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/

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

by observed data (Griffiths and Steyvers, 2004; Noel and Peterson, 2014).

In practice, LDA has been utilized as a very efficient tool to assist topic discovery (Griffiths and Steyvers, 2004) and question answering (Yang et al., 2013), and has also provided aid in analyzing citation networks, data structures, time gaps, content comparison and scientific maps of publications in various areas (Ding, 2011; Jeong and Song, 2014; Chen et al., 2015a; De Battisti et al., 2015; Suominen and Toivanen, 2015; Venugopalan and Rai, 2015). Blei and Lafferty (2006) extended LDA and proposed the dynamic topic model to capture the evolution of topics in a probabilistic perspective. It brings time series and topic modelling together by dividing the data by time slice and modelling the slices with a k-component topic model. Because the predictive power of dynamic topic model declines over the future time slides, and comparatively LDA is more general and easy to apply, the dynamic topic model is seldom used in scientometrics research. In recent years, a number of studies on applying LDA to patent data starts to emerge (Suominen et al., 2016). In Suominen et al.' work, modelling the latent topics and temporal pattern separately, using LDA and time series model, provides better flexibility of empirical technological forecasting. However, since no detailed trend shifts or temporal patterns were given, further discussion and development is still needed to learn how to adjust the representation of the temporal dynamics underlying the discovered topics.

2.3. Piecewise linear representation

Capturing the temporal feature of patenting activities requires an abstract representation of real-world dynamics first. Piecewise linear representation (PLR) proposed by Keogh et al. (2001) is one of the most promising time series simplification approaches to reveal latent trend patterns. In recent studies, owing to its ability to decompose data into compressed segments, PLR has been a useful tool in areas such as stock prediction (Chang et al., 2009; Luo and Chen, 2013) and audio signal analysis (Kimura et al., 2008). The piecewise concept was heuristically introduced into technology trend analysis by Philips (1999) in his work using a piecewise linear regression method to capture price changes.

Because shifts and patterns are easier to be observed when data is simplified, PLR is very suitable for catching short-term tendencies and sudden shifts. Many kinds of segmentation algorithms appear under different names in the research of PLR; however, their implementations have slight differences. Most approximation algorithms can be summarized as one of following three types (Keogh et al., 2001, 2004):

- *Top-down*: The time series is recursively partitioned until certain stopping criteria are met.
- *Bottom-up*: Starting from the finest possible approximation, segments are merged until certain stopping criteria are met.
- *Sliding windows*: A segment is grown until it exceeds an error bound. The process repeats with the next data point not included in the newly-approximated segment.

This paper uses a bottom-up algorithm to segment the patent counts series into a number of straight lines. The bottom-up algorithm has been used extensively to support a variety of time series data mining tasks due to efficiency (Keogh and Pazzani, 1998), especially for high level representation of pattern matching systems (Keogh et al., 2004). This algorithm allows the user to specify a desired value for the number of segments, or the total error of the approximation. Specifically, PLR refers to the approximation of a time series P, of length n, with k straight lines. The algorithm begins by creating the finest possible approximation of the original data, which creates n/2 segments to approximate the n-length series. It then calculates the cost of merging each pair of adjacent segments and starts to iteratively merge the pair with lowest cost, until a stopping criterion is met.

3. Methodology

In this section, we illustrate the complete process of our proposed topic-based technological forecasting approach. Detailed explanations for each step and parameter setting are provided in these subsections.

3.1. Framework

This research applies patent titles and claims as the main data source for topic modelling. It is worth mentioning that as one of the most functional parts of the unstructured segments of a patent document, these claims embody all the significant technological features of an invention, the core inventive idea and the most essential technological terms to define the protection of the invention (Novelli, 2015; Tong and Frame, 1994; Yang and Soo, 2012). Moreover, patent claims are concise and clear, and are always described in precise language and certain words (WIPO, 2002; Xie and Miyazaki, 2013), which makes them the best resource for technological content analysis.

We examine the overall framework, input and output of our proposed approach in Fig. 1. After a target technological area has been determined, search statements relating to analytic requirements are passed to USPTO. All patents that belong to the scope are crawled from webpages and added to a corpus waiting for further analysis. Then, the titles and claims of patent documents, their corresponding patent ID, issue dates, United States Patent Classification (USPC), and the patent publication counts for each month are extracted separately as the input of our approach. The claims and title of each patent constitute one .txt document in our corpus, while the patent ID and issue date of all patents compose a single file, USPC information forms a single file as well, and the patent counts are presented as a sequence of data points.

In Fig. 1, all the detailed modules relating to textual data processing are marked in blue. Textual data containing all the patent titles and claims are first passed to several cleaning and consolidation modules to remove the punctuation, meaningless symbols, stop-words, general words used in claims and high frequency academic words. Subsequently, LDA is applied to generate latent topics and topic distribution from the prepared corpus- a unique term list. We then run LDA on the prepared corpus for r times, and apply USPC information to assist with selecting a suitable topic set that better explains the thematic structure of the corpus from multiple experiments. Meanwhile, the detailed modules referring to patent counts processing are marked in green. The patent counts sequence is first normalized for viewing convenience and then passed to a PLR module, where the original observation is decomposed into a number of segments, strengthening and emphasizing the trend patterns underlying the patenting activities of the target area. A trend turning points matrix will be finally generated in the step of trend turning points and trend segments, and passed to topic-based trend coefficient calculation module. We then use the extracted patent issue date information and results from topic modelling to compute the topic annual weight matrix; at the same time, a topic-based trend coefficient sequence is calculated using the topic set and trend turning point matrix, to illustrate to what degree the topic has contributed to the patenting activities of the whole area. Eventually, we conduct the topic-based trend forecasting by analyzing the annual weight variation and trend coefficient changes of the discovered topics. All modules associated with trend forecasting and analysis are marked in red.

3.2. Trend pattern identification

To analyze the temporal trend of various topics and how they contribute to the patenting activities of the whole area, the trend pattern of the target patents needs to be first quantitatively represented. In this research, PLR is applied to detect the trend patterns. Let $P = \{p_1,p_2,\ldots,p_i,\ldots,p_r\}$ be the patent counts over time, where p_i represent

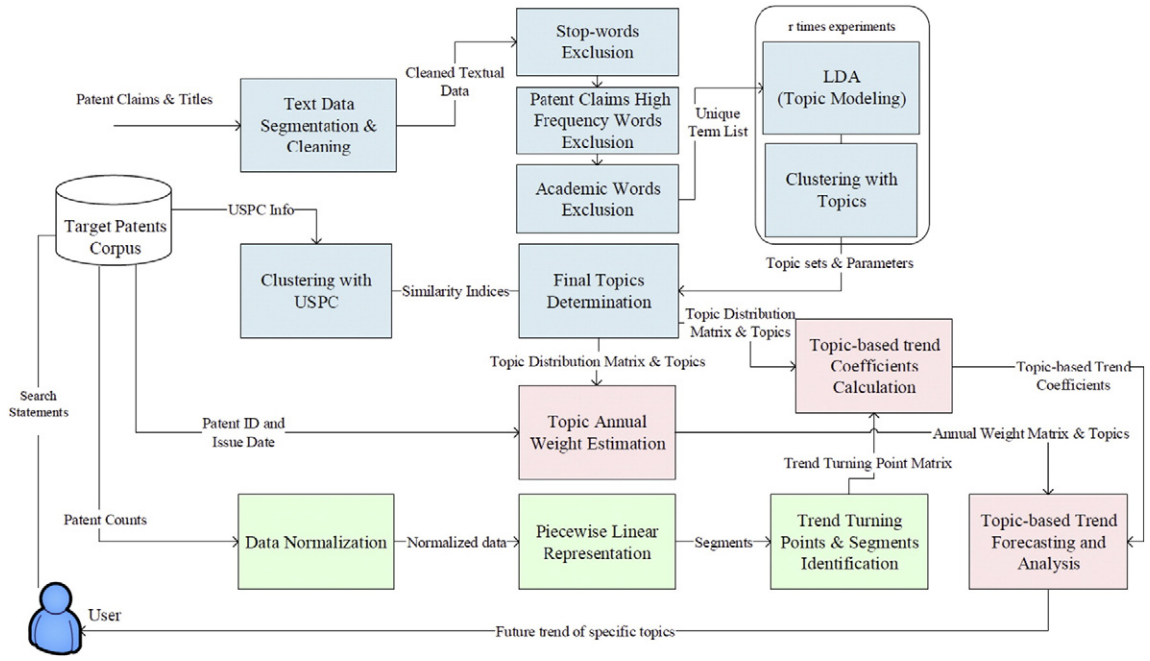

Fig. 1. Framework for the topic-based technological forecasting approach.

