Chapter 11
Identifying Technological Topic Changes in Patent Claims Using Topic Modeling

Hongshu Chen, Yi Zhang and Donghua Zhu

Abstract Patent claims usually embody the core technological scope and the most essential terms to define the protection of an invention, which makes them the ideal resource for patent topic identification and theme changes analysis. However, conducting content analysis manually on massive technical terms is very time-consuming and laborious. Even with the help of traditional text mining techniques, it is still difficult to model topic changes over time, because single keywords alone are usually too general or ambiguous to represent a concept. Moreover, term frequency that used to rank keywords cannot separate polysemous words that are actually describing a different concept. To address this issue, this research proposes a topic change identification approach based on latent dirichlet allocation, to model and analyze topic changes and topic-based trend with minimal human intervention. After textual data cleaning, underlying semantic topics hidden in large archives of patent claims are revealed automatically. Topics are defined by probability distributions over words instead of terms and their frequency, so that polysemy is allowed. A case study using patents published in the United States Patent and Trademark Office (USPTO) from 2009 to 2013 with Australia as their assignee country is presented, to demonstrate the validity of the proposed topic change identification approach. The experimental result shows that the proposed approach can be used as an automatic tool to provide machine-identified topic changes for more efficient and effective R&D management assistance.

Keywords Tech mining · Topic modeling · Patent analysis

H. Chen (✉) · Y. Zhang
Decision Systems & e-Service Intelligence Research Lab,
Centre for Quantum Computation & Intelligent Systems,
Faculty of Engineering and Information Technology, University of Technology Sydney,
Sydney, Australia
e-mail: Hongshu.Chen@uts.edu.au

H. Chen · Y. Zhang · D. Zhu
School of Management and Economics, Beijing Institute of Technology,
Beijing, People's Republic of China

T.U. Daim et al. (eds.), *Anticipating Future Innovation Pathways Through Large Data Analysis*, Innovation, Technology, and Knowledge Management,
DOI 10.1007/978-3-319-39056-7_11

11.1 Introduction

Patent claims are often argued as a valuable source for the detection of technological changes and to gain technological insight (Campbell 1983; Ernst 1997; WIPO 2004). As an important part of unstructured segments of a patent document, claims hold explicit information and implicit knowledge revealing technological concepts, topics, and related R&D activities with concise, but precise language (Xie and Miyazaki 2013; WIPO 2002). Since manually conducting content analysis on massive patent documents is very time-consuming and laborious, in recent years, one of the fundamental changes to research in R&D management is the access to extremely powerful information techniques and a vast amount of digital and textual data (Daim et al. 2011). In particular, for efficient patent analysis, automatic approaches to assist domain experts and decision makers to discover and understand large volumes of patent documents have drawn increasing attention and still are in great demand (Abbas et al. 2014).

Much effort has been devoted to reveal latent knowledge from the textual data of patent documents. Watts and Porter (1997) suggested an approach to investigate terminological trends by tracking the historical change of keywords. Yoon and Park (2005) presented a keyword-based morphology study to identify the detailed configurations of promising technology. Zhang et al. (2014) introduced a term clumping approach based on principal components analysis to explore keywords and main phrases in abstract from scientific literature. In addition, text analytics have already been applied to technology intelligence application *TrendPerceptor* (Yoon and Kim 2012), *Techpioneer* (Yoon 2008), *VantagePoint* (Zhu and Porter 2002), and *Aureka* (Trippe 2003) to determine hidden concepts and relationships, where clustering, classification and mapping techniques were used to support further content analysis of technological documents. However, before most of these applications are applied, usually several sets of keywords need to be defined in advance, which still derive from the opinion and knowledge of domain experts. Moreover, the outcomes of majority traditional text mining techniques are based on single keywords with ranking, yet these words alone are usually too general or misleading for indicating a concept, especially when there are polysemous words actually describing different themes (Tseng et al. 2007).

To overcome the above-mentioned limitations, this research proposes a topic change identification approach using a well-known topic modeling approach, latent dirichlet allocation. Unsupervised topic modeling is applied to vast amounts of target patent claims, providing a corpus structure with minimal human intervention. There is no preset classification or keywords list for this approach, and the results are discovered in a completely unsupervised way. In addition, instead of using single terms, topics are represented by probability distributions over words. The actual semantic meaning of a topic is able to be delivered in this way, and at the same time, the polysemous words, which are actually depicting different concepts, can also be separated. After revealing topics from patent sub-collections of different years, a topic change model is presented to identify topic changes over time.

Finally, to demonstrate the performance of our proposed approach, patents published during years 2009 to year 2013 in the United States Patent and Trademark Office (USPTO) with Australia as their assignee country are selected to present a case study. The experimental result demonstrates that the proposed approach is able to provide machine-identified topic changes automatically without any presetting of keywords. The outcomes of our approach will be used to serve R&D management assistance.

This paper is organized as follows: the first section reviews related research developments by introducing patent data in technological research and latent dirichlet allocation. Methodology Section describes the proposed topic change identification approach step by step. Case Study Section carries out experiments using USPTO patents to demonstrate the proposed approach in a real patent analysis context. The conclusions and future study are addressed in the last section.

11.2 Literature Review

11.2.1 Patent Data in Tech Mining

Patent documents are composed of structured information and unstructured descriptions of inventions. Analytical approaches based on structured data of patents, such as issue date, inventor, assignees, or International Patent Classification, have played the major role in both theoretical and practical research to gain insight of technology development in certain area (Lai and Wu 2005; Sheikh et al. 2011; Nishijima et al. 2013). However, the unstructured data in patent documents, such as abstracts, claims, and descriptions, usually contain much more abundant information than the structured sections, since they contain significant characteristics, detailed functionalities, or major contributions of technologies. Therefore, there has been a lot of interest in applying text mining techniques to conduct tech mining and set domain analysts free from studying and understanding massive amounts of technological content since the last decade (Tseng et al. 2007; Camus and Brancaleon 2003; Porter 2005).

Among all the unstructured segments of a patent file, patent claims play a role of embodying all the important technical features of an invention with the most essential technological terms to define the protection (Tong and Frame 1994). On one hand, they reveal the core inventive topics and the major technological scope of a patent; on the other hand, claims are written in concise, but precise language, which make them the best resource for identifying technological topics and facilitating patent document analysis (Xie and Miyazaki 2013; WIPO 2002; Yang and Soo 2012; Novelli 2014).

A patent claim usually consists of three parts: a preamble that serves as an introductory section to recite the primary purpose, function, or properties; a transition phrase, such as comprising, having including, consisting of, etc.; a “body”

that contains the elements or steps that together describe the invention (Yang and Soo 2012; USPTO 2012; Sheldon 1995). This research utilizes patent claims as the main source of topic change analysis. Among patent databases from different countries, the United States Patent and Trademark Office (USPTO) database is mostly used because patents submitted in other countries are often also simultaneously submitted in the United States (USPTO 2015).

