Technological Forecasting & Social Change 114 (2017) 281–292

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data

Junegak Joung, Kwangsoo Kim *

Department of Industrial and Management Engineering, Pohang University of Science and Technology, 77 Cheongam-ro, Nam-gu, Pohang 790-784, Republic of Korea

ARTICLE INFO

Article history:
Received 3 February 2016
Received in revised form 7 July 2016
Accepted 21 August 2016
Available online 9 September 2016

Keywords:
Patent analysis
Keyword-based model
Technical keyword
Technology planning

ABSTRACT

This paper proposes technical keyword-based analysis of patents to monitor emerging technologies, and uses a keyword-based model in contents-based patent analysis. This study also presents methods to automatically select keywords and to identify the relatedness among them. After using text-mining tools and techniques to identify technical keywords, a technical keyword-context matrix is constructed. The relatedness between pairs of keywords is then identified in a transformation of this matrix. Patent documents are clustered by using a hierarchical clustering algorithm based on patent document vectors. As a result, emerging technologies can be monitored by identifying clusters composed of technical keywords. A case study of mechanisms of electron transfer in electrochemical glucose biosensors is given to demonstrate how the proposed method can monitor emerging technologies.

1. Introduction

Monitoring of emerging technologies can identify incipient technological changes quickly, and is an invaluable component of technology planning, and of development of research and development (R&D) policy by governments and companies (Ashton et al., 1991). By investing in R&D strategically in potentially-important emerging technologies, companies can become market winners or early followers of market leaders (Hamilton, 1985). Given that technology monitoring can help companies' development of new products, technologies or joint ventures, monitoring of emerging technologies provides a starting point to induce radical technological change or mergers and acquisitions (M&A) (Ashton et al., 1994). Furthermore, many methods in technology forecasting predict emerging technologies, but have limited ability to identify possible emerging technologies. Therefore, use of reliable sources (e.g., research organization reports) to comprehend emerging technologies allows examination of how they are specifically realized. To constantly monitor emergence of technologies, patents are the best source of technology information, because they contain technical details. Patent analysis has been considered as a basis for technology assessment to monitor sources of technological knowledge (Ernst, 2003).

Patent documents describe commercialized inventions, and the number of granted patents represents a company's rate of technological advance (Ernst, 2001). For these reasons, analysis of patent data has two major benefits. First, patents are written to protect the right of invention and to prevent overlapping R&D investment, so they are reliable formal sources of unique technical information. Second, patent database systems are well-organized and are supplied with retrieval systems, so large numbers of patents can be examined easily. These systems present up-to-date information, and can rapidly check new applied or granted patents and their applications; moreover, patent information is available to everyone. For example, since 1976, the United States Patent and Trademark Office (USPTO) has provided full-text patents and a retrieval system that can be used free of charge by anyone. For these reasons, stakeholders such as R&D policy makers, R&D managers, technology developers, and R&D planners have used patent information to support identification of world-wide technical evolution (Zhang, 2011; Altuntas et al., 2015b), and to support making in R&D (Thorleuchter et al., 2010; Altuntas et al., 2015a). As a result, users can examine technological trends in which competitors' R&D strategies, technological resources, and external knowledge of technology are embedded.

Many techniques have been used in patent information analysis to monitor technological trends. Engelsman and van Raan (1992) suggested co-word maps to identify declining or emerging fields of technological activities, which were identified by examining keywords at the meso-level and micro-level. Yoon et al. (2002) proposed patent maps that displayed patents in two or three-dimensional space according to similarity of keywords; an absence of patents in the map was considered as a starting point to identify emerging technologies. Yoon and Park (2004) developed keyword-based patent network and identified up-to-date trends of high technologies by using network analysis. In addition, the k-Means algorithm (Kim et al., 2008), a formal concept analysis-based approach (Lee et al., 2011), and a novelty detection technique (Geum et al., 2013) have been used in keyword-patent matrices

* Corresponding author.
E-mail address: kskim@postech.ac.kr (K. Kim).

http://dx.doi.org/10.1016/j.techfore.2016.08.020

to capture technological flows and emerging patterns. More-advanced techniques have been described, such as an approach based on subject-action-object (SAO) relationships. Choi et al. (2011) suggested using an SAO-based network to identify technology trends; the authors constructing a noun-by-verb relationship matrix, so emerging technologies could be inferred by low density and high cohesion in the network. Wang et al. (2015) conducted SAO-based technology roadmapping to comprehend developing trends.

However, previous research to monitor technological trends has some limitations in the process of keyword selection and in identification of relatedness of keywords that all keyword-based patent analysis methods share. These processes rely too much on the intervention of experts, so the reliability of the analysis can be greatly affected by the experts' opinions. Expert-dependent analysis is also expensive because it is time-consuming and laborious. Keyword selection is the most crucial factor, but the main keywords are chosen based on the subjective judgment of experts (Tseng et al., 2007; Noh et al., 2015). Although some research (Yoon and Park, 2004; Lee et al., 2009; Geum et al., 2013) suggested term frequency to guide experts in evaluating significant terms, experts must still take time to eliminate common words. Moreover, previous extraction of keywords could not effectively analyze multiple-phrase words even if they articulated a patent document well. Identification of the relatedness of keywords as synonyms, hypernyms, and hyponyms raises the quality of semantic processing when comparing patent documents, but assessment of their relatedness has been completely dependent on technical experts. Choi et al. (2011) and Wang et al. (2015) proposed using a word ontology such as WordNet (Miller, 1995) to understand the relatedness of keywords, but WordNet uses only a hierarchy of generic terms, so it does not consider technical terms; therefore, it does not produce a good solution because patents include many technical terms.