the counts of the i^{th} month and r indicates the total month number in m years. P is normalized into $P\sim = \{p\sim_1, p\sim_2, \ldots, p\sim_i, \ldots, p\sim_r\}$ where $P\sim$ has a value between 0.0 and 1.0. Then $P\sim$ is decomposed by PLR into s segments,

$$P\sim_{PLR} = \{L_1(p\sim_1, p\sim_2, \ldots, p\sim_{t_1}), L_2(p\sim_{t_1+1}, p\sim_{t_1+2}, \ldots, p\sim_{t_2}), \ldots, L_i(p\sim_{t_{i-1}+1}, p\sim_{t_{i-1}+2}, \ldots, p\sim_{t_i}), \ldots, L_s(p\sim_{t_{s-1}+1}, p\sim_{t_{s-1}+2}, \ldots, p\sim_r)\},$$

where $P\sim_{PLR}$ denotes the combination of s segments and $L_i(p\sim_{t_{i-1}+1}, p\sim_{t_{i-1}+2}, \ldots, p\sim_{t_i})$ indicates the i^{th} $(1 < i < s)$ segment of $P\sim_{PLR}$ (Keogh et al., 2001). In the same way, $P\sim_{PLR}$ is presented as s straight lines, which present a number of observable trend shifts. Specifically, the joint points between adjacent segments exhibit the detailed change of trends (Chen et al., 2015b). The calculation of trend turning points is presented in the matrix below, where each row of matrix TP indicates a start and an end of a trend state.

$$TP = \begin{bmatrix} 1, & t_1 \\ t_1+1, & t_2 \\ \vdots & \vdots \\ t_{i-1}+1, & t_i \\ \vdots & \vdots \\ t_{s-1}+1, & r \end{bmatrix}$$

Here, the parameter s is a threshold of PLR. It directly affects the sensitivity of the trend pattern extraction. A comparatively smaller value of s produces larger trend segments that present the trend more explicitly, despite slight fluctuations; conversely, a larger value of s makes it more sensitive when trend segments are determined. In existing research, parameter s can be determined by using a Genetic Algorithm (GA) on stock benefit records in stock trading point predictions. However, in the context of patenting trend identification, GA is not suitable because we do not have the 'evaluation criteria', like stock benefit, to evaluate patenting records.

In this research, experiments show that the discrete data of the residual sum of squares (RSS) value between $P\sim_{PLR}$ and $P\sim$ is gradually declining, while s is rising, fast to slow. A smaller s provides more obvious trend features than a larger one. However, it will also produce a quite large RSS, which means the PLR model is less representative. To balance the explicitness of trend shifts and the representability of the model, we select s where the declining rate of RSS starts to obviously slow down, as the preferable one. The approximate derivative (AD) of a series of RSS produced by their corresponding s is calculated for threshold determination,

$$AD_{s_{preferable}} = max\left|\frac{\Delta RSS}{\Delta s}\right|,$$

where $s_{preferable}$ provides the maximum absolute value of AD of the RSS series.

After PLR segmentation, slight jitters are noticeably removed from the original observation. The original data is transformed to s straight lines with only identifiable trend turning points maintained. We then convert $P\sim_{PLR}$ into corresponding trend segments $TS = \{ts_1, ts_2, \ldots, ts_s\}$ to quantitatively depict the temporal pattern of patenting activities. The mean values of straight lines are calculated to present the trend segments between every two trend turning points,

$$TS_i = (ts_{t_{i-1}+1}, ts_{t_{i-1}+2}, \ldots, ts_{t_i}),$$

$$ts_{t_{i-1}+1} = ts_{t_{i-1}+2} = \ldots = ts_{t_i} = mean\, L_i\left(p\sim_{t_{i-1}+1}, p\sim_{t_{i-1}+2}, \ldots, p\sim_{t_i}\right)\},$$

where TS_i denotes the i^{th} $(1 < i < r)$ segment, indicating a trend slice from time $t_{i-1} + 1$ to t_i. The values of all the data points from $ts_{t_{i-1}+1}$ to ts_{t_i} in TS_i, equal the mean value of the i^{th} segment of $P\sim_{PLR}$, $L_i(p\sim_{t_{i-1}+1}, p\sim_{t_{i-1}+2}, \ldots, p\sim_{t_i})$.

Fig. 2 explains the process of transforming original data to trend segments step by step. The x-axis indicates the number of time intervals, for example, days or months; the y-axis stands for the normalized values of patent counts. The transformation between $P\sim_{PLR}$ and TS aggregates and merges data points on a same piecewise linear segment into one trend state, which provides an abstract quantitative representation of the real-world patenting dynamics.

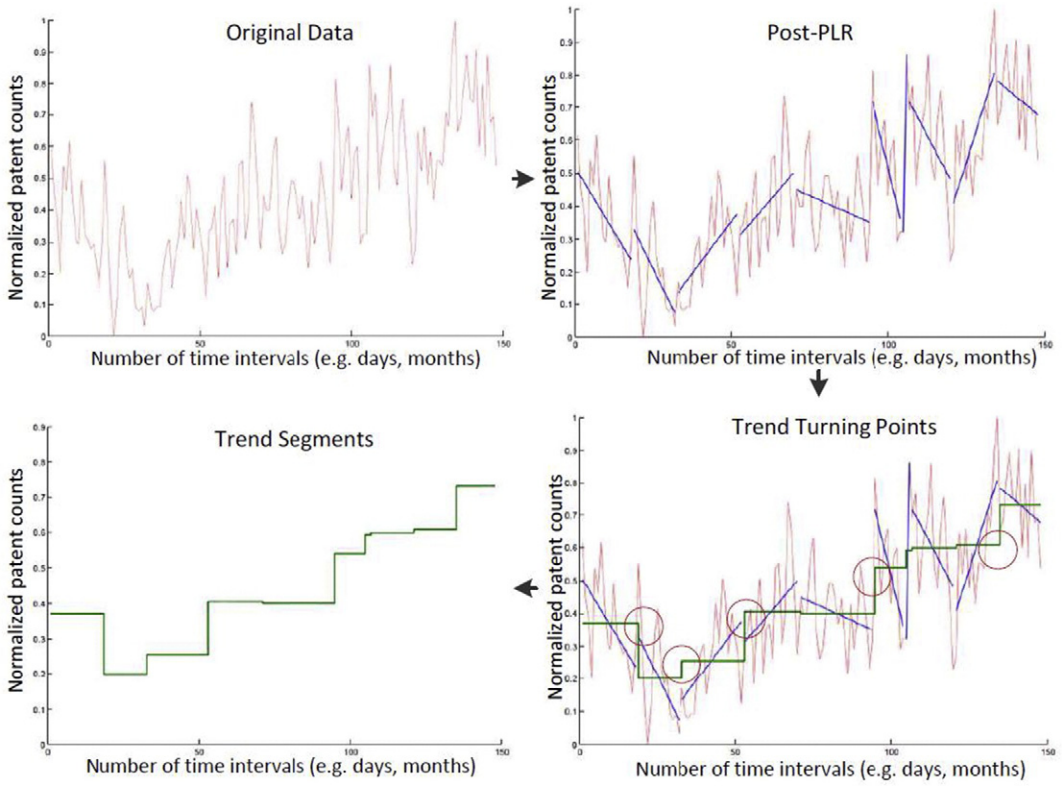

Fig. 2. An example of transforming original data to trend segments step by step.

3.3. Patent topic discovery

3.3.1. Patent claims text cleaning

Patent claims are a special kind of textual data, one can expect plenty of technical terms; they also contain specific words serving as transition phrases and numerous academic-related words that describe invention outcomes. Although they are written with concise, but precise language, the textual data of patent claims still need to be cleaned in order to maintain words that provide the most meaningful information reflecting technological topics only. Before topic discovery, we clean the target claims to retrieve technical terms by performing words cleaning and consolidation.

A patent claim usually consists of three parts: a preamble that serves as an introductory section to outline the primary purpose, function or properties; a transition phrase, such as comprising, having, including, consisting of, and so forth; a "body" that contains the elements or steps that together describe the invention (Sheldon, 1995; USPTO, 2012; Yang and Soo, 2012). Facing its specific characteristic, we use three modules to remove general words from the corpus of patents as follows:

- Stop words such as *the, that, these*;
- High frequency words in patent claims such as *claimed, comprising, invention*;
- General academic words such as *research, approach, data*.

The stop words list we applied is from an information retrieval resources link from Stanford University (Lewis et al., 2004); the patent claim commonly used phrases are summarized from a transitional phrase page on Wikipedia (2014); the general academic words list is provided by the University of Nottingham; we select the top 100 most frequent academic words and remove them from our final corpus (Haywood, 2003).

3.3.2. Topic modelling

From the perspective of probabilistic topic modelling, the corresponding patent document collection is associated with multiple technological topics. The concept 'topic' here is a cluster of words that has a higher possibility of showing up together in a collection of documents. Before topic modelling, we know nothing about the word distributions composing the topics or the topic distributions composing the documents D, so assumptions need to be first drawn to determine the parameters K,α,β of LDA. This research sets $K = 50$, $\alpha = 0.5$ and $\beta = 0.1$ to balance the topical granularity, convenience of understanding and time consumption. We then apply 2000 iterations of Gibbs sampling to infer the needed distributions. Different parameter settings may improve modelling performance, but optimizing these parameters is beyond the scope of this paper.