11.2.2 Latent Dirichlet Allocation

Latent dirichlet allocation (LDA) (Blei et al. 2003) is a probabilistic model that aims to estimate the properties of multinomial observations by unsupervised learning. It gives an estimation of the latent semantic topics hidden in large archives of documents and calculates the probabilities of how various documents belong to different topics. LDA has been used as an efficient tool to assist topic discovery and analysis, in practice. For example, Griffiths and Steyvers (2004) applied LDA-based topic modeling to discover the hot topics covered by papers in Proceedings of the National Academy of Sciences of the United States of America (PNAS); Yang et al. (2013) proposed a topic expertise model (TEM) based on LDA to jointly model topics and expertise for community question answering (CQA) with stack overflow data; Kim and Oh (2011) proposed a framework based on LDA to identify important topics and their meaningful structure within the news archives on the Web.

The graphical model of LDA is presented in Fig. 11.1, showing three rectangular plates where: D denotes the overall documents in a corpus; K indicates the topic numbers for D; and N_d stands for the term number of dth document in document collection D. Each node in the figure stands for a random variable in the generative process of LDA, while the plates indicate replication. In the left part of the figure, $\vec{\vartheta}_d$ stands for the topic proportions for the dth document. For document

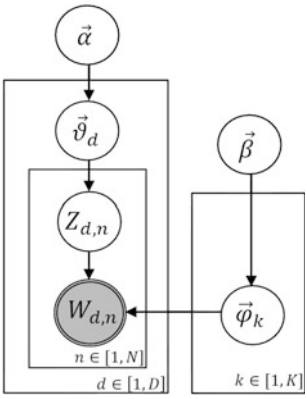

Fig. 11.1 The graphical model of latent dirichlet allocation

d, the topic assignments are Z_d, where $Z_{d,n}$ indicates the topic assignment of the nth word in the dth document. On the right of the figure, the topics themselves are illustrated by $\vec{\varphi}_{1:K}$, where each $\vec{\varphi}_k$ is a distribution over vocabularies. All of the unshaded circles indicate hidden nodes. The shaded circles, on the contrary, are observable nodes, where $W_{d,n}$ stands for the nth word in document d. Finally, α and β are two hyperparameters that determine the amount of smoothing applied to the topic distributions for each document and the word distributions for each topic (Blei et al. 2003; Steyvers and Griffiths 2007; Blei 2012; Heinrich 2005).

The parameters of LDA need to be estimated by an iterative approach. Among existing approaches, Gibbs sampling is one of the most commonly used methods. It is an approximate inference algorithm based on the Markov chain Monte Carlo (MCMC) and has been widely used to estimate the assignment of words to topics by observed data (Griffiths and Steyvers 2004; Noel and Peterson 2014; Lukins et al. 2010). Gibbs sampling produces different results each time in executing LDA, so that the topic estimations are slightly different even with exactly the same setting of input and parameters; yet on the whole, the results of different experiments will not change much.

11.3 Methodology

This section explains the details of our proposed topic change identification approach. The framework is given first; each detailed step is illustrated subsequently.

11.3.1 Framework

The overall framework of our proposed topic change identification approach is shown in Fig. 11.2. First of all, users need to initiate a search statement to declare their domain analytic requirements and address a group of target patents in USPTO database. Patent ID, title, claims, issue time, assignees, United States Patent Classification (USPC), and other information of target patents are then crawled into a database waiting for further analysis. To identify topic changes over time, the whole patent collection is divided into several sub-collections first and labeled with their corresponding issue year. Subsequently, for each sub-collection, patent claims and titles, embodying essential technical terms, and USPC, providing a general understanding of the domain classification, are extracted from the target patents database separately. The two plates in the figure indicate replication.

Textual data composed by claims and titles, after data segmentation and cleaning, are then placed into a series of words exclusion modules to filter out the most common function words, high-frequency words that commonly appeared in

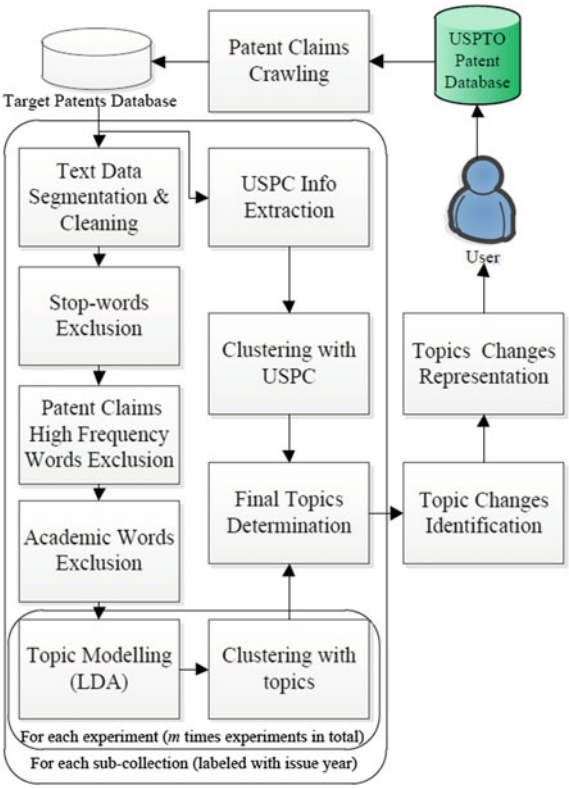

Fig. 11.2 The framework of the proposed topic change identification approach

patent claims, and academic words with vague and general meanings. Then, the prepared text will be passed to the topic modeling module. Meanwhile, the USPC information of the corresponding patents is extracted to assist final topic determination. As mentioned, the randomness introduced by the initiation of the sampling will affect the final result of LDA. To acquire the most reliable topics of the corpus, we utilize USPC as a measurement to evaluate results from *m* times experiments. Patents are clustered with both their USPC and topic proportions. The final topic modeling result is the one trial that provides the most similar clusters to the USPC clustering outcome. Finally, with all the topics estimated from patent sub-collections of different years, topic changes over time can be identified and presented to users.

11.3.2 Patent Corpus Text Cleaning

Patent claims are a special kind of textual data that contain plenty of technical terms, specific words serving as transition phrases, and numerous academic words

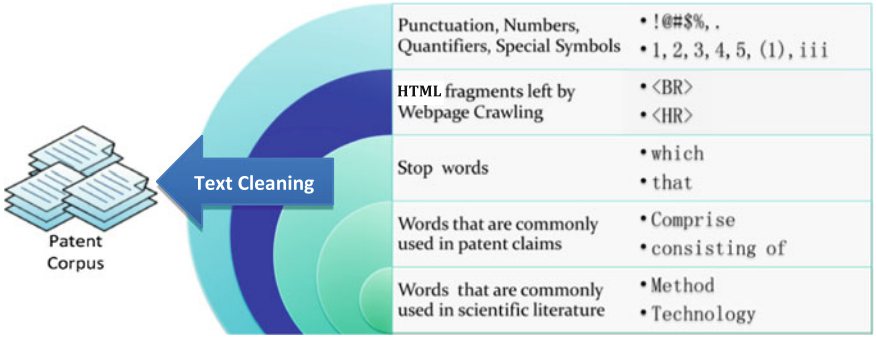

Fig. 11.3 Relationships between sub-collections and topics

that describe invention outcomes. Among all the terms that one claim may contain, only technical terms provide the most meaningful information that reflects technological topics and innovations. Therefore, for our patent corpus, each sub-collection, as shown in Fig. 11.3, before modeling topics with LDA, except all the punctuations, numbers, and HTML fragments left by webpage crawling, we also utilize three modules to remove general words from the corpus of patents as follows:

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

The stop words list we applied is from an information retrieval Resources link from Stanford University (David 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; Zhang et al. 2014).

