Technological Forecasting & Social Change 119 (2017) 39–52

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014

Hongshu Chen [a,*], Guangquan Zhang [a], Donghua Zhu [b], Jie Lu [a]

[a] *Decision Systems & e-Service Intelligence Lab, Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, PO Box 123, Broadway, NSW 2007 Sydney, Australia*

[b] *School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China*

ARTICLE INFO

Article history:
Received 19 December 2015
Received in revised form 3 March 2017
Accepted 6 March 2017
Available online 19 March 2017

Keywords:
Technological forecasting
Text mining
Topic modelling
Topic analysis

ABSTRACT

The study of technological forecasting is an important part of patent analysis. Although fitting models can provide a rough tendency of a technical area, the trend of the detailed content within the area remains hidden. It is also difficult to reveal the trend of specific topics using keyword-based text mining techniques, since it is very hard to track the temporal patterns of a single keyword that generally represents a technological concept. To overcome these limitations, this research proposes a topic-based technological forecasting approach, to uncover the trends of specific topics underlying massive patent claims using topic modelling. A topic annual weight matrix and a sequence of topic-based trend coefficients are generated to quantitatively estimate the developing trends of the discovered topics, and evaluate to what degree various topics have contributed to the patenting activities of the whole area. To demonstrate the effectiveness of the approach, we present a case study using 13,910 utility patents that were published during the years 2000 to 2014, owned by Australian assignees, in the United States Patent and Trademark Office (USPTO). The results indicate that the proposed approach is effective for estimating the temporal patterns and forecast the future trends of the latent topics underlying massive claims. The topic-based knowledge and the corresponding trend analysis provided by the approach can be used to facilitate further technological decisions or opportunity discovery.

1. Introduction

Patents are one of the most valuable indicators of technological trend detection and forecasting. They hold explicit technical information and implicit knowledge that indicate technological concepts, topics and related R&D activities, which can be used to support decision making, early warning signals for subsequent market shifts, or to promote future competition (Campbell, 1983; Ernst, 1997; Griliches, 1990; WIPO, 2004). Over the last decade, the continuous growth of patents has given rise to technological knowledge than ever before. However, it has also created information overload, whereby researchers face difficulties in understanding and analyzing massive data and their trends (Cunningham et al., 2006). Manually conducting content analysis on patent documents can be very time consuming and laborious (Tseng et al., 2007). Machine learning-based text analysis has been applied to change the status of traditional patent data analysis approaches and methods (Suominen et al., 2016).

Much effort has been devoted to the study of empirical technological forecasting based on patenting activities. From a temporal perspective, growth curves (S-curves) (Chen et al., 2011; Young, 1993), time series analysis (Porter and Cunningham, 2004), chaos-like behavior analysis (Modis and Debecker, 1992), non-linear regression fitting (Baskurt, 2011), smoothed trajectory (Krampen et al., 2011), Hidden Markov models (HMM) (Lee et al., 2011) and other promising approaches have been utilized to deal with trend forecasting tasks of a particular industry. Nevertheless, when it comes to estimating the underlying trend of detailed topics in large volumes of patent documents, text mining techniques are required to uncover the latent trends from a semantic perspective. As Zhu and Porter (2002) concluded, a managerially usable empirical technological forecasting first needs to have the capability to efficiently exploit massive textual data. Existing research has also made large strides in using text mining to support trend analysis. Kim et al. (2012) proposed a technology trend analysis and forecasting model based on ontology for systematic information analysis; Choi and Hwang (2014) incorporated both network-based and the keyword-based patent analysis methods for effective trend

* Corresponding author at: Faculty of Engineering and Information Technology, University of Technology, Sydney (UTS), 15 Broadway, Ultimo, NSW 2007, Australia.

E-mail addresses: hongshu.chen@uts.edu.au (H. Chen), guangquan.zhang@uts.edu.au (G. Zhang), zhudh111@bit.edu.cn (D. Zhu), jie.lu@uts.edu.au (J. Lu).

http://dx.doi.org/10.1016/j.techfore.2017.03.009
0040-1625/© 2017 Elsevier Inc. All rights reserved.

analysis; Chang et al. (2010) monitored the technological trends in an emerging field of technology by constructing maps using keywords and key phrases. In addition, morphology analysis, property-function research, semantic analysis approach, rule based methods and other text mining approaches have also been utilized as efficient tools to assist with more effective technological trend analysis (Abbas et al., 2014; Lee et al., 2013; Shih et al., 2010; Yoon and Park, 2005; Yoon and Kim, 2012).

However, from a temporal perspective, using the most accepted method, restrained fitting models, on patent counts, only provides a rough tendency estimation of the corresponding industry or technical area. In real-life situations, one patent document may contain a number of different technological topics. From a semantic perspective, as has been pointed out by many other researchers, the subjectiveness embedded in the patent classification process has been a limitation of using and analyzing patent data (Venugopalan and Rai, 2015). It brings drawbacks of clustering and presenting technological concepts using pre-defined categories but not actual topics discussed in patent documents. Moreover, the outcome of keyword-frequency-based text mining techniques are keywords with rankings; yet these words alone are usually too general or ambiguous to indicate a concept, especially when there are polysemous words actually describing different topics (Tseng et al., 2007). It is very difficult to track the temporal patterns of keywords for trend forecasting purpose as well.