In practice, Gibbs sampling produces subtly different results each time even with exactly the same input and parameter settings. Facing this problem, USPC is used to help select a comparatively more suitable topic set that better explains the actual thematic structure of the corpus. As a predefined classification hierarchy built on domain expert judgments, although USPC brings subjectiveness into the classification process of patents, it also provides a general understanding of the technical area of concern to one patent. Generally speaking, patents covering similar topics are usually assigned to a same main USPC. Specifically, we denote the main USPC of all d documents in our corpus as $U = (u_1,u_2,u_3,\cdots,u_i,\cdots,u_d)$, where u_i is the USPC of the i^{th} document. After performing each run of LDA, patents are clustered, with their estimated topic distributions Θ and main USPC U, using the hierarchical clustering algorithm (Steinbach et al., 2000). The closer the two clustering results are, the more reliable the topic modelling result is. Specifically, the values of indices Jaccard, Folkes & Mallows and F1 of r times experiments are used to measure the similarity between clustering results based on two different attributes (Halkidi et al., 2001). The three

indices are listed as follows:

$$J = a/a + b + c,$$

$$FM = a/\sqrt{r_1 \cdot r_2},$$

$$F_\beta = \frac{\left(\beta^2 + 1\right) \cdot r_1 \cdot r_2}{\beta^2 \cdot r_1 + r_2},$$

where J stands for Jaccard coefficient, FM indicates Folkes & Mallows index, F_β presents the F1 index. In the equations $r_1 = a/(a + b)$, $r_2 = a/(a + c)$, a represents the number of patents that belong to the same cluster of topics and to the same USPCs, b is the number of patents that are assigned to the same cluster of topics but to different USPCs, and c is the number of patents that are associated with different clusters of topics but to the same subject USPCs. The topic modelling result that provides the highest index values is the optimal one.

3.4. Topic-based trend forecasting and analysis

After topic modelling, we discovered K latent topics expressing D documents, which are presented by their top ranked words, the words' corresponding probabilities, and the topic distribution matrix Θ with D rows and K columns. Each row of the matrix indicates how different topics are distributed over one single document in the corpus, with the summation being equal to 1. The sum values of each column, however, are different. For each topic, the summation of its corresponding column can be seen as an indicator to determine the weight of this topic in the whole topic collection. We select a number of the most weighted topics using the sum of the columns.

Since the patents are issued along a time line, while topic modelling, by processing all the documents with an ascending order of their issue ID, we can obtain a topic distribution matrix in chronological order, as shown in Fig. 3. Then we add up a group of elements in a column that is associated with patents published in the same year, and use the summation to present the annual weight of the corresponding topic. Specifically, we set matrix $W = (w_{ij})_{m \times k}$ to represent the annual weight of all K topics that appeared during m years, where w_{ij} stands for the weight of the j^{th} topic in the i^{th} year.

To estimate the future trend, in a least-squares sense, the annual weight values of each topic are fitted to a univariate quadratic polynomial, $y = ax^2 + bx + c$, where y stands for the topic weight, and x represents the year. We utilize the coefficients a and b to forecast the developing trends of different topics, since a controls the speed of increase (or decrease) of the quadratic function, $-b/2a$ controls the axis of symmetry. For instance, if coefficient a is positive and the symmetry is on the left of the y-axis, we consider the corresponding topic has a growing trend where the greater a is, the faster the growth will be. Table 1 lists the details of using values of a and b to forecast the developing trends of topics. If the value of a is positive and $-b/2a < m$, indicating that the parabola opens up, the future trend of the topic is upward developing, which means it has development potential in the future and has been attracting increasing attentions; when the value of a is positive but $-b/2a < m$, the parabola still opens up, yet the future trend will be downward declining for the corresponding topic, indicating it becomes comparatively less vigorous than other topics in the next few years; if the value of a is negative and $-b/2a < m$, means the parabola opens down, in such case the future trend of the topic is downward declining; when the value of a is negative but $-b/2a < m$, the parabola opens up, under such circumstances, the future trend of the corresponding topic is upward growing.

Table 1
Trend forecasting indicators and future trend estimation.

Value of a	Symmetry	Future trend
Positive	$-b/2a < m$	Upward
Positive	$-b/2a > m$	Downward
Negative	$-b/2a < m$	Downward
Negative	$-b/2a > m$	Upward

Furthermore, for more specific trend analysis, we then integrate the identified trend segments and discovered topics, to compute a sequence of contribution coefficients and evaluate how different topics contributed to the patenting activities of the whole target area, as shown in Algorithm 1.

Algorithm 1. Topic-based trend coefficients estimation.

input: Trend turning points matrix TP and topic distributions Θ

output: A sequence of topic-based trend coefficients for each prominent topic (n topics), TC

input: Trend turning points matrix TP and topic distributions Θ
output: A sequence of topic-based trend coefficients for each prominent topic (n topics), TC
set $ws_k = \sum_{i=1}^{d} \vartheta_{ik}$
select n topics with top largest ws_k as prominent topic set $\vec{N}$
set Θ in chronological order
for topic n in $\vec{N}$
    $tc_{in} = \sum_{t_{i-1}+1}^{t_i} \vartheta_{in}$
    where $[t_{i-1} + 1, t_i]$ is the i^{th} row of matrix TP
end for
$TC_n = (tc_1, tc_2, tc_3, \ldots, tc_s)$
end

For the n^{th} selected topics, let $TC_n = (tc_1,tc_2,tc_3,\ldots,tc_s)$ be the contribution coefficients, where tc_s indicates the topic weight on the s^{th}

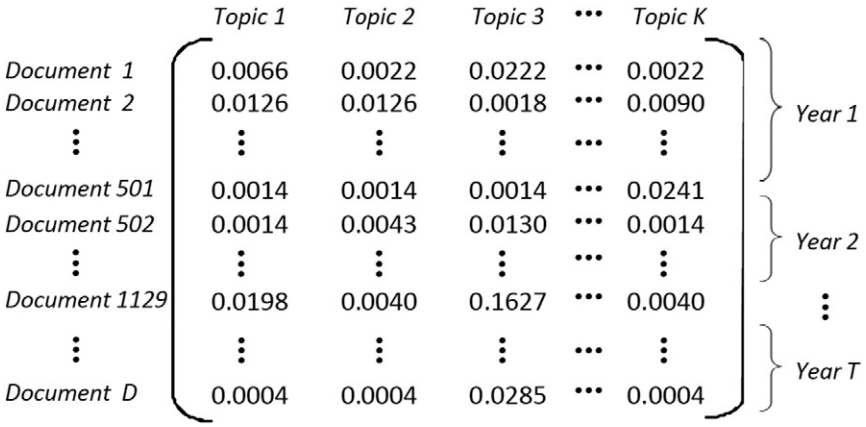

Fig. 3. An example of a topic distribution matrix in chronological order.

trend segment. These topic-based trend coefficients are used to serve the detailed analysis of the historical topic trend, thus revealing the most and least contributing trend segments, which integrates the temporal patterns of patenting activities and semantic topics together to provide topic-based technological trend explanation.

4. Case study and discussion

Our goal is to forecast the developing trend of specific topics underlying a large volume of patent documents, and find to what degree each topic has contributed to the patenting activities of the whole area. In this section, a case study using USPTO utility patent is provided to demonstrate the effectiveness of the proposed topic-based technological forecasting approach.

4.1. Data collection

Utility patents published during the years 2000 to 2014 in USPTO (http://www.uspto.gov/) with Australia as their assignee country were selected as our target patents. Their patent ID, titles, issue time, inventors, assignee, United States Patent Classification (USPC) and most importantly, their claims, were crawled from USPTO and placed in a patent database for further processing. In total, we collected 13,910 utility patents covering 374 different main USPC. The IDs and the issue time of all the target patents formed one single file, while the claims and title for each patent constitute one document in our corpus, with a total of 13,910 documents. Altogether, in the target corpus, we found 103,935 unique vocabularies containing the technological topics of inventions owned by Australian assignees over the past 15 years.

4.2. Trend pattern identification

We collected the published patent counts for each month to generate a counts sequence and normalized it to values between 0.0 and 1.0. After calculating the approximate derivative of a series of RSS produced by segment numbers from 3 to 22, we chose the value that produced the maximum absolute value of the approximate derivative, $s = 5$, as the optimal pieces number. As shown in Fig. 4, the normalized data was decomposed into five trend segments to quantitatively reveal and highlight a group of main trend shifts. In the figure, the original observation is displayed with blue lines, while the PLR segments are marked in red, and the final trend segments are illustrated with green lines.

We can observe directly from Fig. 4 that the trend turning points are January 2006, January 2009, September 2010 and March 2012. On the whole, the trends for patents owned by Australian assignees have experienced an approximate ladder-type growth. In the six years between 2000 and 2005, the trend maintains a low and stable status. Then, an important trend turning point appeared in January 2006 when a sharp upward transition occurred, implying a breakthrough in R&D activities or the expansion of existing technological topics. After this trend turning point, the publication of patents almost doubled. In January 2009, the next trend turning point occurred, indicating another round of rapid growth in patents. From September 2010 to February 2012, lasting one and half years, the trend has reached a peak for the time being, and has started to decline. In the follow-up trend segment, from March 2012 to the end of 2014, the main trend declined to the level of approximately three years ago, implying the importance of some technological topics has diminished. Table 2 illustrates the details of all trend turning points, trend segments and the document numbers belonging to each trend segment. Trend segment 1 covered more documents than the others with 3,706 patents; trend segment 3 contained the least a number of documents, with a total of 1,898 patents.