11.3.3 Topic Modeling

LDA utilizes a probability distribution over words, instead of a single term, to define a concept, delivering better semantic meaning of the topic and, at the same time, allowing polysemy. Thus, it is very suitable for "understanding" the content of large corpuses such as emails, news, scientific papers, and our main data source here, patent claims. After removing all commonly used words from the corpus, we utilize LDA to generate several groups of topics for each patent sub-collection in the corpus, which is labeled by its corresponding issue year. In a sub-collection, the claims and title of each patent constitute one document, and the number of

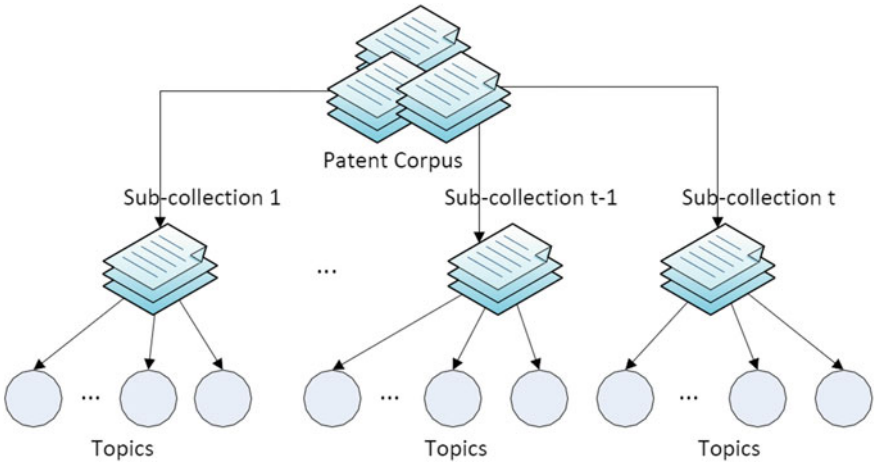

Fig. 11.4 Relationships between sub-collections and topics

documents equals the number of patents; the USPC and other structural information are stored alone in a single file to assist further topic determination. All the textual documents in the corpus are seen as mixtures of a number of topics; each topic is seen as a distribution over various vocabularies. Here, we present the global topics as $\vec{P}_{1:t} = (\vec{P}_1, \vec{P}_2, \ldots, \vec{P}_i, \ldots, \vec{P}_t)$, where $\vec{P}_i$ stand for the topics of the ith sub-collection of the corpus. The relationship between sub-collections and topics is illustrated in Fig. 11.4.

Since we know nothing about the word distributions composing the topics and the topic distributions composing the documents, before topic modeling, assumptions need to be first drawn to determine the parameters k, α, β of LDA. According to previous research, hyperparameters α, β of the dirichlet distribution in LDA have a smoothing effect on multinomial parameters; that is, the lower the values of α and β are, the more decisive topic associations there will be (Heinrich 2005). This research sets $\alpha = 0.5$ and $\beta = 0.1$, which are commonly used in LDA applications (Koltcov et al. 2014). For the setting of K, higher K will reduce the topical granularity but increase the processing time significantly. Therefore, during the implementation, K needs to be decided case by case, balancing user requirement and time consumption. Different parameter settings may improve modeling performance, yet optimizing these parameters is beyond the scope of this paper.

11.3.4 Final Topics Determination

We then apply Gibbs sampling to infer the needed distributions in LDA. Since the initial values of variables are determined randomly in Gibbs sampling, the outputs of LDA in multiple experiments with a same corpus are slightly different. To ensure

the final topic modeling estimation as reliable as possible, evaluation criteria will be needed for the topics finalization. In this research, we select USPC as the criteria. As a predefined classification hierarchy built on domain expert judgments, USPC provides a general understanding of the technical domain of concern to one patent. Because patents covering similar topics are usually assigned to a same main USPC, thus here we use the main USPC to judge which estimation is closer to the actual topic structure.

For a sub-collection of corpus, multiple LDA experiments will produce a number of topic distribution matrixes, each indicating the topic distribution proportions of patent documents in the corresponding trial. As shown in the approach framework, Fig. 11.2, there will be m times experiments for every sub-collection; and after performing each time run, patents in the sub-collection are clustered with their calculated topic distributions using the hierarchical clustering approach (Steinbach et al. 2000). Meanwhile, the same group of patents will be also clustered with USPC information. The closer the two clustering results are, the more reliable the topic modeling result is.

Specifically, the values of indexes Jaccard et al. and F1 of m times experiments are used to measure the similarity of the two clustering results, one by topics and the other by USPC. The three indices are listed as follows (Halkidi et al. 2001):

$$J = a/(a+b+c), \tag{11.1}$$

$$\mathrm{FM} = a/\sqrt{r_1 \cdot r_2}, \tag{11.2}$$

$$F_\beta = \frac{\left(\beta^2+1\right)\cdot r_1 \cdot r_2}{\beta^2 \cdot r_1 + r_2}, \tag{11.3}$$

where J stands for Jaccard coefficient, FM indicates Folkes & Mallows index, F_β presents the $F1$ indice. In addition, $r_1 = a/(a+b)$, $r_2 = a/(a+c)$, where a represents the number of patents that belong to the same cluster of topics and to the same USPC in our case, b is the number of patents that belong to the same cluster of topics but to different USPC, and c is the number of patents that belong to different clusters of topics but to the same USPC. The topic modeling result that provides the highest index values is the optimal one.

11.3.5 Topic Change Identification

After locating the final topics and words underlying the sub-collections of our corpus, we are able to identify the topic change over time. As show in Fig. 11.5, we compare two groups of topics deriving from different corpus sub-collections, calculating the similarity of words between each topic in $\vec{P}_i$ and all the topics in $\vec{P}_{i-1}$, in a traversal way. If two topics under different sub-collections contain

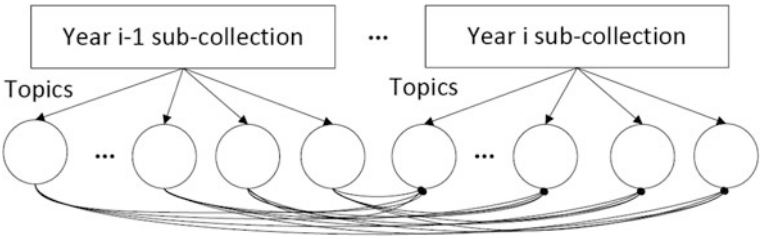

Fig. 11.5 Topic change identification model

approximately the same group of words, then we believe that these two topics are actually one topic evolving from year to year. However, if the majority of words comprising two topics are very different, then we believe these are two different topics. Finally, for documents sub-collection of year i, if there is no similar topic can be matched in the previous year, year $i-1$, then the unmatched topic in the later year can be seen as a newly important one, which means it became more hot in the year i.

