FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY

Topic-based Analysis for Technology Intelligence

Hongshu Chen

A thesis submitted for the Degree of

Doctor of Philosophy

University of Technology, Sydney
January, 2016

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

This thesis is the result of a research candidature conducted jointly with another University as part of a collaborative Doctoral degree. I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as part of the collaborative doctoral degree and/or fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

_____

ACKNOWLEDGEMENTS

First and foremost I would like to express my sincere gratitude to my supervisors, Prof. *Jie Lu* and Prof. *Guangquan Zhang*, who not only offered me a unique opportunity to study in UTS, but also have provided tremendous guidance and support for my research and life in the past four years. I still remember the first day I met Prof. *Jie Lu* in her office, she knew that I was so nervous to step into a new area of research, thus told me, "doing research is just like climbing up a mountain, the higher you go, the more beautiful scenery you will see". Her enthusiasm for research and the excellent example she has provided as a successful researcher were so motivational, has become my life treasure. I also would like to thank Prof. *Guangquan Zhang* for all his continuous guidance, contribution of ideas, time, and discussions during my study. Both of my supervisors have supported me academically and emotionally through the road to finish my PhD study. Their respectful personality and precise academic attitude have benefited me so much. Without their excellent supervision and continuous encouragement, this thesis could not have been finished on time. I really appreciate all the kind help from them.

As a dual-degree PhD student, I would like to express my earnest thanks to my supervisor in the Beijing Institute of Technology (BIT), Professor *Donghua Zhu*, who gave me an enormous amount of help and support since the year 2009. I have significantly benefited from his invaluable suggestions and support. Without his kind help and professional supervision, this research could not have been accomplished either.

My experience at the Decision Systems and e-Service Intelligent (DeSI) Lab in the centre for Quantum Computation and Intelligent Systems (QCIS), has been more than amazing. The great lab members of mine have contributed immensely to my professional

and personal time at UTS. I would like to thank all the friends that provide me with kind support during my PhD study. Especially, I would like to acknowledge Dr. *Dianhuang Wu*, Mr. *Junyu Xuan*, Mr. *Fan Dong* and Mr. *Yi Zhang* for their help and advice on my experiments. In addition, I am grateful to all members of the Knowledge Management and Data Analysis lab, School of Management & Economics, BIT. Thanks for all their support and help.

I gratefully acknowledge the funding sources that made my research possible. Special thanks go to China Scholarship Council and UTS. Furthermore, thanks to QCIS Travel Fund, FEIT Travel Fund, Vice-Chancellor's Postgraduate Conference Fund for providing me with financial support for international conferences travelling. I also would like to express my since thanks to Ms. *Sue Felix* and Ms, *Jemima Moore* for helping me to correct English presentation problems in my publications.

I want to dedicate this thesis to my grandfather, who has passed away in year 2005, just before I received my college acceptance letter. He has always told me how much he treasured me, trusted me and loved me, and how important for a person to be educated, which shaped my values and made me the person that I am today.

Last but not least, I would like to thank my family for their continuous support during my PhD study. I have amazing parents, who have provided me with unconditional support and encouragement all these years. My family is my energy source and the harbour of my heart.

I love both of you, Mom and Dad.

Hongshu Chen

7th January 2016
At Sydney

ABSTRACT

Since the past several decades, scientific literature, patents and other semi-structured technology indicators have been generating and accumulating at a very rapid rate. Their growth provides a wealth of information regarding technology development in both the public and private domain. However, it has also caused increasingly severe information overload problems whereby researchers, analysts and decision makers are not able to read, summarize and understand massive technical documents and records manually. The concept and tools of technology intelligence aims to handle this issue.

In the current technology intelligence research, one of the big challenges is that, the frameworks and applications of existing technology intelligence conducted semantic content analysis and temporal trend estimation separately, lacking a comprehensive perspective on trend analysis of the detailed content within an area. In addition, existing research of technology intelligence is mainly constructed on the fundamentals of semantic properties of the semi-structured technology indicators; however, single keywords and their ranking alone, are too general or ambiguous to represent complex concepts and their corresponding temporal patterns. Thirdly, systematic post-processing, forecasting and evaluation on both content analysis and trend identification outputs are still in great demand, for diverse and flexible technological decision support and opportunity discovery.