4.3. Topic modelling and prominent topic selection

After identifying the trend turning points of the whole area, we continued to process the textual data. The stop words, and the high-frequency common phrases used in patent claims and general academic vocabularies were first excluded from our data collection. We applied LDA parameters $K = 50$, $\alpha = 0.5$ and $\beta = 0.1$ to conduct topic modelling. In total, we performed 5 ($r = 5$) runs, with 2000 iterations of Gibbs sampling to decide the final topic set. After clustering all 13,910 patents using both topic distributions and their USPC, we selected the trial with highest values of Folkes & Mallows, Jaccard, and F1 similarity indices. Fifty latent semantic topics were estimated, and each was presented by the top 10 ranked words and their corresponding probabilities. For reading convenience, the details of all topics, the top 10 ranked words and the probabilities of each word belonging to a topic are listed in Table 1 of the Appendix A.

The top 10 most weighted topics are selected using the topic distribution matrix. Table 3 lists the topic weight of all estimated 50 topics, and highlights the selected 10 topics in bold. These prominent topics in utility patents owned by Australian assignees over the past 15 years includes: printhead (topic 37), nozzle (topic 12), axis drive shaft (topic 17), wall body (topic 5), sensing device (topic 6), fluid valve

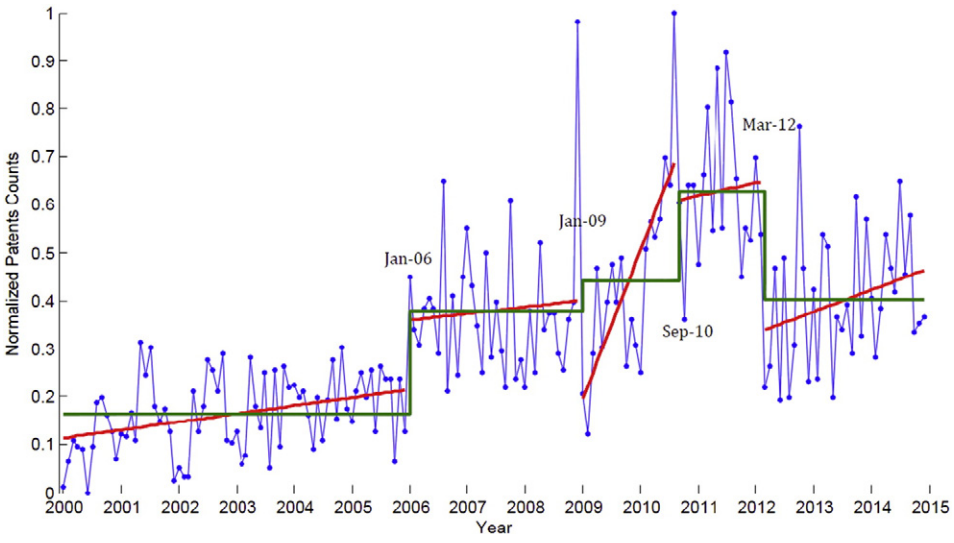

Fig. 4. The trend turning points and trend segments generated from patenting activities.

Table 2
The trend turning points, document numbers, and term numbers for each trend segment.

Trend segment	Trend turning start	Trend turning end	Trend segment value	Doc no.
1	Jan-00	Dec-05	0.163	3706
2	Jan-06	Dec-08	0.379	3064
3	Jan-09	Aug-10	0.442	1898
4	Sep-10	Feb-12	0.628	2232
5	Mar-12	Dec-14	0.401	3010

(topic 10), amino acid sequence (topic 15), composite material (topic 31), antibody composition (topic 33) and signal & circuit (topic 13).

4.4. Topic annual weight matrix and topic-based trend coefficients estimation

Since the 13,910 utility patent claims documents we crawled from USPTO were published following a strict time line, the smaller the patent ID, the earlier it was published. In the case study, we named all the files according to their patent ID. While topic modelling, the documents were processed in ascending order of their name tag. After identifying both prominent topics and trend turning points, we then generated the annual weight matrix to illustrate annual weight changes in each topic, as shown in Table 4.

To further evaluate how the various topics contributed to the patenting activities of the entire target area, we calculated the topic-based trend coefficients using Algorithm 1 and provide the results in Table 5, where TSV stands for trend segment value, TS1 indicates trend segment 1, T37 means Topics 37, and so on. Although some segments cover comparatively more documents than others, not all topics contribute to these segments significantly. For example, trend segment 5 contains 3010 documents and its segment value is higher than segments 1 and 2, yet the contribution of topic 37 to segment 5 is only 22.91, which is much lower than its contribution to segments 1 and 2. In summary, different latent topics contribute differently to the trend changes of patenting activities differently, and we can use these topic-based trend coefficients to measure the varying degrees of their involvement.

4.5. Topic-based trend forecasting and analysis

The latent topics we generated from the document collection have their very own trends and different contribution levels to the patenting activities of the whole area. We then forecast the weight changes of each prominent topic and forecast their future trend in a least-squares prospective. Fig. 5 presents the fitting curve for the10 selected prominent topics.

We can observe from the figure that printhead (topic 37) and nozzle (topic 12), were two more important topics that Australian assignees owned in the past 15 years; both experienced a high speed development stage and showed a downward trend between the 2012 and 2014. The graph for topic 37 appears more closed than topic 12, indicating that it experienced greater variation while increasing and decreasing. On the whole, these two topics have just gone through a boom period, and they may become comparatively less vigorous than other topics in the next few years. The significance of topics axis drive shaft (topic 17) and wall body (topic 5), on the contrary, are gradually growing, indicating that the two topics have development potential in the future. Among the rest of the topics, sensing device (topic 6) and composite material (topic 31) have a downward trend. Yet, the decline of topic 6 was more dramatic than topic 31, which largely remained steady with just a slight reduction. Fluid valve (topic 10), amino acid sequence (topic 15), antibody composition (topic 33) and signal & circuit (topic 13) all show upward growing trends. In particular, the topic importance of antibody composition has a faster increasing trend than other topics. It displayed quite obvious growth in the past five years, indicating it has the potential to continue to grow in future patent publications.

In Table 6, we examine all the quadratic polynomial fitting coefficients, and provide a summary of the topic-based trend forecasting and trend segments that a topic most and least contributed to. As mentioned, an important trend turning point appeared in January 2006 when a sharp upward transition occurred, which implies expansion or breakthroughs for existing technological topics. Since topics 37, 12 and 6 all significantly contributed to trend segment 2, and more than any other segment, we learn that the development of printhead, nozzle and sensing device from years 2006 to 2009 increased patent publications for the whole area. The significance of these three topics, however,

Table 3
The top 50 topics generated from the patent claims collection and their weight indicator.

Topic no.	Topic	Weight	Topic no.	Topic	Weight
Topic 37	Printhead	1250.873	Topic 30	Pressure vent	193.640
Topic 12	Nozzle	824.743	Topic 28	Laser beam	187.874
Topic 17	Axis drive shaft	751.278	Topic 40	Plunger module	176.804
Topic 5	Wall body	750.291	Topic 18	Optical fibre	175.717
Topic 6	Sensing device	720.324	Topic 29	Vehicle break	175.428
Topic 10	Fluid valve	539.187	Topic 44	Heart rate sensor	173.578
Topic 15	Amino acid sequence	485.246	Topic 49	Nucleic acid	171.474
Topic 31	Composite material	467.351	Topic 3	Structure roof	167.466
Topic 33	Antibody composition	441.376	Topic 25	Radiation detector	160.021
Topic 13	Signal&Circuit	384.553	Topic 43	Optical lens	153.859
Topic 7	Polymer agent	363.251	Topic 20	Temperature control	145.019
Topic 39	Support frame	352.747	Topic 45	Solar heat	140.127
Topic 50	Vessel material	338.813	Topic 47	Memory search	132.249
Topic 16	Gaming controller	334.080	Topic 27	Semiconductor	112.792
Topic 23	Camera image	314.639	Topic 24	3d Fin	111.528
Topic 36	Alkyl compound	314.055	Topic 41	Magnetic impeller	102.972
Topic 9	Resin material	258.562	Topic 34	Glyphosate formulation	100.900
Topic 22	Transmission security	251.222	Topic 32	Oligonucleotide	98.174
Topic 21	Electrode carrier	232.779	Topic 14	Delivery conveyor	97.169
Topic 35	Wireless communications	228.530	Topic 19	Explosives	97.017
Topic 46	Conduction device	222.701	Topic 48	Headgear/strap	94.996
Topic 4	Channel symbol	217.575	Topic 2	Humidifier	84.664
Topic 26	Respiratory connector	213.356	Topic 38	Benzyl illumination	82.162
Topic 42	Tubular actuator	200.995	Topic 8	c.sub.1-c.sub.10	68.461
Topic 1	Hearing prosthesis	196.602	Topic 11	Payment settlement	50.790

Table 4
The annual weight matrix of the selected top 10 significant topics.