11.3.6 Topic-Based Trend Estimation

If we already identified a topic evolving from year to year, besides discovering how the detailed content of the topic evolves from year to year with the above model, we can also use the topic distribution matrix to generate historical topic-based trend and forecast future trend. As an important part of LDA outcomes, the topic distribution matrix $\vec{\vartheta}$ provides the estimated result that how all the topics distribute over the document collection. The summation of each row of the matrix equals 1. The sum values of each column, however, are different. The larger the sum of a column, the more important the corresponding topic is. Since the patents are issued along a time line, if we add up a group of elements in a column that associates with patents published in a same time interval (month or year), the summation can be used to present the weight of the topic in that time frame. Thus we can then get a temporal-weight matrix to reveal the importance of selected topics in different month or years.

After the temporal-weight matrix is achieved, we calculate the weight changes in a least-squares sense to estimate the general trend of the target topics. The temporal-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 time. We utilize the coefficients a and b to measure developing trends of topics, since a controls the speed of increase (or decrease) of the quadratic function, $-b/2a$ control the axis of symmetry. For instance, if coefficient a is positive and the symmetry is on the left of y-axis, we consider the corresponding topic has a growing trend where the greater a is, the faster the growth will be.

11.4 Case Study

11.4.1 Data Collection

To demonstrate the performance of our proposed approach, patents published during years 2009 to year 2013 in USPTO (http://www.uspto.gov/) with Australia as their assignee country are selected to present a case study. There are 7071 target patents covering 343 different main USPC[1,2]. Their patent ID, titles, issue time, inventors, Assignees, United States Patent Classification (USPC), International Patent Classification (IPC), and most importantly, their claims are clawed from USPTO and placed in a patents tool for further processing. The claims and title for each patent constitute one document in our corpus, which totals 7071 documents on the whole. Then, the whole document collection was divided into five sub-collections to present technological feature and essential terms of inventions by Australia assignees in the past five years. The detailed documents number was published every year from 2009 to 2010; the term number and USPC number in each corresponding sub-collection are shown in Table 11.1. Although the number of documents declined from year 2011, the term number kept rising, which implies that the average complexity of patent claims description is increasing in the resent three years. We also observe that the number of USPC in 2010 had a visible growth, suggesting that there may be a group of new topics appearing in year 2010 comparing with year 2009.

11.4.2 Topic Set Determination

Before topic modeling, as mentioned, a number of parameters need to be set first, including the number of topics K and α, β of dirichlet distribution. In the case study, we applied $K = 10$ with model hyperparameters $\alpha = 0.5, \beta = 0.1$ to our target documents, to balance the topical granularity, convenience of understanding, and

Table 11.1 The number of documents, terms, and USPC of patents published each year

Year	Doc No.	Term No.	USPC No.
2009	1174	19,796	199
2010	1613	24,726	233
2011	1746	23,757	228
2012	1256	25,102	233
2013	1282	29,714	227

[1]Data accessed in March 2014.

[2]All plant patents are seen as having one same USPC for calculation convenience.

Table 11.2 Indexes information for the final chosen experiment result

Year	Index	E 1	E 2	E 3	E 4	E 5
2009	FM	0.2376	0.2803	**0.2845**	0.2739	0.1948
2009	DJC	0.1217	0.1500	**0.1505**	0.1436	0.0962
2009	F1	0.2169	0.2608	**0.2616**	0.2511	0.1755
2010	FM	0.2668	0.2152	0.2253	0.3125	**0.3688**
2010	DJC	0.1357	0.1037	0.1077	0.1634	**0.2017**
2010	F1	0.2389	0.1880	0.1944	0.2809	**0.3356**
2011	FM	0.2521	0.2484	0.2334	**0.2604**	0.2541
2011	DJC	0.1334	0.1300	0.1166	**0.1342**	0.1294
2011	F1	0.2354	0.2301	0.2089	**0.2366**	0.2292
2012	FM	0.3060	**0.3202**	0.2773	0.2820	0.2686
2012	DJC	0.1756	**0.1853**	0.1539	0.1632	0.1521
2012	F1	0.2987	**0.3127**	0.2667	0.2806	0.2640
2013	FM	0.2984	0.2989	**0.3356**	0.3177	0.3086
2013	DJC	0.1753	0.1749	**0.1986**	0.1876	0.1794
2013	F1	0.2983	0.2977	**0.3313**	0.3159	0.3042

the speed of processing. There are 10 topics describing the essential technological content and feature for each year; and every topic is presented with 10 words given highest probability by this topic.

Indices Folkes & Mallows (FM), Jaccard (DJC), and F1 are calculated after we clustered the patents using both topic assignment and main USPC information. Observation for each year was performed 5 ($m = 5$) runs with 2000 iterations of Gibbs sampling. The detailed index values of five times experiments are listed in Table 11.2, where we can observe directly that the 3rd experiment (E3) of documents sub-collection in 2009, the 5th experiment of documents sub-collection in 2010 (E5), the 4th experiment of documents sub-collection in 2011 (E4), the 2nd experiment (E2) of documents sub-collection in 2012, and the 3rd experiment (E3) of documents sub-collection in 2013 have the largest value of all three indexes among all experimental trials. We believe that these models can fit the observation better and the topics and parameters provided by the five trials are our final topic modeling result.

Since there is no preset classification or domain knowledge assistance needed, the topic modeling results are discovered in an unsupervised way. In the past five years, patents owned by Australia assignees cover several important technological topics, such as print head and nozzle, alkyl compound, pressure apparatus, and antibody sequence. The more the topic words are taken into consideration to describe a topic, the more clear and specific the topical semantic meaning will be. Specifically, the topics for each year are presented as follows. The order of the topics is random, and the numbers behind words are the probability values of corresponding topic words. Details of all the topics, the top 10 ranked words and their corresponding probabilities, are shown in Table 11.3 in the Appendix.