This research aims to handle these three challenges in both theoretical and practical aspects. It first quantitatively defines and presents temporal characteristics and semantic properties of typical semi-structured technology indicators. Then this thesis proposes a framework of topic-based technology intelligence, with three main functionalities, including data-driven trend identification, topic discovery and comprehensive topic

evaluation, to synthetically process and analyse technological publication count sequence, textual data and metadata of target technology indicators. To achieve the three functionalities, this research proposes an empirical technology trend analysis method to extract temporal trend turning points and trend segments, which help with producing a more reasonable time-based measure; a topic-based technological forecasting method to first discover and characterize the semantic knowledge underlying in massive textual data of technology indicators, meanwhile estimating the future trends of the discovered topics; a comprehensive topic evaluation method that links metadata and discovered topics, to provide integrated landscape and technological insight in depth. In order to demonstrate the proposed topic-based technology intelligence framework and all the related methods, this research presents case studies with both patents and scientific literature. Experimental results on Australian patents, United States patents and scientific papers from Web of Science database, showed that the proposed framework and methods are well-suited in dealing with semi-structured technology indicators analysis, and can provide valuable topic-based knowledge to facilitate further technological decision making or opportunity discovery with good performance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND

The increasingly rapid technology development has dramatically accelerated the emergence and accumulation of data and information that indicate technologies, such as patents, scientific literatures, national research and development (R&D) project records, gross domestic R&D expenditure and so forth, and has brought the research on technological management and decision making to the age of Big Data (McAfee et al. 2012). In the intense technological competition nowadays, either for academic or commercial purposes, it is all very important to understand the overall technological landscape and gain technical insight effectively and intelligently, thus to determine what and how to gain from these changes (Camus & Brancaleon 2003; Porter & Cunningham 2004; Ketata, Sofka & Grimpe 2015). With the purpose of obtaining potential opportunities in the intense technological competition nowadays, the demand for efficient analysis on large volumes of technology indicating documents is becoming increasingly important for future development of organizations ranging from individual ones like companies to the multinational level like government unions (TFAMWG 2004; Daim, Kocaoglu & Anderson 2011; Daim 2015).

The rapidly shifting technology environment raises severe information overload, which introduces big challenges but also rich opportunities for modern society. Since manually conducting content analysis and trend identification on massive and multiple science & technology information resources is very time consuming and laborious (Tseng, Lin & Lin 2007), technology intelligence, which covers intelligent methods and tools to

automatically reveal implicit knowledge, is used to handle the issue and assist further technological decision making (Chen, Zhang, Zhu, et al. 2015; Daim 2015; Safdari Ranjbar & Tavakoli 2015). Just like business intelligence to support business decision-making, technology intelligence is promising to turn "data" found in patents or scientific literatures into "knowledge", and help users survive data tsunamis and eventually, to succeed in strategy making (Kerr et al. 2006; Niu, Lu & Zhang 2008; Mortara et al. 2009b).

The concept of technology intelligence was first presented in supplier management research, to keep a finger on the pulse of new technology worldwide (Shapiro 1985). Its related research then gradually expanded from statistics-based to data-mining-based and has become more intelligent in recent years, where an increasing number of researchers focused on the use of extremely powerful information technologies and a vast amount of available data (Daim, Kocaoglu & Anderson 2011; Zhang et al. 2015). However, the partial process of technology intelligence has become one of the most important reasons that organizations' failures in confrontation with radical technological changes (Lichtenthaler 2007), which has led the discussion on how to systematically structuring the amount of information that is available these days (Schuh, Brakling & Apfel 2014). Specifically, one of the big challenges is that, the frameworks and applications of existing technology intelligence usually conduct the semantic content analysis and temporal trend estimation separately, lacking a comprehensive perspective (Veugelers, Bury & Viaene 2010; Safdari Ranjbar & Tavakoli 2015). In addition, while conducting analysis, although much effort has been devoted to identify valuable terms and phrases from semi-structured technology indicators using computer-aided text mining, keyword-based morphology, term clumping and so forth (Watts & Porter 2003; Yoon & Park 2005; Cascini & Russo 2007; Zhang et al. 2014), the outcomes of text mining are mostly keywords with ranking. These words alone, however, are usually too general or misleading to indicate complex concepts and their corresponding temporal patterns, especially when there are polysemous words actually describing different topics (Tseng, Lin & Lin 2007). Thirdly, to achieve the aim of understanding valuable latent knowledge from multiple data resources, and improving the methods of doing so in technology intelligence applications, after emblematic content detection, systematic post-processing,

forecasting and evaluation on both content analysis and trend identification, outputs needs to be contacted to provide more meaningful decision support (De Battisti, Ferrara & Salini 2015).