Year	Topic37	Topic12	Topic17	Topic 5	Topic 6	Topic10	Topic15	Topic31	Topic33	Topic13
2000	4.948	7.146	45.062	44.740	3.522	33.725	21.226	16.337	16.445	12.145
2001	20.274	41.399	47.202	44.355	5.883	31.987	39.488	25.476	27.685	12.064
2002	22.955	40.158	36.791	38.311	5.791	33.869	27.871	22.267	24.774	14.991
2003	31.666	33.218	44.023	38.164	10.640	31.137	23.996	29.267	20.817	23.503
2004	34.048	32.019	43.477	41.368	34.697	30.968	15.031	31.106	17.740	24.346
2005	62.299	47.867	40.804	46.256	35.969	31.569	18.023	30.142	12.659	19.324
2006	126.983	98.659	49.701	42.766	72.082	37.298	31.962	40.343	20.720	30.054
2007	150.042	84.661	50.609	40.772	77.192	30.371	29.372	35.444	20.144	30.130
2008	199.104	98.232	54.905	39.844	74.694	32.522	23.483	30.705	30.166	24.988
2009	130.780	91.007	44.239	39.792	60.803	26.813	38.640	28.134	29.481	25.073
2010	223.097	101.882	62.789	65.746	101.605	52.418	40.633	40.568	35.724	33.633
2011	197.481	120.112	56.470	69.065	143.483	62.093	44.615	49.901	38.413	36.146
2012	36.276	18.149	64.982	70.585	67.854	38.180	44.192	28.496	43.917	27.281
2013	8.216	6.198	54.834	64.382	21.440	32.744	37.700	28.542	50.176	35.181
2014	2.702	4.038	55.390	64.145	4.666	33.492	49.016	30.623	52.516	35.693

all dropped quickly on the fifth trend segment, which indicates that their developing potential, compared with other topics, is limited. Topic 33, antibody composition, appeared to have quite an opposite trend. It contributed mainly to the last trend segment. Since 2005, the significance of this topic started growing continuously, from which we learn that the research and patenting for the topic of antibody composition is increasing over the past 15 years, and this topic has the most potential among all generated latent topics. Specifically, this topic related to: human antibody, peptide binding, peptide fragment and peptide bond amino acid. Details of the topic content can be found in the Appendix A.

4.6. Discussion

A patent document collection is actually associated with multiple underlying technological topics. These latent topics have their very own trends and different contribution levels to the patenting activities of the whole area. If we only model the trend shift of the whole area year or season, it is very difficult to learn the trend patterns of a technological topic. From a methodological perspective, the main contributions of this paper are: (1) it proposes a stepwise approach to quantitatively identify temporal trend patterns and semantic topics, that integrates these two features in different dimensions, to provide topic-based technological trend forecasting; (2) this research estimates the developing trends for specific latent topics, rather than a broad technological area, that also evaluates to what degree various topics have contributed to the patenting activities of an entire area and how it will perform in the future.

From an application viewpoint, the proposed topic-based trend forecasting approach can be used to automatically uncover the thematic structure of massive patent data in a technological area of interest, and estimate the detailed developing trends of each detected topic with a forward-looking estimation, thereby assisting decision making for potential opportunity identification, technical strategy formation, and decision making support. For instance, a full understanding of the underlying technological topics distribution and trends in the target area are essential for both newly created innovative enterprises and venture capitalists (VCs). This understanding enables entrepreneurs to prepare appropriate technical proposals with potential while at the same time providing VCs with the confidence to support companies with a better understanding of the current situation in a certain industry (Holst et al., 2010).

Potentially, the proposed approach can be applied to assist in building content-based indicator for radical innovation. Comparing with incremental technology development, the radical innovation will reshape the well-defined and predicable trajectories (Arts et al., 2013), which is highly possible to show sudden shifts in a trend. The trend turning point and topic annual weight matrix proposed in this paper open a window of opportunity to first detect the sudden shift like sharp increase, then further track the trend of all related topics, and eventually identify the one topic or several topics leading the sudden shifts. In addition, the topic-based trend provides a link between patenting activities and content of patents, thus makes it possible to use the rich textual data in patent documents to support technological change discovery.

There are also some possible limitations of using the content-based indicator. In the previous studies, one patent indicates one radical innovation, however in the perspective of topic modelling, what we can identify, is a radical topic. The precondition of successfully identifying a novel topic based on its trend is that, we have discovered this topic in the first place. In a document collection, the strong topics usually are the mainstream topics that have been discussed a lot. Thus it will be very important to construct comparatively more refined document collection as the input, and adjust the topic granularity in the empirical analysis, to increase the possibility of catching and presenting all the novel topics. In the existing studies, Verhoeven et al. (2016) has conceptualized technological novelty and applied patent-based indicators based on classification and citation information for technological breakthrough identification. Combining the patent-based and content-based indicator will be potentially very helpful to better delimitate the target patent documents and detect important technological inventions.

5. Conclusions and future work

With technological advances and the accumulation of patent publications, manually conducting content analysis and trend forecasting

Table 5
Topic-based trend coefficients for all 10 prominent topics.

	Doc no.	TSV	T37	T12	T17	T 5	T6	T10	T15	T31	T33	T13
TS 1	3706	0.163	176.19	201.81	257.36	253.19	96.50	193.26	145.63	154.60	120.12	106.37
TS 2	3064	0.379	476.13	281.55	155.22	123.38	223.97	100.19	84.82	106.49	71.03	85.17
TS 3	1898	0.442	283.31	171.00	86.27	88.69	121.73	64.48	65.78	55.15	51.81	45.57
TS 4	2232	0.628	292.33	155.71	91.48	98.01	211.58	86.87	65.71	69.88	56.69	56.66
TS 5	3010	0.401	22.91	14.67	160.96	187.01	66.54	94.39	123.31	81.23	141.73	90.77

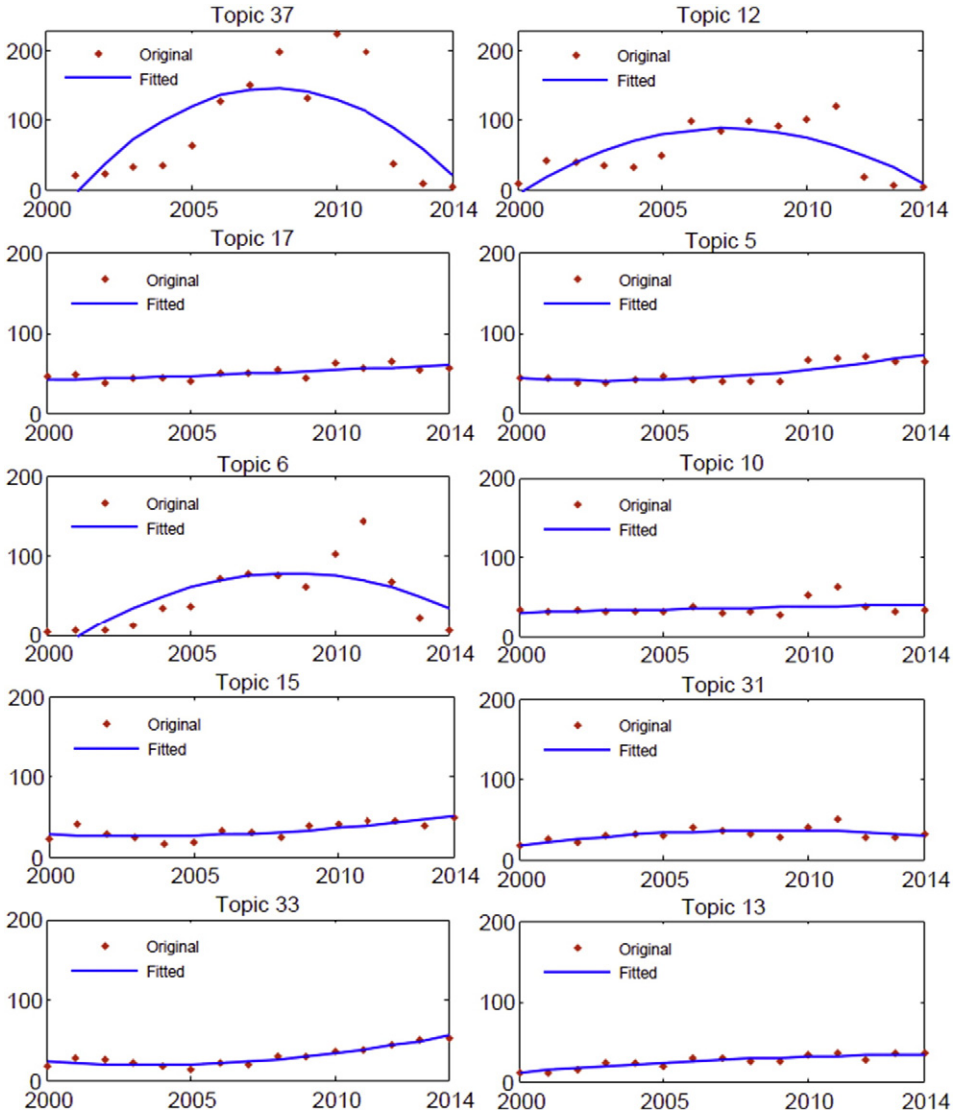

Fig. 5. The curve fitting result for topic trend estimation of the 10 selected topics.

on massive patent documents has become increasingly time-consuming and laborious. Thus automatic topic discovering and trend estimating approaches will continue to be emphasized. This research proposes an empirical topic-based technological forecasting approach to generate topics from massive patent claims documentation, then forecast their very own trends and different contribution levels for the patenting activities of the entire target area.