Table 11.3 The top 10 ranked words of all the topics from years 2009 to 2013 and their corresponding probabilities

Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Year 2009									
Topic 1		*Topic 2*		*Topic 3*		*Topic 4*		*Topic 5*	
Printhead	0.0418	Device	0.0244	Ink	0.0442	Step	0.0116	Portion	0.0246
Ink	0.0353	Image	0.0217	Ejection	0.0336	Composition	0.0095	Body	0.0150
Print	0.0333	Coded	0.0209	Nozzle	0.0334	Gas	0.0088	Assembly	0.0132
Printer	0.0252	System	0.0195	Inkjet	0.0307	Leach	0.0081	Surface	0.0110
Media	0.0229	Sensing	0.0181	Printhead	0.0245	Material	0.0065	Extending	0.0092
Cartridge	0.0138	Digital	0.0132	Drop	0.0229	Acid	0.0064	Wall	0.0091
Module	0.0137	Computer	0.0105	Apparatus	0.0224	Fuel	0.0063	Mask	0.0081
Printing	0.0135	Camera	0.0101	Actuator	0.0220	Water	0.0059	Adapted	0.0076
Assembly	0.0132	Identity	0.0092	Element	0.0191	Polymer	0.0058	Substantially	0.0072
Configured	0.0124	Position	0.0086	Chamber	0.0189	Ph	0.0055	Support	0.0069
Topic 6		*Topic 7*		*Topic 8*		*Topic 9*		*Topic 10*	
Support	0.0152	Compound	0.0183	System	0.0116	Signal	0.0278	Antibody	0.0379
Roller	0.0142	Formula	0.0111	Material	0.0090	Sensor	0.0108	Fragment	0.0246
Device	0.0122	Alkyl	0.0109	Game	0.0088	Signals	0.0107	Sequence	0.0220
Drive	0.0109	Independently	0.0102	Plurality	0.0087	Frequency	0.0089	Human	0.0219
Assembly	0.0101	Layer	0.0098	Computer	0.0079	Device	0.0087	Acid	0.0177
Mechanism	0.0082	Optionally	0.0095	Gaming	0.0073	Input	0.0084	Peptide	0.0175
Surface	0.0080	Base	0.0088	Entry	0.0072	Output	0.0081	Mature	0.0164
Frame	0.0075	Detector	0.0087	Torque	0.0063	Apparatus	0.0081	Cell	0.0157
Position	0.0071	Substituted	0.0087	Object	0.0058	Processing	0.0071	Binding	0.0138
Mounted	0.0067	Reflector	0.0087	Service	0.0054	Power	0.0067	Amino	0.0133

(continued)

Table 11.3 (continued)

Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Year 2010									
Topic 1		*Topic 2*		*Topic 3*		*Topic 4*		*Topic 5*	
Portion	0.0217	Signal	0.0240	Ink	0.0518	**Material**	**0.0144**	**Memory**	**0.0253**
Surface	0.0126	Light	0.0131	Printhead	0.0476	**Step**	**0.0136**	**Computer**	**0.0191**
Outer	0.0095	System	0.0121	Nozzle	0.0214	**Water**	**0.0101**	**Plurality**	**0.0161**
Assembly	0.0090	Optical	0.0104	Inkjet	0.0183	**Layer**	**0.0101**	**Network**	**0.0155**
Body	0.0088	Device	0.0104	Print	0.0176	**Metal**	**0.0088**	**Single**	**0.0143**
Extending	0.0086	Image	0.0083	Assembly	0.0172	**Polymer**	**0.0081**	**Application**	**0.0141**
Wall	0.0080	Power	0.0076	Printer	0.0156	**Form**	**0.0070**	**Program**	**0.0133**
Support	0.0076	Frequency	0.0076	Media	0.0127	**Defined**	**0.0067**	**System**	**0.0117**
Upper	0.0073	Output	0.0069	Ejection	0.0126	**Composition**	**0.0066**	**Local**	**0.0103**
Frame	0.0071	Sensor	0.0067	Configured	0.0110	**Concentration**	**0.0063**	**Computers**	**0.0097**
Topic 6		*Topic 7*		*Topic 8*		*Topic 9*		*Topic 10*	
Device	0.0269	Acid	0.0199	Apparatus	0.037	**Compound**	**0.0184**	**System**	**0.0175**
Coded	0.0252	Sequence	0.0172	Air	0.0214	**Substituted**	**0.0183**	**Device**	**0.0154**
System	0.0245	Plant	0.0159	Pressure	0.0164	**Independently**	**0.0140**	**Electrode**	**0.0146**
Print	0.0190	Nucleic	0.0152	Fluid	0.0148	**Alkyl**	**0.0096**	**Apparatus**	**0.0107**
Computer	0.0168	Seq	0.0146	Valve	0.0144	**Formula**	**0.0094**	**Signal**	**0.0105**
Sensing	0.0161	Cell	0.0136	Flow	0.0140	**Optionally**	**0.0092**	**Configured**	**0.0095**
User	0.0149	Antibody	0.0117	Chamber	0.0131	**Aryl**	**0.0065**	**Euphorbia**	**0.0095**
Media	0.0115	Fragment	0.0088	System	0.0129	**Moiety**	**0.0051**	**Array**	**0.0079**
Mobile	0.0109	Binding	0.0086	Inlet	0.0083	**Composition**	**0.0049**	**Patient**	**0.0074**
Indicative	0.0101	Polypeptide	0.0086	Outlet	0.0071	**Hydrogen**	**0.0046**	**Processing**	**0.0071**

(continued)

Table 11.3 (continued)

Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Year 2011									
Topic 1		*Topic 2*		*Topic 3*		*Topic 4*		*Topic 5*	
Material	0.0188	Portion	0.0260	Ink	0.0579	Sequence	0.0234	Optionally	0.0228
Layer	0.0166	Assembly	0.0202	Printhead	0.0457	Acid	0.0201	Substituted	0.0224
Step	0.0130	Mask	0.0113	Nozzle	0.0282	Seq	0.0179	Compound	0.0159
Composition	0.0083	Support	0.0110	Inkjet	0.0170	Amino	0.0138	Alkyl	0.0142
Range	0.0070	Frame	0.0105	Assembly	0.0163	Cell	0.0130	Lens	0.0102
Polymer	0.0064	Surface	0.0095	Chamber	0.0118	Plant	0.0120	Independently	0.0089
Coating	0.0060	Outer	0.0087	Integrated	0.0116	Gene	0.0113	Optical	0.0079
Metal	0.0058	Wall	0.0084	Printer	0.0113	Fragment	0.0096	Aryl	0.0074
Solution	0.0057	Extending	0.0071	Fluid	0.0107	Cells	0.0085	Zone	0.0070
Forming	0.0056	Body	0.0069	Plurality	0.0103	Isolated	0.0084	Lower	0.0067
Topic 6		*Topic 7*		*Topic 8*		*Topic 9*		*Topic 10*	
Apparatus	0.0226	Signal	0.0203	Print	0.0449	System	0.0289	**System**	**0.0108**
Flow	0.0191	Light	0.0133	Media	0.0296	Coded	0.0211	**Step**	**0.0099**
Air	0.0180	Power	0.0120	Printer	0.0177	Device	0.0207	**Apparatus**	**0.0096**
Gas	0.0180	Device	0.0114	Image	0.0170	Computer	0.0186	**Plurality**	**0.0084**
Water	0.0178	Wireless	0.0103	Controller	0.0148	Memory	0.0140	**Pressure**	**0.0078**
Pressure	0.0161	Apparatus	0.0090	Module	0.0141	Sensing	0.0130	**Determining**	**0.0076**
Valve	0.0158	Source	0.0090	Game	0.0131	Plurality	0.0114	**Processing**	**0.0066**
Device	0.0129	Plurality	0.0078	Gaming	0.0129	Identity	0.0109	**Monitoring**	**0.0058**
Fluid	0.0124	Electrical	0.0078	Configured	0.0127	Indicative	0.0101	**Time**	**0.0057**
Humidifier	0.0110	Optical	0.0074	Printing	0.0120	Position	0.0086	**Determined**	**0.0055**