In the past five years, in order to provide more managerially utilizable extraction and informative representations of technological concepts, topic model-based approaches have attracted an increasing research interest. In particular, as a well-known probabilistic topic model, Latent Dirichlet Allocation (LDA) has provided aid in analysing citation networks, time gaps, content comparison and scientific maps of scientific publications in various areas (Ding 2011; Jeong & Song 2014; De Battisti, Ferrara & Salini 2015; Suominen & Toivanen 2015). However, a fixed topic number is needed for LDA implementation, which raises a new requirement for researchers to understand the possible latent thematic structure of the corpus before topic modelling (Suominen & Toivanen 2015). In addition, Gibbs sampling, one of the most commonly used approximate inference algorithm for parameter estimation, produces different results each time in executing LDA, making the determination on the final topic set even more complicated.

1.2 RESEARCH QUESTIONS AND OBJECTIVES

Facing the limitations and challenges in existing research, this research aims to develop a topic-based technology intelligence framework and a set of data-driven methods to improve the design and applications of existing research, especially for comprehensive analysis on semi-structured technology indicators.

1.2.1 RESEARCH QUESTIONS

To summarize, the following research questions will be answered by this research:

Question 1. How should the temporal characteristics and semantic property of semi-structured technology indicators be quantitatively defined and measured?

From a semantic perspective, many studies has been attempted to identify valuable terms and phrases from scientific publications and patent documents. Nevertheless, the

outcomes of text mining-based techniques applied to assist topic extraction are mostly keywords with a ranking. These words alone are usually too general or misleading to indicate complex concepts. From a temporal perspective, much effort has also been devoted to the study of empirical technology trend analysis, focusing on applying restrained fitting models to provide tendency estimation. However, existing temporal trend measurement used in technology forecasting is based on the assumption of technology life cycle and carried out by curve fitting. The fitting curves do not have any noticeable trend turning points to express detailed trend patterns. Thus, how to quantitatively define and measure the temporal characteristic and semantic property of semi-structured technology indicators is the first research question.

Question 2. How to systematically process technological publication count sequences, textual data and metadata of typical semi-structured technology indicators in technology intelligence?

Typical technology indicators, such as scientific literature, patents or technical project reports, share very similar semi-structured organization, comprising structured metadata, and unstructured textual data. Moreover, their publication activities, shown as publication count sequences, contain temporal trend patterns indicating the positive relationship between the publication activities and R&D activities. To comprehensively deal with the temporal characteristics and semantic property of semi-structured technology indicators, their publication count sequences, textual data and metadata need to be processed and analysed synthetically.

Question 3. How to extract detailed trend pattern from technological publication count sequences and conduct empirical technology trend analysis?

Temporal characteristics underlying in publication activities are seldom taken into consideration while processing data, in technology intelligence framework or application research. Existing model-driven empirical technological trend analysis usually suffers from model choosing and limitations that a certain model may bring. It is quite difficult to reveal all the trend patterns by only restrained models. The fitting curves used to present temporal patterns do not have any noticeable trend turning points to correlate

semantic extraction with a specific time interval, either. Under conditions of rapid technologies development, budgetary constraint and time limitation, an increasing number of demands have created new requirements for the technological trend analysis process especially in technology intelligence research, that is, to be more effective and less time consuming.

Question 4. How to dynamically link the temporal characteristics and semantic property in technology intelligence application, to analyse the future trend for each topic?

Although text mining techniques such as information extraction, text cleaning, text clustering have been used to access knowledge hidden in massive documents, it is usually very difficult to provide a corresponding temporal measurement to the text mining outputs. In real-life situations, even one patent document may contain a number of different technological topics. It is also difficult to reveal the trend of specific topics using keyword-based text mining techniques, since single keywords are usually too general to represent concepts and reflect their corresponding temporal pattern.