To overcome these limitations, this research proposes a topic-based technological forecasting approach to discover and estimate the trends for specific topics underlying large volumes of patent claims using Latent Dirichlet Allocation (LDA). We bring the thematic analysis of patents and trend forecasting together to (1) identify temporal trend patterns and semantic topics quantitatively; and (2) integrate the two features in different dimensions to provide valuable topic-based knowledge and corresponding trend forecasting to facilitate further decision making and opportunity discovery. The trend patterns are first quantitatively learned using a piecewise approach and presented by a trend turning points matrix. Then for each discovered topic, a topic annual weight matrix and a sequence of topic-based trend coefficients are generated to estimate its developing trend. We then continue to evaluate to what degree various topics have contributed to the patenting activities of the whole area. Finally, a case study, using 13,910 Australian utility patents published during the years 2000 to 2014 in the United States Patent and Trademark Office (USPTO), is presented to demonstrate the effectiveness of the proposed approach. A number of strong topics with upward developing trends are identified and analyzed. The case study result shows that our proposed approach can be used to automatically uncover the thematic structure of massive patent data in a technological area of interest, and then estimate the detailed developing trend of each detected topic, thereby assisting decision making for potential opportunity identification, decision support and technical strategy formation.

This paper is organized as follows: Related Work reviews research related to our topic-based patent technological forecasting, by discussing empirical technological forecasting, Latent Dirichlet Allocation in patent analysis and piecewise linear representation. The Methodology section describes the full process of the proposed technological forecasting approach. The Case Study and Discussion present experiments using USPTO patents to conduct an examination of the approach and then explains how to use it in a real patent analysis context. Finally, the Conclusion and Future Work section summarizes this study and outlines future research directions.

2. Related work

2.1. Empirical technological forecasting

Empirical technology trend forecasting aims to build a bridge between trend patterns and the observations derived from technology indicators such as patents, scientific literature and R&D expenditure (Porter and Cunningham, 2004). An abstract representation of real-world dynamics in such circumstances is necessary to learn trend trajectories, shift and patterns, so that future trends can be estimated. Combining bibliometric analysis and curve fitting-based approaches are the most accepted and adopted empirical technology forecasting methods (Carrillo and González, 2002; Baskurt, 2011; Bengisu and Nekhili, 2006; Chen et al., 2011), in which the counts of patents, publications, or citations are used to measure and interpret technological advances (Watts and Porter, 2003). These model-based methods depict the characteristics of technology throughout their life cycles thus allow researchers to make strategic decision (Martino, 1993). They provide simple computation and straightforward presentation which are quite workable for general trend identification; however in real-world tasks, it is not common that the true saturation value of one technology or a group of technologies is known beforehand. In addition, when an innovation manifests in sudden shifts in a trend line (Phillips and Linstone, 2016), these detailed patterns need to be captured by more data-based approximations. In order to learn the patterns more efficiently, machine learning-based approaches start to be increasingly evolved into trend forecasting tasks. Suominen et al. (2016) applied a grouped time series model proposed by Hyndman and Athanasopoulos (2014) to forecast the future developments of target topics, creating a forward looking aspect central to technology management. Hidden Markov Model (HMM) approach was also used to model stair-like patterns of innovation and then cluster technologies with similar patterns (Lee et al., 2011, 2012). It brought machine learning to the technology trend analysis area, however the modeled patterns of technologies were only applied to assist subsequent clustering, not forecasting, thus further trend prediction is still needed.

2.2. Latent dirichlet allocation in patent analysis

Facing the limitation brought by the subjectiveness embedded in the classification process of patents, topic modelling-based approaches, represented by LDA, have become increasingly attractive to researchers due to their promising ability to automatically discover and present latent topics. LDA by Blei et al. (2003) is a probabilistic topic model that uses unsupervised learning to estimate the properties of multinomial observations. It provides an estimation of the latent semantic topics in massive documents and the probabilities of how various documents belong to different topics (Blei, 2012).

In the generative process of LDA, the overall documents are denoted as D, the topic numbers for D is K, the term number of the d^{th} document in the collection D is N_d and the n^{th} word in document d is $W_{d,n}$. The topic proportions for the d^{th} document is defined as $\vec{\vartheta}_d$. For document d, the topic assignments are Z_d, where $Z_{d,n}$ indicates the topic assignment of the n^{th} word in the d^{th} document. The topics themselves are illustrated by $\vec{\varphi}_{1:K}$, where each $\vec{\varphi}_k$ is a distribution over vocabularies. In addition, there are two hyper-parameters that determine the amount of smoothing applied to the topic distributions for each document and the word distributions for each topic, α and β. In summary, the generative process of LDA can be denoted by the joint distribution of the random variables as follows (Blei et al., 2003; Heinrich, 2005; Steyvers and Griffiths, 2007),

$$p\left(\vec{w}_d, \vec{z}_d, \vec{\vartheta}_d, \phi | \vec{\alpha}, \vec{\beta}\right) \\ = \prod_{n=1}^{N_d} p\left(w_{d,n} | \vec{\varphi}_{z_{d,n}}\right) p\left(z_{d,n} | \vec{\vartheta}_d\right) p\left(\vec{\vartheta}_d | \vec{\alpha}\right) p\left(\phi | \vec{\beta}\right).$$

The required parameters of LDA need to be estimated using an iterative approach. Among existing approaches, Gibbs sampling, which is one of the most commonly used methods, is an approximate inference algorithm based on the Markov Chain Monte Carlo (MCMC) method and has been widely used to estimate the assignment of words to topics