Our future work will continue to focus on discovering and analyzing the development and changes in the estimated latent semantic topics, and also further discuss the possibility and detailed method to apply the topic-based forecasting approach to radical innovation identification. In addition, while developing topic-based trend forecasting, we observed that some topics shows very similar trend patterns and turning points, which implies they may have strong correlations on a semantic

Table 6
The details of trend estimation for the 10 selected topics.

Topic no.	Topic	a	b	Symmetry	Most contributing	Least contributing	Topic-based future trend
Topic 37	Printhead	−3.25	56.99	8.78	TS 2	TS 5	Downward
Topic 12	Nozzle	−1.76	29.25	8.30	TS 2	TS 5	Downward
Topic 17	Axis drive shaft	0.03	0.87	−14.33	TS 1	TS 3	Upward
Topic 5	Wall body	0.26	−2.03	3.89	TS 1	TS 3	Upward
Topic 6	Sensing device	−1.45	27.52	9.47	TS 2	TS 5	Downward
Topic 10	Fluid valve	−0.02	1.07	23.52	TS 1	TS 3	Upward
Topic 15	Amino acid sequence	0.21	−1.64	4.01	TS 1	TS 4	Upward
Topic 31	Composite material	−0.24	4.73	9.83	TS 1	TS 3	Downward
Topic 33	Antibody composition	0.32	−2.75	4.34	TS 5	TS 3	Upward
Topic 13	Signal&Circuit	−0.09	3.03	17.01	TS 1	TS 3	Upward trend

level. We will further explore the topic network in our future research. Since judgments on certain states, relations or tendencies are often expressed by linguistic terms in a real-life situation, such as 'growing', 'stable', 'have potential' and so forth, we will also introduce fuzzy sets to deal with the vagueness of these terms in future work to provide a better understanding of the trend estimation.

Acknowledgement

The work presented in this paper is partly supported by the Australian Research Council (ARC) under Discovery Project DP140101366 and the National High Technology Research and Development Program of China (Grant No. 2014AA015105).

Appendix A

Table 1
The top 10 ranked words of 50 topics in USPTO patents with Austrian assignees from 2000 to 2014 and their corresponding probabilities.

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Configured	0.0347	Humidifier	0.0483	Structure	0.0856	Values	0.0272	Wall	0.0241
Hearing	0.0337	Flow	0.0412	Roof	0.0216	Symbol	0.0173	Upper	0.0202
Signal	0.0320	Respiratory	0.0335	usb	0.0162	Channel	0.0153	Surface	0.0199
Stimulation	0.0318	Tub	0.0299	Plurality	0.0155	Time	0.0141	Body	0.0183
Prosthesis	0.0280	Generator	0.0245	Mirror	0.0122	Plurality	0.0110	Container	0.0176
Audio	0.0253	Configured	0.0227	Clock	0.0121	Parameter	0.0106	Lower	0.0170
Signals	0.0158	Lid	0.0186	Coupled	0.0118	Model	0.0105	Outer	0.0135
Sound	0.0155	Apparatus	0.0181	Tunnel	0.0116	Vector	0.0103	Panel	0.0134
Level	0.0150	Heater	0.0174	Barrier	0.0109	Parameters	0.0095	Substantially	0.0116
Auditory	0.0137	Base	0.0157	Respective	0.0100	Function	0.0095	Adjacent	0.0108

Topic 6		Topic 7		Topic 8		Topic 9		Topic 10	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Coded	0.0401	Composition	0.0584	Series	0.0163	Material	0.0114	Fluid	0.0369
Sensing	0.0388	Weight	0.0194	–c.sub.1-c.sub.10	0.0140	lrc	0.0095	Valve	0.0309
Device	0.0359	Agent	0.0161	Graphical	0.0105	Tread	0.0089	Chamber	0.0301
Identity	0.0179	Polymer	0.0145	Position	0.0090	Gravity	0.0088	Flow	0.0283
Indicative	0.0165	Water	0.0084	–c.sub.2-c.sub.20	0.0089	Ion	0.0086	Inlet	0.0246
Position	0.0148	Amount	0.0080	Sensor	0.0089	Water	0.0082	Air	0.0231
Indicating	0.0143	Gel	0.0072	Alkyl	0.0082	Concentration	0.0082	Water	0.0212
Interface	0.0128	Acid	0.0066	Independently	0.0079	Acid	0.0078	Outlet	0.0176
Product	0.0109	Active	0.0062	Output	0.0074	Leach	0.0073	Gas	0.0170
Surface	0.0097	Polymeric	0.0048	Alkenyl	0.0073	Resin	0.0054	Line	0.0157

Topic 11		Topic 12		Topic 13		Topic 14		Topic 15	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Payment	0.0803	Nozzle	0.0434	Signal	0.0913	Mode	0.0403	Sequence	0.0554
Settlement	0.0773	Ink	0.0371	Frequency	0.0252	Delivery	0.0187	Acid	0.0535
Customer	0.0650	Ejection	0.0286	Signals	0.0220	Consumer	0.0185	Plant	0.0299
Bank	0.0426	Heater	0.0232	Output	0.0216	Therapy	0.0175	Cell	0.0254
Amount	0.0402	Actuator	0.0227	Input	0.0172	Dormant	0.0148	Amino	0.0195
Funds	0.0360	Printhead	0.0211	Circuit	0.0153	Conveyor	0.0135	Nucleic	0.0176
Incentive	0.0227	Inkjet	0.0198	Control	0.0130	Device	0.0130	Nucleotide	0.0141
Agreement	0.0226	Chamber	0.0167	Electrical	0.0116	Marketing	0.0128	Protein	0.0140
Message	0.0179	Drop	0.0146	Digital	0.0090	Time	0.0126	Molecule	0.0125
Bank	0.0157	Substrate	0.0125	Phase	0.0081	Sampling	0.0122	Isolated	0.0096

Topic 16		Topic 17		Topic 18		Topic 19		Topic 20	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Game	0.0762	Position	0.0188	Apparatus	0.2790	Material	0.0425	Power	0.0610
Gaming	0.0578	Drive	0.0167	Optical	0.0589	Blasting	0.0406	Temperature	0.0543
Controller	0.0320	Mounted	0.0117	Fibre	0.0217	Body	0.0356	Predetermined	0.0375
Plurality	0.0317	Axis	0.0112	Plurality	0.0189	Blast	0.0248	Control	0.0372
Symbols	0.0257	Movement	0.0108	Communication	0.0117	Explosives	0.0165	Heated	0.0307
Award	0.0221	Shaft	0.0099	Waveguide	0.0108	Respective	0.0159	Heating	0.0298
Symbol	0.0217	Rotation	0.0099	Enclosure	0.0106	Biometric	0.0153	Supply	0.0289
Machine	0.0189	Mechanism	0.0090	Joint	0.0091	Test	0.0152	Voltage	0.0232
Player	0.0182	Locking	0.0087	Detection	0.0089	Detonator	0.0115	Pap	0.0210
Outcome	0.0176	Housing	0.0084	Light	0.0087	Blastholes	0.0106	Circuit	0.0210

Topic 21		Topic 22		Topic 23		Topic 24		Topic 25	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Electrode	0.0571	Security	0.0167	Image	0.0809	Fin	0.0289	Material	0.0295
Carrier	0.0284	Code	0.0165	Camera	0.0142	Plug	0.0251	Particles	0.0292
Field	0.0245	Transmission	0.0163	Capturing	0.0094	Viewing	0.0160	Trailer	0.0213
Coil	0.0212	Authentication	0.0153	Video	0.0082	Space	0.0155	Radiation	0.0166
Array	0.0205	Key	0.0151	Input	0.0076	Plurality	0.0153	Detector	0.0142
Imaging	0.0197	Stored	0.0117	Depth	0.0075	Grid	0.0149	Particulate	0.0089

(continued on next page)

Table 1 (*continued*)

Topic 21		Topic 22		Topic 23		Topic 24		Topic 25	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Implantable	0.0146	Memory	0.0117	Capture	0.0075	3d	0.0115	Microwave	0.0088
Surface	0.0144	Terminal	0.0115	Feature	0.0066	Processor	0.0107	Size	0.0084
Magnetic	0.0120	Secure	0.0111	Overview	0.0064	Visual	0.0101	Unit	0.0080
Contact	0.0115	Message	0.0099	Pixel	0.0059	Module	0.0086	Carrying	0.0072