Question 5. How to effectively identify prominent topics underlying in massive technical textual data and comprehensively correlate metadata to provide evaluation in depth?

To achieve the aim of improving the approaches of understanding valuable thematic knowledge from massive documents, there are two main phases which need to be considered. The first one is automatically detecting latent knowledge from textual data more accurately. The second one is assisting further thematic evaluation using discovered topics or emblematic keywords. In real-life situations, further evaluation is needed to identify prominent topics, providing a more accurate presentation of the semantic knowledge, by connecting it well with the descriptive metadata of the target technology indicators.

1.2.2 RESEARCH OBJECTIVES

This research aims to achieve the following objectives, which are expected to answer the above research questions:

Objective 1. To develop a three-dimensional schematic structure of the semantic property and temporal characteristics of semi-structured technology indicators, such as scientific literature and patents.

This objective corresponds to the research questions 1 and 2. To face the difficulties of quantitatively measuring and integrating the semantic property and temporal characteristics, this research uses 'topics' to represent the semantic property of all target technology indicators, and use 'trend segments' and 'trend turning points' to define the temporal characteristics. The concept 'topic' here is represented as a group of soft-clustered words that frequently show up together in a collection of documents; while 'trend turning points' and 'trend segment' indicate the occurrence time of significant changes of the publication activities and how long the changes take. Considering the entire knowledge presentation in a semantic space and a temporal dimension, concepts in the semantic space actually have different levels of development. To provide decision makers with more comprehensive awareness of the relationship between the semantic property and temporal characteristics, this research first aims to develop a three-dimensional schematic structure of the semantic property and temporal characteristics of target technology indicators.

Objective 2. To propose a topic-based technology intelligence framework that systematically processes and analyses technological publication count sequences, textual data and metadata of typical semi-structured technology indicators.

This objective corresponds to the research question 2. After defining the semantic properties and temporal characteristics of target technology indicators, this research designs and establishes a topic-based technology intelligence framework, to face the challenges of processing not only textual data, but also publication count sequence and metadata, for technological decision making support or opportunity discovery. Automatic identification and quantitative analysis of topics, trend turning points and trend segments lay the important foundation of this thesis.

Objective 3. To propose a data-driven empirical technological trend analysis method.

This objective corresponds to the research question 3. From a temporal perspective, to capture detailed trend patterns of publication activities and proceed with future trend estimation, at the same time to overcome the limitations of using only restrained models, this research aims to propose a data-driven empirical technological trend analysis method, which mainly explains the trend identification functionality of the proposed topic-based technology intelligence framework. Publication count sequences serve as the main input of technological trend analysis research, since they are the observation of historical publication activities of technology indicators.

Objective 4. To propose a topic-based technological forecasting method and conduct afterwards content analysis.

This objective corresponds to the research question 4. In real-life situations, even one scientific paper or one patent document may contain a number of different technological topics. To integrate the temporal trend patterns and semantic topics quantitatively, this research aims to propose a topic-based technological forecasting method and conduct afterwards content analysis, to discover and estimate the trends for specific topics underlying large volumes of patent claims.

Objective 5. To propose an empirical topic detection and evaluation method based on topic modelling.

This objective corresponds to the research question 5. While detecting latent knowledge from textual data and conducting thematic evaluation, there are several limitations to be considered in a real world application, when applying topic modelling into the proposed topic-based technology intelligence. This research aims to identify a prominent topic set to represent the semantic knowledge of a target technical corpus and link descriptive metadata, propose an empirical topic detection and evaluation method based on topic modelling, to eventually access effective post-topic-modelling evaluation for a comprehensive overview of the technological landscape and an in-depth insight of scientific details.

1.3 RESEARCH SIGNIFICANCE

The significance of this research work can be summarized from the theoretical, technical and practical aspects as follows:

1.3.1 THEORETICAL SIGNIFICANCE

Significance 1: theoretically, the research develops a framework of Topic-based Technology Intelligence with three main functionalities: (1) trend identification functionality; (2) topic discovery functionality; (3) comprehensive topic evaluation functionality.