(continued)

Table 11.3 (continued)

Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Year 2012									
Topic 1		*Topic 2*		*Topic 3*		*Topic 4*		*Topic 5*	
Signal	0.0325	Fluid	0.0209	Portion	0.024	**Gaming**	**0.0513**	**Light**	**0.0145**
Configured	0.0165	Gas	0.0172	Assembly	0.0213	**Game**	**0.0504**	**Plurality**	**0.0114**
Frequency	0.0132	Flow	0.0151	Support	0.0126	**System**	**0.0205**	**System**	**0.0107**
Optical	0.0116	Chamber	0.0145	Mask	0.0106	**Symbols**	**0.0190**	**Site**	**0.0075**
Sound	0.0116	System	0.0132	System	0.0087	**Symbol**	**0.0186**	**Pattern**	**0.0070**
System	0.0103	Valve	0.0129	Element	0.0080	**Plurality**	**0.0185**	**Registration**	**0.0070**
Power	0.0092	Water	0.0121	Nasal	0.0073	**Controller**	**0.0172**	**Respective**	**0.0068**
Control	0.0090	Inlet	0.0099	Adapted	0.0072	**Machine**	**0.0166**	**Lens**	**0.0067**
Electrical	0.0088	Pressure	0.0097	Frame	0.0071	**Player**	**0.0157**	**Symbol**	**0.0063**
Device	0.0087	Liquid	0.0078	Extending	0.0066	**Jackpot**	**0.0127**	**Image**	**0.0063**
Topic 6		*Topic 7*		*Topic 8*		*Topic 9*		*Topic 10*	
Time	0.0112	Material	0.0196	Portion	0.0164	System	0.0202	Substituted	0.0204
Determining	0.0107	Layer	0.0119	Apparatus	0.0101	Computer	0.0202	Optionally	0.0190
Signal	0.0104	Polymer	0.0100	Surface	0.0101	Memory	0.0150	Sequence	0.0162
Test	0.0093	Metal	0.0093	Device	0.0098	Device	0.0139	Compound	0.0157
Sensor	0.0093	Surface	0.0092	Body	0.0088	User	0.0128	Acid	0.0151
Flow	0.0089	Electrically	0.0074	Upper	0.0088	Plurality	0.0081	Seq	0.0095
Waveform	0.0085	Step	0.0067	Extending	0.0087	Coded	0.0078	Nucleic	0.0084
Pressure	0.0085	Conductive	0.0064	Lower	0.0081	Content	0.0078	Composition	0.0079
Predetermined	0.0070	Cell	0.0057	Container	0.0081	Printed	0.0071	Amino	0.0072
Plant	0.0068	Component	0.0056	Assembly	0.0073	Image	0.0069	Antibody	0.0069

(continued)

Table 11.3 (continued)

Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
Year 2013									
Topic 1		*Topic 2*		*Topic 3*		*Topic 4*		*Topic 5*	
Portion	0.0200	Game	0.0555	Signal	0.0206	Cushion	0.0345	Composition	0.0234
Assembly	0.0122	Gaming	0.0451	Configured	0.0181	Mask	0.0287	Seq	0.0184
Body	0.0107	Symbol	0.0322	Apparatus	0.0145	Portion	0.0285	Acid	0.0167
Surface	0.0091	Plurality	0.0274	Device	0.0139	Assembly	0.0191	Sequence	0.0158
Extending	0.0079	Symbols	0.0238	Stimulation	0.0105	Frame	0.0186	Amino	0.0102
Wall	0.0073	Controller	0.0226	Signals	0.0097	Support	0.0168	Antibody	0.0091
Housing	0.0072	Player	0.0189	System	0.0096	Structure	0.0154	Cell	0.0076
Position	0.0070	System	0.0177	Power	0.0096	Full-face	0.0124	Nucleic	0.0071
Relative	0.0063	Arranged	0.0152	Flow	0.0091	Nasal	0.0122	Polypeptide	0.0068
Outer	0.0062	Machine	0.0128	Electrical	0.0086	Underlying	0.0121	Binding	0.0066
Topic 6		*Topic 7*		*Topic 8*		*Topic 9*		*Topic 10*	
Device	0.0286	Material	0.0135	Image	0.0236	**System**	**0.0272**	Substituted	0.0583
Wireless	0.0132	Layer	0.0120	Oligonucleotide	0.0120	**Computer**	**0.0260**	Optionally	0.0513
System	0.0115	Fluid	0.0102	Lens	0.0098	**User**	**0.0154**	Compound	0.0160
Plurality	0.0112	Gas	0.0094	Optical	0.0095	**Program**	**0.0112**	Alkyl	0.0132
Sensor	0.0109	Flow	0.0084	Antisense	0.0086	**Message**	**0.0103**	Independently	0.0129
Signal	0.0092	Water	0.0083	Light	0.0085	**Access**	**0.0088**	Formula	0.0084
Processing	0.0088	Liquid	0.0081	Plurality	0.0077	**Vehicle**	**0.0071**	Alkenyl	0.0084
Control	0.0088	Surface	0.0075	System	0.007	**Code**	**0.0061**	Salt	0.0076
Devices	0.0087	Step	0.0067	Laser	0.0063	**Storage**	**0.0060**	Alkynyl	0.0066
Component	0.0082	Electrode	0.0066	Step	0.0062	**Device**	**0.0059**	Acceptable	0.0065

- The topics of year 2009 include printhead (0.0418) cartridge (0.0353), image (0.0217) device (0.0244), ink (0.0442) nozzle (0.0334), composition (0.0095) material (0.0065), portion (0.0246) assembly (0.0132), roller (0.0142) device (0.0122), alkyl (0.0109) compound (0.0183) formula (0.0111), computer (0.0079) gaming (0.0088), signal (0.0278) sensor (0.0108), and antibody (0.0379) sequence (0.0220).
- The topics of year 2010 contain portion (0.0217) assembly (0.0090), light (0.0131)/optical (0.0104) device (0.0104), ink (0.0518) printhead (0.0476), layer (0.0101) material (0.0144), computer (0.0191) memory (0.0253) plurality (0.0161), coded (0.0252) device (0.0269), antibody (0.0117) sequence (0.0172), pressure (0.0164) apparatus (0.0370), alkyl (0.0096) compound (0.0184), and electrode (0.0146) system (0.0175).
- The topics of year 2011 include layer (0.0166) material (0.0188), portion (0.0260) assembly (0.0202), ink (0.0579) printhead (0.0457), acid (0.0201) sequence (0.0234), alkyl (0.0142) compound (0.0159), pressure (0.0161) apparatus (0.0226), light (0.0133) device (0.0114), image (0.0170) print (0.0449), coded (0.0211) device (0.0207), and plurality (0.0084) apparatus (0.0096).
- The topics of year 2012 cover configured (0.0165) signal (0.0325), fluid (0.0209) chamber (0.0145), portion (0.0240) assembly (0.0213), gaming (0.0513) system (0.0205), light (0.0145) lens (0.0067), signal (0.0104) sensor (0.0093), layer (0.0119) material (0.0196), portion (0.0164) apparatus (0.0101), computer (0.0202) memory (0.0150), and acid (0.0151) sequence (0.0162).
- The topics of year 2013 comprise portion (0.0200) assembly (0.0122), gaming (0.0451) controller (0.0226), configured (0.0181) signal (0.0206), cushion (0.0345) mask (0.0287), acid (0.0167) sequence (0.0158), wireless (0.0132) signal (0.0092) sensor (0.0109), layer (0.0120) material (0.0135), optical (0.0095) lens (0.0098), message (0.0103) system (0.0272), and alkyl (0.0132) compound (0.0160).