To summarize, the selection of keywords and identification of their relatedness are important to provide reliable and objective results of patent analysis, but current algorithm are not sufficient to assist experts to select significant keywords or to grasp the relatedness among them (Joung and Kim, 2015). To reduce these limitations, measures that reduce dependence on experts are required.

Therefore, this paper suggests a reliable way to monitor emerging technologies. This method uses analysis of technical keywords extracted from patents, and improves on the keyword-based patent analysis. Technical keyword based analysis allows monitoring of whether patents embody emerging technologies. For this purpose, methods to choose technical keywords and to be aware of their relatedness are presented. Technical keywords are extracted using a commercial NLP software package (*Alchemy*™), then selected using a term frequency-inverse document frequency (TF-IDF) function. The relatedness between pairs of technical keywords is established by distributional similarity. Next, a dissimilarity matrix is built as the complement of the similarity of patent documents that was identified on the basis of the relatedness of keywords. Then a hierarchical clustering algorithm clusters the patents by considering the relationships encoded by this dissimilarity matrix. Emerging technologies can be identified by monitoring clusters composed of technical keywords. To demonstrate the validity of the proposed approach, a case study of mechanisms of electron transfer in electrochemical glucose biosensors is given.

The rest of this paper is organized as follows. Section 2 briefly provides theoretical background of proposed methodology. Section 3 explains the proposed methodology. Section 4 describes a case study of mechanisms of electron transfer in electrochemical glucose biosensors to apply the proposed approach. Section 5 provides conclusions and future work.

2. Theoretical background

2.1. Keyword-based patent analysis to identify emerging technologies

Keyword-based patent analysis has been applied in various techniques such as text mining, data reduction, clustering, and network analysis to identify emerging technologies. Engelsman and van Raan (1992) provided a co-word map based on keywords extracted by experts, and used it to visualize developments in fields of technology by using co-occurrences. Similarly, Courtial et al. (1993) identified emerging technologies by using degree of density and centrality in co-word analysis. Zhu and Porter (2002) offered automatic extraction of keywords by using text mining, and developed a commercial patent-analysis program, *VantagePoint*™. Given that text mining could extract keywords automatically, Yoon et al. (2002) proposed a self-organizing feature-map-based patent map that uses a keyword-patent matrix to produce simplified images of multi-dimensional patent data, and to visualize the dynamic pattern of technological advancement. Kim et al. (2008) developed a patent map to visualize emerging technologies and to anticipate future technological trends on the basis of a semantic network of keywords by clustering patent documents that share keywords. Lee et al. (2011) suggested a concept-analysis-based approach to monitor technological changes that organizes objects with shared properties based on a keyword-patent matrix. Geum et al. (2013) also applied novelty detection techniques that identify new or unusual data that existing systems did not perceive, and detected new and emerging pattern of patents. These existing papers recommended term frequency to assist experts to select keywords, but experts were still required to remove common words; this is also time-consuming work (Yoon et al., 2002; Lee et al., 2011; Geum et al., 2013).

In addition to these studies, Yoon and Park (2004) proposed a text mining-based patent network to analyze up-to-date trends of high technologies by calculating the distance between patents on the basis of a keyword-patent matrix. They then discovered emerging technologies by developing new network analysis indexes (centrality index, cycle index) and keyword clusters. Similarly, Chang et al. (2010) concentrated on cluster network analysis to identify key technologies in carbon nanotube field-emission displays. Zhang et al. (2014) provided term clumping that includes automatic keyword selection by exploiting term frequency-inverse document frequency (TF-IDF), but did not consider many variants of TF and IDF.

To improve the efficiency of keyword-based patent analysis, Yoon et al. (2011) suggested property-function that is composed of 'adjective + keyword' and 'verb + keyword' that are extracted using the Stanford dependency parser, and built a network based on co-occurrence matrix. Emerging properties and functions were then identified by analyzing small and highly dense sub-networks. Similarly, Yoon and Kim (2011) constructed a network on the basis of similarity matrix that compared patents by using SAO structures extracted by *Knowledgist*™, and in this way detected clusters that include up-to-date technology. Choi et al. (2011) built a network composed of noun nodes and verb nodes based on a noun-by-verb matrix, and identified emerging technologies by using density and a cohesion index to analyze sub-networks. Wang et al. (2015) proposed technology roadmapping based on SAO analysis to identify technology development trends and future directions of the technology domain. These studies suggested a thesaurus called WordNet to perform semantic processing while comparing patent documents (Yoon and Kim, 2011; Choi et al., 2011; Wang et al., 2015), but WordNet uses a glossary of generic terms and therefore cannot take technical terms into account.

Although previous research has been useful to look into emerging technologies, it has limitations, including too much dependence on expert intervention during keyword selection and identification of the relatedness of keywords. Consequently, improvement of keyword-based patent analysis requires algorithms to better secure experts' objectivity and reliability. For these reasons, algorithms to guide selection of technical keywords and to identify relatedness among them will be considered next.

2.2. Technical keywords & their relatedness

Development of text mining has resulted in ways to extract keywords automatically, so contents analysis has become possible; this