This research defines and measures the temporal characteristics and semantic property of typical technology indicators quantitatively with a three-dimensional schematic structure and develops a topic-based technology intelligence framework to process and analyse technological publication count sequence, textual data and metadata synthetically. In addition, this research specifies the input, output and detailed components of the framework.

Significance 2: technically, the research develops a data-driven technological trend analysis method to capture the underlying trend patterns of technological publication activity and further technology forecasting.

This research proposes a technological trend analysis method with technological trend identification, analysis and forecasting functionalities. The contributions of this method include: from a data-driven perspective, developing an innovative solution for empirical technology trend patterns identification and future trend forecasting by quantitatively identifying and depicting the concept “trend” with trend segments; overcoming the limitations of model choosing and upper limits estimating which is experienced by existing empirical technology forecasting approaches; learning valuable trend patterns from historical patent counts record, then using the learned trend turning points and trend segments to predict future technology trend.

Significance 3: technically, the research develops a topic-based technological forecasting method, which is designed to uncover the latent topics and temporal trends underlying massive technical documents, and to evaluate to what degrees various topics have contributed to the patenting activities of the whole area.

The contributions of this method include: it proposes a stepwise methodology to quantitatively integrate temporal trend patterns and semantic topics in different dimensions, to provide topic-based technological trend forecasting; it estimates the developing trends for specific latent topics, instead of a broad technological area; it overcomes the limitation where simple keywords and rankings are too general or misleading to indicate a concept, and reflect its corresponding temporal pattern.

Significance 4: technically, the research develops an empirical topic detection and evaluation method to discover and characterize the estimated topic after applying topic modelling to a target technical corpus.

The contributions of this method include: it selects the most representative topic set to explain a target corpus of technology indicator, as well as linking metadata with the estimated topics to provide valuable evaluation on various themes; it presents three indices, which comprehensively bring the metadata of publications into consideration, and quantitatively characterizes the weight, developing trend and activeness of all discovered topics; it conducts topic-based citation and the corresponding citation distribution analysis to feature the influence and potential usefulness of each topic.

1.3.2 PRACTICAL SIGNIFICANCE

From the viewpoint of application, the proposed topic-based technology intelligence framework and a set of methods can be used for comprehensive analysis on all semi-structured technology indicators. This research presents case studies with both patents and scientific literature, to show how to apply the proposed methods in real cases. Experimental results on Australian patents, United States patents and scientific papers from Web of Science database, showed that the proposed framework and methods are well-suited in dealing with semi-structured technology indicators analysis, and can

provide valuable topic-based knowledge to facilitate further technological decision making or opportunity discovery with good performance. Since the approach is data-driven, time and effort can be saved from model choosing and saturation stage estimating in real-world analysis and forecasting tasks.

A full understanding of the underlying technological topics distribution and trends in target areas are essential for both newly created innovative enterprises and venture capitalists. This understanding enables entrepreneurs to prepare appropriate technical proposals with potential while at the same time providing venture capitalists with the confidence to support companies with a better understanding of a certain industry. From the perspective of academic use, a full understanding of topic-based research strength and interest are significant for seeking potential funding and collaboration opportunities. The framework and methods proposed in this research can be used to automatically uncover the thematic structure of massive technical data, estimate the detailed developing trend of each detected topic, and link the metadata with discovered topics, thereby assisting decision making for potential opportunity identification, technical strategy formation, and so forth, for both industrial and academic purposes.

1.4 Research Methodology and Process

This research mainly belongs to the information system domain. The concept of research methodology in this domain is defined as follows, “collections of problem solving methods governed by a set of principles and a common philosophy for solving targeted problems” (Gallupe 2007). A number of research methodologies, such as case study, field study, design research, archival research, field experiment, laboratory experiment, survey and action research, have been proposed and applied in the information system domain (Niu 2009; Wu 2014; Vaishnavi & Kuechler 2015).

1.4.1 Research Methodology

In this study, design research is utilized as the research methodology according to the analysis of the research questions and objectives. As the research has focused on the development of new framework, methods, algorithms and strategies in the technology

intelligence design and application, and the soundness of these methods, techniques or algorithms must be supported by the results from the experimentations and evaluations. Thus the experimental approach integrated with the standard information system research cycle was chosen as the proposed research method. The methodology of design research is illustrated in Figure 1-1.