11.4.3 Topic Change Identification

After discovering main topics underlying in patent claims of each year's document collection, we then use the topic change model to identify the topic variation from years 2009 to 2013. For different groups of topics associated with two consecutive years, we conduct traversal comparison between the topics that belong to the later year with the topics related to the previous year. Topics that contain very similar words are considered as the same topic experiencing innovation; while topics that cannot match any existing ones count as new topics. Figure 11.6 illustrates the important topics that arose each year after 2009, by presenting the top 10 words for each topic using Pajek (Batagelj and Mrvar 2004).

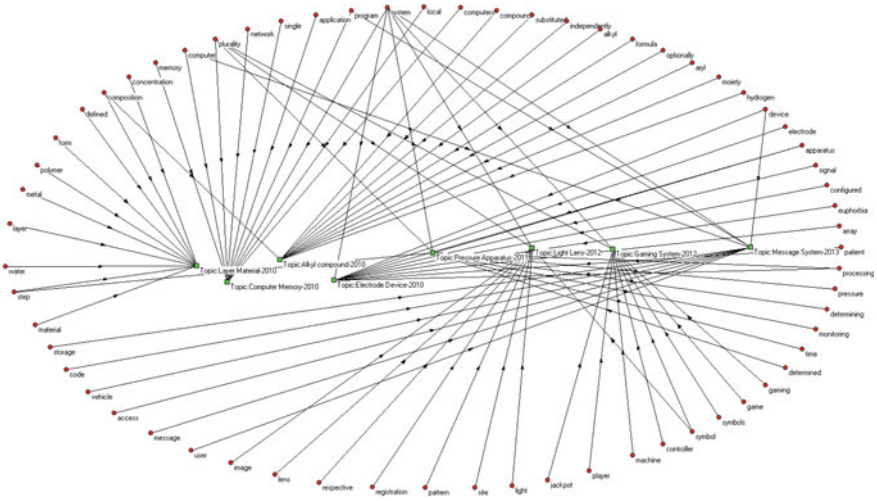

Fig. 11.6 Topics became newly important in each year of 2010–2013 and topmost frequent words of each topic

In year 2010, there are four different topics appeared compared with year 2009, including layer material that related to metal and polymer composition, electrode device, computer memory, and alkyl compound. In year 2011, one newly important topic appeared, pressure apparatus. Then, year 2012 introduced two new topics including light lens and gaming system/controller compared with the previous year. Finally, for year 2013, computer system related to vehicle and message appeared as a new theme. All the topics above were identified without assistance of preset domain knowledge. The detailed words and their corresponding probabilities of the new topics mentioned above are highlighted in boldface in Table 11.3 of the Appendix.

11.4.4 Topic-Based Trend Estimation

As mentioned, we can use the proposed approach to discover how the detailed content of a certain topic evolves from year to year and forecast the topic-based trend using historical status. In the case study, topic antibody fragment/sequence is chosen as an example. As shown in Fig. 11.7, we observe that the word distribution composing the topic develops over time. In year 2009, human and peptide were in the top words list, yet after this, the stress of the topic itself moved to plant, amino acid, nucleic acid, and polypeptide. The word "acid," instead of "antibody," ranked higher from year 2010 to 2013, which means they have larger probability of belonging to this topic as time goes on. The variation of the content of this topic may suggest that, in this area, the key point of technological research and development has shifted to amino/nucleic acid sequence.

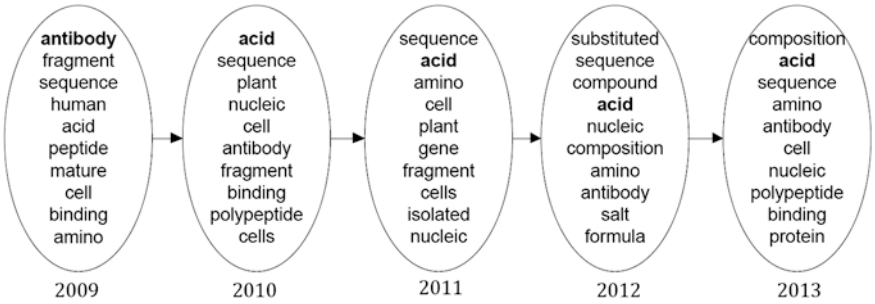

Fig. 11.7 An example of the topic "antibody" evolving over time

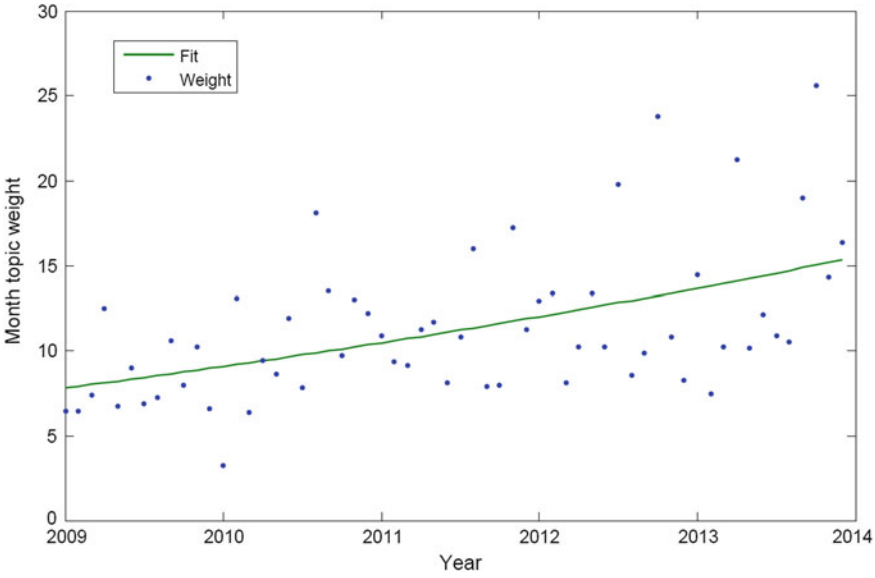

Fig. 11.8 An example of the topic-based trend estimation of the theme "antibody"

To estimate topic-based trend of this topic, we then generate its temporal-weight matrix with one month as time interval. Each element in the matrix presents the weight of the topic in a corresponding time frame, from January 2009 to December 2013. We calculate the weight changes in a least-squares sense to estimate the general trend of the target topic. Figure 11.8 shows the final result of topic-based trend estimation of the theme "antibody." We can obverse directly that this topic appeared to have an upward trend. The significance of this topic kept growing continuously, from which we learn that the research and patenting for the topic of antibody is increasing over the past 5 years, and the importance of this topic has the potential to keep growing in future.

