4. The usefulness and effectiveness of the system in a real-life environment have been studied in Section 5 through a practical implementation of a prototype at a selected reference site. Section 6 is an overall conclusion of the paper, and some suggested areas for further work are also explored in this section.

2. Related work

Most studies in the patent analysis are on document search and classification (Fall et al., 2003; Larkey, 1999; Lai and Wu, 2005; Tsakalidis et al., 2002). A patent document contains a lot of items for analyses. Some of them are structured data. They are uniform in semantics and in format across patents such as patent number, filing date, or assignees. Some are unstructured data. They are free texts of various lengths and contents, such as claims, abstracts, or descriptions of the invention. Patent analyses based on structured information have been most practised approaches and have been found in the literature for years (Archibugi and Pianta, 1996; Be'de'carrax and Huot, 1994; Ernst, 1997; Lai and Wu, 2005).

These structured data can be analyzed by bibliometric methods, data mining techniques, or well-established database management tools such as On-Line Analytical Processing (OLAP). One type of structured data analyses is determined by citations (Karki, 1997). For example, if a patent is cited by a large number of other patents, this cited patent is possibly a foundation of those citing patents and is thus important. In another example, since two patents related to the same invention tend to cite the same patent and are cited by the same patent, interpatent similarity can be determined by cocitation analysis. Using patent citation analysis, Narin et al. (1984) evaluated corporate technological performance while Chakrabarti et al. (1993) analyzed the diffusion of technological information among different organizations. Lai and Wu (2005) used citation-based interpatent similarity to perform a patent classification. Recently, there has been an interest in applying text mining techniques to assist the task of patent analysis and patent mapping (Lent et al., 1997; Fattori et al., 2003; Yoon and Park, 2004).

Text mining can be viewed as data mining extended to text data to find implicit, previously unknown, and potentially useful patterns from a large text repository (Fayyad et al., 1996). It is an interdisciplinary area involving machine learning and data mining, statistics, information retrieval and natural language processing (NLP) (Grobelnik et al., 2002). Text mining can work with unstructured datasets such as full-text documents, HTML files, emails, etc. NLP techniques are commonly used as the first step in text mining for converting the unstructured text into a structured format (Milic-Frayling, 2005). The document can be featured by keywords that are extracted through text mining algorithms, and then other data mining techniques can be applied to retrieve interesting patterns. In relation to patent analysis, text mining is used as a data-processing and information-extraction tool. Since the original patent documents are expressed in text (natural language) format, it is necessary to transform raw data into structured data. Then, the process of keyword extraction is employed to identify keywords and to measure similarity between patents.

In the present study, a Semantic-based Intellectual Property Management System (SIPMS) is presented which attempts to address the shortcoming of the prior research in several respects. It applies text mining to analyze the unstructured part of the patent documents. It extracts the key concepts of the patent documents and discovers the relations among the concepts based on the syntactic structure of the documents. This approach is able to extract concepts and relationships from the document itself rather than retrieved from predefined ontologies. Hence, it enables the approach to be applied in any domains without the need for the capture of prior knowledge. It also ensures that the intentions of the author of the document can be preserved using the words generated from the document itself. Furthermore, the proposed algorithm produces concepts and relationships based on multi-word text structures instead of individual words which makes the concept and relationship labels to be more complete. Based on the proposed approach, the results of classification of the patents can achieve a higher accuracy.

Prior research into keyword extraction is summarized in three categories, which are dictionary (e.g. Braam and Moed, 1991; Callon et al., 1991; Zitt and Bassecoulard, 1994), statistical (e.g. Cutting et al., 1992; Eisen et al., 1998; Karypis et al., 1994), and linguistic approaches (e.g. Rajaraman and Tan, 2002; SanJuan and Ibekwe-SanJuan, 2006).

The dictionary approach utilizes a dictionary, which contains the forms, meanings and relationships between words and phrases. By matching the dictionary with the words of sentences in each article, the concepts in the article are extracted. Relationships between the concepts are then coordinated based on the dictionary. The major advantages of this approach are its efficiency of execution and ease of implementation. However, the prior keyword list is external to the documents. It reflects a general understanding of the domain instead of the intention of the author of the document. Since words in the target articles cannot consist of new words, it requires further operations for handling these new words. In the statistical approach, words are selected based on term weighting indices such as Inverse Document Frequency (IDF) or Mutual Information (MI). It also eliminates the low frequency words so as to reduce the number of words being extracted. However, this also results in the drastic elimination of more than half of the initial data from the analysis. Moreover, Price and Thelwall (2005) have demonstrated the usefulness of low frequency words for scientific web intelligence (SWI). Removal of low frequency words results in documents becoming more general and similar. The linguistic approach makes use of semantic knowledge bases, heuristics, or rules to extract concepts. However, from a conventional linguistic approach, it is hard to make linguistic generalizations that can be applied reliably due to the occurrence of ambiguous words and ambiguous sentences structures.

In addition, patent analysis is a complex task which consists of different objectives or goals. A multi-agent system (MAS) architecture has been developed which consists of a collection of autonomous agents which have defined their own goals and actions and can interact and collaborate among each other through XML communication. These agents act collectively and collaborate to achieve their own individual goals as well as the common goal.

3. The architecture of Semantic-based Intellectual Property Management System (SIPMS)

The architecture of SIPMS is shown in Fig. 1. It is a knowledge-based system which consists of three major processes which are pre-processing, patent analysis, and invention support. The combination of the algorithms is presented here for the first time for supporting technological innovation.

3.1. Pre-processing

There are different kinds of intelligent agents in the processes. The pre-processing process consists of extraction agent, segmentation