Topic 26		Topic 27		Topic 28		Topic 29		Topic 30	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Mask	0.0423	Cardiac	0.0186	Light	0.0451	Vehicle	0.0487	Pressure	0.0528
Connector	0.0311	Semiconductor	0.0132	Beam	0.0303	Wheel	0.0297	Vent	0.0277
Patient	0.0307	Regions	0.0132	Laser	0.0214	Plurality	0.0257	Mask	0.0250
Interface	0.0237	Time	0.0107	Source	0.0190	Speed	0.0155	Flow	0.0245
Cushion	0.0214	Conductive	0.0104	Plurality	0.0107	Brake	0.0106	Cpap	0.0237
Nasal	0.0178	Cardiogenic	0.0088	Dicarba	0.0080	Graphics	0.0102	Insert	0.0218
Frame	0.0149	Filament	0.0082	Attribute	0.0078	Suspension	0.0098	Apparatus	0.0177
Respiratory	0.0132	Storage	0.0077	Database	0.0065	Load	0.0097	Treatment	0.0173
Seal	0.0127	Terminals	0.0072	Band	0.0065	Arrangement	0.0093	Patient	0.0169
Strap	0.0126	Light	0.0070	Materials	0.0061	Respective	0.0093	Gas	0.0158

Topic 31		Topic 32		Topic 33		Topic 34		Topic 35	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Layer	0.0590	Antisense	0.0407	Antibody	0.0290	Input	0.0399	Network	0.0352
Material	0.0336	Oligonucleotide	0.0319	Composition	0.0203	Formulation	0.0312	Communications	0.0259
Metal	0.0234	Executable	0.0314	Peptide	0.0146	Glyphosate	0.0305	Wireless	0.0216
Substrate	0.0186	Code	0.0293	Human	0.0141	Term	0.0192	Node	0.0187
Surface	0.0171	Combination	0.0290	Fragment	0.0122	Solid	0.0189	Access	0.0110
Formed	0.0110	Fragments	0.0171	Binding	0.0120	Local	0.0153	File	0.0098
Composite	0.0099	Charge	0.0162	Subject	0.0108	Acid	0.0141	Received	0.0096
Layers	0.0094	Battery	0.0153	Amino	0.0098	Index	0.0133	Plurality	0.0093
Forming	0.0093	Determined	0.0148	Administering	0.0093	Basis	0.0125	Service	0.0087
Membrane	0.0088	Modified	0.0112	Amount	0.0084	Equipment	0.0116	Location	0.0077

Topic 36		Topic 37		Topic 38		Topic 39		Topic 40	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Substituted	0.0589	Printhead	0.0618	Block	0.0257	Support	0.0657	Module	0.0354
Compound	0.0422	Ink	0.0549	Substituted	0.0242	Frame	0.0345	Plunger	0.0279
Formula	0.0159	Print	0.0406	Illumination	0.0235	Base	0.0198	Carrier	0.0206
Alkyl	0.0113	Printer	0.0331	Intermediate	0.0226	Mask	0.0170	Spiral	0.0198
Independently	0.0110	Media	0.0302	Protecting	0.0224	Edge	0.0153	Tube	0.0197
Aryl	0.0070	Cartridge	0.0148	Position	0.0193	Forehead	0.0146	Barrel	0.0195
Pharmaceutically	0.0069	Integrated	0.0146	Groups	0.0191	Arm	0.0121	Cartridge	0.0178
Salt	0.0066	Unit	0.0123	Benzyl	0.0178	Clip	0.0103	Needle	0.0138
Acceptable	0.0065	Plurality	0.0111	Switch	0.0168	Locking	0.0095	Separator	0.0136
Composition	0.0051	Configured	0.0101	Tilt	0.0161	Pair	0.0093	Syringe	0.0115

Topic 41		Topic 42		Topic 43		Topic 44		Topic 45	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Magnetic	0.0487	Tubular	0.0490	Zone	0.0424	Sensor	0.0821	Energy	0.0419
Impeller	0.0297	Tool	0.0373	Lens	0.0402	Change	0.0246	Heat	0.0377
Bearing	0.0253	Elongate	0.0219	Optical	0.0396	Rate	0.0189	Medium	0.0178
Pump	0.0234	Distal	0.0215	Lifting	0.0192	Condition	0.0175	Solar	0.0169
Axial	0.0169	Body	0.0189	Central	0.0189	Heart	0.0157	Wall	0.0147
Position	0.0137	Actuator	0.0168	Surface	0.0179	Reservation	0.0152	Transfer	0.0147
Cavity	0.0122	Steering	0.0154	Peripheral	0.0178	Processor	0.0138	Collection	0.0135
Packaging	0.0114	Sheath	0.0130	Eye	0.0154	Failure	0.0138	Body	0.0099
Plurality	0.0104	Handle	0.0130	Power	0.0132	Configured	0.0124	Exchanger	0.0089
Heart	0.0098	Fastener	0.0125	Mantle	0.0127	Indicator	0.0122	Regulating	0.0087

Topic 46		Topic 47		Topic 48		Topic 49		Topic 50	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Device	0.2205	Application	0.0480	Portions	0.0421	Sample	0.0376	Material	0.0205
Surface	0.0232	Output	0.0331	Strap	0.0356	Nucleic	0.0190	Gas	0.0187
Bone	0.0186	Feature	0.0272	Vent	0.0286	Control	0.0155	Stream	0.0123
Mold	0.0157	Event	0.0237	Headgear	0.0231	Acid	0.0144	Vessel	0.0117
Conduction	0.0140	Search	0.0171	Media	0.0152	Precession	0.0111	Carbon	0.0113
Rod	0.0129	Memory	0.0150	Holes	0.0130	Primer	0.0104	Metal	0.0105
Configured	0.0123	Recorded	0.0135	Titanium	0.0129	Target	0.0094	Feed	0.0102
Central	0.0111	Parameters	0.0132	Top	0.0127	cpg-containing	0.0079	Liquid	0.0084
Liquid	0.0110	Representing	0.0127	Front	0.0120	Substance	0.0077	Treatment	0.0066
Seal	0.0101	Respective	0.0118	Flow	0.0111	Piece	0.0075	Water	0.0060

References

Abbas, A., Zhang, L., Khan, S.U., 2014. A literature review on the state-of-the-art in patent analysis. World Patent Inf. 37, 3–13.

Arts, S., Appio, F.P., Van Looy, B., 2013. Inventions shaping technological trajectories: do existing patent indicators provide a comprehensive picture? Scientometrics 97 (2), 397–419.

Baskurt, O., 2011. Time series analysis of publication counts of a university: what are the implications? Scientometrics 86 (3), 645–656.

Bengisu, M., Nekhili, R., 2006. Forecasting emerging technologies with the aid of science and technology databases. Technol. Forecast. Soc. Chang. 73 (7), 835–844.

Blei, D.M., 2012. Probabilistic topic models. Commun. ACM 55 (4), 77–84.

Blei, D.M., Lafferty, J.D., 2006. Dynamic topic models. Proceedings of the 23rd International Conference on Machine Learning, ACM, pp. 113–120.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Campbell, R.S., 1983. Patent trends as a technological forecasting tool. World Patent Inf. 5 (3), 137–143.

Carrillo, M., González, J.M., 2002. A new approach to modelling sigmoidal curves. Technol. Forecast. Soc. Chang. 69 (3), 233–241.

Chang, P.C., Fan, C.Y., Liu, C.H., 2009. Integrating a piecewise linear representation method and a neural network model for stock trading points prediction. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 39 (1), 80–92.

Chang, P.L., Wu, C.C., Leu, H.J., 2010. Using patent analyses to monitor the technological trends in an emerging field of technology: a case of carbon nanotube field emission display. Scientometrics 82 (1), 5–19.

Chen, Y.H., Chen, C.Y., Lee, S.C., 2011. Technology forecasting and patent strategy of hydrogen energy and fuel cell technologies. Int. J. Hydrog. Energy 36 (12), 6957–6969.

Chen, H., Zhang, Y., Zhang, G., Zhu, D., Lu, J., 2015a. Modelling technological topic changes in patent claims. 2015 Portland International Conference on Management of Engineering and Technology (PICMET). IEEE, pp. 2049–2059.

Chen, H., Zhang, G., Zhu, D., Lu, J., 2015b. A patent time series processing component for technology intelligence by trend identification functionality. Neural Comput. & Applic. 26 (2), 345–353.

Choi, J., Hwang, Y.S., 2014. Patent keyword network analysis for improving technology development efficiency. Technol. Forecast. Soc. Chang. 83, 170–182.

Cunningham, S.W., Porter, A.L., Newman, N.C., 2006. Special issue on tech mining. Technol. Forecast. Soc. Chang. 73 (8), 915–922.

De Battisti, F., Ferrara, A., Salini, S., 2015. A decade of research in statistics: a topic model approach. Scientometrics 103, 413–433.

Ding, Y., 2011. Topic-based pagerank on author cocitation networks. J. Assoc. Inf. Sci. Technol. 62 (3), 449–466.

Ernst, H., 1997. The use of patent data for technological forecasting: the diffusion of CNC-technology in the machine tool industry. Small Bus. Econ. 9 (4), 361–381.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proc. Natl. Acad. Sci. U. S. A. 101, 5228–5235.

Griliches, Z., 1990. Patent statistics as economic indicators: a survey. J. Econ. Lit. 28 (4), 1661–1707.

Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2001. On clustering validation techniques. J. Intell. Inf. Syst. 17 (2-3), 107–145.

Haywood, S., 2003. Academic Vocabulary. Nottingham University http://www.nottingham.ac.uk/alzsh3/acvocab/wordlists.htm (Accessed January 2015).

Heinrich, G., 2005. Parameter Estimation for Text Analysis. Fraunhofer IGD, Darmstadt, Germany.

Holst, H., Nguyen, H., Wikander, J., 2010. Innovation Driven Research Education: Volume I: An Introduction. Product Innovation Engineering Program, Stockholm, Sweden.

Hyndman, R.J., Athanasopoulos, G., 2014. Optimally reconciling forecasts in a hierarchy. Foresight 35, 42–48.

Jeong, D.H., Song, M., 2014. Time gap analysis by the topic model-based temporal technique. J. Informetr. 8 (3), 776–790.

Keogh, E., Pazzani, M., 1998. An enhanced representation of time series which allowsfast and accurate classification, clustering and relevance feedback. Proceedings of the 4th International Conference of Knowledge Discovery and Data Mining. AAAI Press, pp. 239–241.

Keogh, E., Chu, S., Hart, D., Pazzani, M., 2001. An online algorithm for segmenting time series. Proceedings IEEE International Conference on Data Mining, pp. 289–296.

Keogh, E., Chu, S., Hart, D., Pazzani, M., 2004. Segmenting time series: a survey and novel approach. Data mining in time series databases 57, pp. 1–22.

Kim, J., Hwang, M., Jeong, D.H., Jung, H., 2012. Technology trends analysis and forecasting application based on decision tree and statistical feature analysis. Expert Syst. Appl. 39 (16), 12618–12625.

Kimura, A., Kashino, K., Kurozumi, T., Murase, H., 2008. A quick search method for audio signals based on a piecewise linear representation of feature trajectories. IEEE Trans. Audio Speech Lang. Process. 16 (2), 396–407.

Krampen, G., Eye, A., Schui, G., 2011. Forecasting trends of development of psychology from a bibliometric perspective. Scientometrics 87 (3), 687–694.

Lee, H.J., Lee, S., Yoon, B., 2011. Technology clustering based on evolutionary patterns: the case of information and communications technologies. Technol. Forecast. Soc. Chang. 78 (6), 953–967.

Lee, S., Lee, H.J., Yoon, B., 2012. Modelling and analyzing technology innovation in the energy sector: patent-based HMM approach. Comput. Ind. Eng. 63 (3), 564–577.

Lee, C., Song, B., Park, Y., 2013. How to assess patent infringement risks: a semantic patent claim analysis using dependency relationships. Tech. Anal. Strat. Manag. 25 (1), 23–38.

Lewis, D., Yang, Y., Rose, T., Li, F., 2004. SMART stopword list. Journal of Machine Learning Research. MIT Press http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop Accessed January 2015.

Luo, L., Chen, X., 2013. Integrating piecewise linear representation and weighted support vector machine for stock trading signal prediction. Appl. Soft Comput. 13 (2), 806–816.

Martino, J.P., 1993. Technological Forecasting for Decision Making. McGraw-Hill, Inc.

Modis, T., Debecker, A., 1992. Chaoslike states can be expected before and after logistic growth. Technol. Forecast. Soc. Chang. 41 (2), 111–120.

Noel, G.E., Peterson, G.L., 2014. Applicability of latent dirichlet allocation to multi-disk search. Digit. Investig. 11 (1), 43–56.

Novelli, E., 2015. An examination of the antecedents and implications of patent scope. Res. Policy 44 (2), 493–507.

Philips, F., 1999. A method for detecting a shift in a trend. Portland International Conference on Management of Engineering and Technology 231, p. 238.

Phillips, F., Linstone, H., 2016. Key ideas from a 25-year collaboration at technological forecasting & social change. Technol. Forecast. Soc. Chang. 105, 158–166.

Porter, A.L., Cunningham, S.W., 2004. Tech Mining: Exploiting new Technologies for Competitive Advantage vol. 29. John Wiley & Sons.

Sheldon, J.G., 1995. How to Write a Patent Application. Practising Law Institute.

Shih, M.J., Liu, D.R., Hsu, M.L., 2010. Discovering competitive intelligence by mining changes in patent trends. Expert Syst. Appl. 37 (4), 2882–2890.

Steinbach, M., Karypis, G., Kumar, V., 2000. A comparison of document clustering techniques. KDD Workshop on Text Mining. ACM SIGKDD, Boston.

Steyvers, M., Griffiths, T., 2007. Probabilistic topic models. In: Landauer, D.M.T., Dennis, S., Kintsch, W. (Eds.), Latent Semantic Analysis: A road to Meaning. Laurence Erlbaum.

Suominen, A., Toivanen, H., 2015. Map of science with topic modelling: comparison of unsupervised learning and human-assigned subject classification. J. Assoc. Inf. Sci. Technol. http://dx.doi.org/10.1002/asi.23596.

Suominen, A., Toivanen, H., Seppänen, M., 2016. Firms' knowledge profiles: mapping patent data with unsupervised learning. Technological Forecasting and Social Change http://dx.doi.org/10.1016/j.techfore.2016.09.028.

Tong, X., Frame, J.D., 1994. Measuring national technological performance with patent claims data. Res. Policy 23 (2), 133–141.

Tseng, Y.H., Lin, C.J., Lin, Y.I., 2007. Text mining techniques for patent analysis. Inf. Process. Manag. 43 (5), 1216–1247.

USPTO, 2012. Manual of patent examining procedure: claim interpretation. Patent Laws, Regulations, Policies & Procedures, Chapter 2100, Section 2111.

Venugopalan, S., Rai, V., 2015. Topic based classification and pattern identification in patents. Technol. Forecast. Soc. Chang. 94, 236–250.

Verhoeven, D., Bakker, J., Veugelers, R., 2016. Measuring technological novelty with patent-based indicators. Res. Policy 45 (3), 707–723.

Watts, R.J., Porter, A.L., 2003. R&D cluster quality measures and technology maturity. Technol. Forecast. Soc. Chang. 70 (8), 735–758.

Wikipedia, 2014. Transitional phrase. Wikipedia. http://en.wikipedia.org/wiki/Transitional_phrase (Accessed January 2015).

WIPO, 2002. Patent cooperation treaty (PCT) article 6: claims. Claims. WIPO, Washington.

WIPO, 2004. WIPO Intellectual Property Handbook: Policy, law and use. second ed. pp. 17–40.

Xie, Z., Miyazaki, K., 2013. Evaluating the effectiveness of keyword search strategy for patent identification. World Patent Inf. 35 (1), 20–30.

Yang, S., Soo, V., 2012. Extract conceptual graphs from plain texts in patent claims. Eng. Appl. Artif. Intell. 25 (4), 874–887.

Yang, L., Qiu, M., Gottipati, S., et al., 2013. CQArank: jointly model topics and expertise in community question answering. Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM, pp. 99–108.

Yoon, J., Kim, K., 2012. TrendPerceptor: a property-function based technology intelligence system for identifying technology trends from patents. Expert Syst. Appl. 39 (3), 2927–2938.

Yoon, B., Park, Y., 2005. A systematic approach for identifying technology opportunities: keyword-based morphology analysis. Technol. Forecast. Soc. Chang. 72 (2), 145–160.

Young, P., 1993. Technological growth curves: a competition of forecasting models. Technol. Forecast. Soc. Chang. 44 (4), 375–389.

Zhu, D., Porter, A.L., 2002. Automated extraction and visualization of information for technological intelligence and forecasting. Technol. Forecast. Soc. Chang. 69 (5), 495–506.

Hongshu Chen received her Ph.D. degree in software engineering from the Faculty of Engineering and Information Technology, University of Technology Sydney, Australia, in 2016. Her main research interests include text mining, technology intelligence, especially technological forecasting and topic analysis with the combination of qualitative and quantitative methodologies.

Guangquan Zhang received his Ph.D. degree in applied mathematics from Curtin University of Technology, Perth, Australia, in 2001. He is currently an Associate Professor in the Faculty of Engineering and Information Technology, and the Co-director of the Decision Systems and e-Service Intelligence Research Laboratory, Centre for Quantum Computation and Intelligent Systems, at University of Technology Sydney. His main research interests include multi-objective and group decision making, decision support system tools, fuzzy measure and optimization, and uncertain information processing.

Donghua Zhu is currently a Professor of the School of Management and Economics, and the Director of the Knowledge Management and Data Analysis Laboratory, at Beijing Institute of Technology, China. His main academic research fields include science and technology data mining, technology innovation management, technology forecasting and management. His current research emphasises big data analytics.

Jie Lu received the Ph.D. from Curtin University of Technology, Perth, Australia, in 2000. She is currently a Professor and the Associate Dean Research in the Faculty of Engineering and Information Technology, and the Director of the Decision Systems and e-Service Intelligence Research Laboratory, Centre for Quantum Computation and Intelligent Systems, at University of Technology Sydney, Australia. Her main research interests include decision making modelling, decision support system tools, uncertain information processing, recommender systems, and e-Government and e-Service intelligence.