11.5 Conclusion and Future Work

This paper proposed an unsupervised topic change identification approach for patent mining using latent dirichlet allocation. Patent claims that embody the most significant technological terms are chosen as the main textual data source of our research. To improve the usage of LDA in patent topic extraction, we utilize USPC as a measurement of different estimations, to select the optimal model of topic modeling. Machine-identified topics are then placed into a topic change model to locate topic variation over time. Since there is no need to define any keywords in advance and all topics are automatically identified in an unsupervised way, this approach is able to set domain experts and analysts free from reading, understanding and summarizing massive technical documents and records. Finally, a case study, using USPTO patents published during the years 2009–2013 with Australia as their assignee country, is presented. The experimental results demonstrate that the proposed approach can be used as an automatic tool to extract topics and identify topic changes from a large volume of patent documents. From the application perspective, the discovered topic variations can be utilized to assist further decision making in R&D management, especially for newly created innovative enterprises, for example, to provide a full understanding of the topic structure of a certain industry, seek technological opportunities, and so on.

As patents and other technological indicators are generating and accumulating in an increasing rate, approaches for automatically identifying topic changes using data mining and machine learning methods will continue to be emphasized. In future work, we will keep focusing on locating topic changes that associate with more meaningful temporal segmentation, like trend-turning intervals (Chen et al. 2015), to identify and analyze the context that contributes to trend changing of patenting activities.

Acknowledgments 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

See Table 11.3.

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.

Batagelj, V., & Mrvar, A. (2004). *Pajek—Analysis and visualization of large networks*. Berlin: Springer.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*, 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning research, 3*, 993–1022.

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

Camus, C., & Brancaleon, R. (2003). Intellectual assets management: From patents to knowledge. *World Patent Information, 25*, 155–159.

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

Daim, T. U., Kocaoglu, D. F., & Anderson, T. R. (2011). Using technological intelligence for strategic decision making in high technology environments. *Technological Forecasting and Social Change, 78*, 197–198.

David, D., Lewis, Y. Y., Rose, T. G., Li, F. (2004). *SMART stopword list* [Online]. Cambridge: MIT Press. Available: http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

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

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 5228–5235.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems, 17*, 107–145.

Haywood, S. (2003). *Academic vocabulary* [Online]. Nottingham: Nottingham University. Available: http://www.nottingham.ac.uk/alzsh3/acvocab/wordlists.htm, 2014

Heinrich, G. (2005). *Parameter estimation for text analysis*, version 2.9 ed. Darmstadt, Germany: Fraunhofer IGD.

Kim, D., & Oh, A. (2011). Topic chains for understanding a news corpus. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing*. Berlin, Heidelberg: Springer.

Koltcov, S., Koltsova, O., & Nikolenko, S. (2014). Latent dirichlet allocation: stability and applications to studies of user-generated content. In *Proceedings of the 2014 ACM conference on Web science*. Bloomington, Indiana, USA: ACM.

Lai, K.-K., & Wu, S. J. (2005). Using the patent co-citation approach to establish a new patent classification system. *Information Processing and Management, 41*, 313–330.

Lukins, S. K., Kraft, N. A., & Etzkorn, L. H. (2010). Bug localization using latent Dirichlet allocation. *Information and Software Technology, 52*, 972–990.

Nishijima, Y., Anzai, T., & Sengoku, S. (2013). Application of bibliometric analysis to market analysis. In Proceedings of the 2013 Portland International Conference on Management of Engineering & Technology (pp. 2365–2377).

Noel, G. E., & Peterson, G. L. (2014). Applicability of Latent Dirichlet Allocation to multi-disk search. *Digital Investigation*.

Novelli, E. (2014). An examination of the antecedents and implications of patent scope. *Research Policy*.

Office U.S.P.A.T. (2015). *United States Patent and Trademark Office* [Online]. Available: http://www.uspto.gov/

Porter, L. A. (2005). QTIP: Quick technology intelligence processes. *Technological Forecasting and Social Change, 72*, 1070–1081.

Sheikh, N., Gomez, F. A., Yonghee, C., & Siddappa, J. (2011). Forecasting of advanced electronic packaging technologies using bibliometric analysis and Fisher-Pry diffusion model. In Proceedings of the 2011 Portland International Conference on Management of Engineering & Technology (pp. 1–20).

Sheldon, J. G. (1995). *How to write a patent application*. Practising Law Institute.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. KDD workshop on text mining (pp. 525–526), Boston.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Ed.), *Latent semantic analysis: A road to meaning*. Laurence Erlbaum.

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

Trippe, A. J. (2003). Patinformatics: Tasks to tools. *World Patent Information, 25*, 211–221.

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing and Management, 43*, 1216–1247.

USPTO. (2012). *Manual of patent examining procedure: Claim interpretation* [Online]. USPTO. Available: http://www.uspto.gov/web/offices/pac/mpep/s2111.html

Watts, R. J., & Porter, A. L. (1997). Innovation forecasting. *Technological Forecasting and Social Change, 56*, 25–47.

Wikipedia. (2014). *Transitional phrase* [Online]. Wikipedia. Available: http://en.wikipedia.org/wiki/Transitional_phrase, 2014.

WIPO. (2002). *Patent cooperation treaty (PCT) Article 6* [Online]. Washington: WIPO. Available: http://www.wipo.int/pct/en/texts/articles/a6.htm

WIPO. (2004). *WIPO intellectual property handbook: Policy, law and use*.

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

Yang, L., Qiu, M., Gottipati, S., Zhu, F., Jiang, J., Sun, H., & Chen, Z. (2013). Cqarank: Jointly model topics and expertise in community question answering. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 99–108). ACM.

Yang, S., & Soo, V. (2012). Extract conceptual graphs from plain texts in patent claims. *Engineering Applications of Artificial Intelligence, 25*, 874–887.

Yoon, B. (2008). On the development of a technology intelligence tool for identifying technology opportunity. *Expert Systems with Applications, 35*, 124–135.

Yoon, B., & Park, Y. (2005). A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. *Technological Forecasting and Social Change, 72*, 145–160.

Yoon, J., & Kim, K. (2012). TrendPerceptor: A property–function based technology intelligence system for identifying technology trends from patents. *Expert Systems with Applications, 39*, 2927–2938.

Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change, 85*, 26–39.

Zhu, D., & Porter, A. L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. *Technological Forecasting and Social Change, 69*, 495–506.