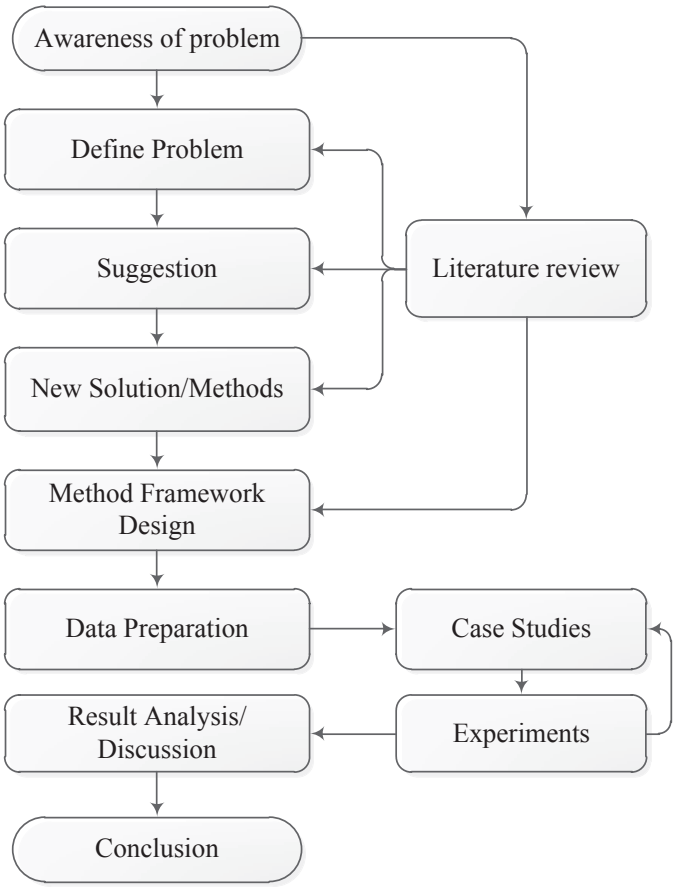


Figure 1-1. The methodology of design research of this thesis

Awareness of problems is the starting point of a design research, at which limitations of existing applications are examined and meaningful research problems are identified (Vaishnavi & Kuechler 2015). This research then defines the problems, makes suggestions and proposes new methods. Suggestion is a creative process during which new concepts, models and functions of artefacts are demonstrated (Wu 2014). Based on the new ideas, the fundamental work of this research is to design a framework of the topic-based technology intelligence. Under the direction of the framework, three main

functionalities are developed, where there are methods targeting on processing and analysing technological publication count sequence, textual data and metadata synthetically are proposed. This research uses scientific literature analysis and patent analysis as background, thus after data collection, the case studies are mainly focused on these two domains. While doing experiments in case studies, this research adjusts the experimental procedure and conduct result analysis and discussion. Finally, a conclusion is the final step of a design research effort (Vaishnavi & Kuechler 2015).

1.4.2 RESEARCH PROCESS

This research was planned according to the methodology of design research. First of all, a subject of technology intelligence was chosen as a research topic of this research. A literature review of existing research in this area was conducted, and existing literature was retrieved and critically reviewed. The results of the literature review provide useful information and suggestions on identifying valuable research questions, which are directly addressed this research. As the research questions grew clearer and more definite, more literature closely related to the research questions was reviewed (Wu 2014).

Because the existing work in the literature lacks the ability to deal with technology intelligence framework and methods to systematically process and analyse semi-structured technology indicators, this research proposed a three-dimensional schematic structure to measure the semantic property and temporal characteristics of technology indicators. The proposed models and approaches were implemented and evaluated. According to the methodology of design research, this research is an iterative process. The output of each research step might be fed back to its previous step when deviations between expectations and evaluation results are found. Through the feedback, research outcomes are progressively improved until satisfying results are drawn from evaluations. The developed methods were then implemented in a real case of scientific literature analysis and patent analysis. Finally, writing up the PhD thesis is done at the end of the research.

1.5 THESIS STRUCTURE

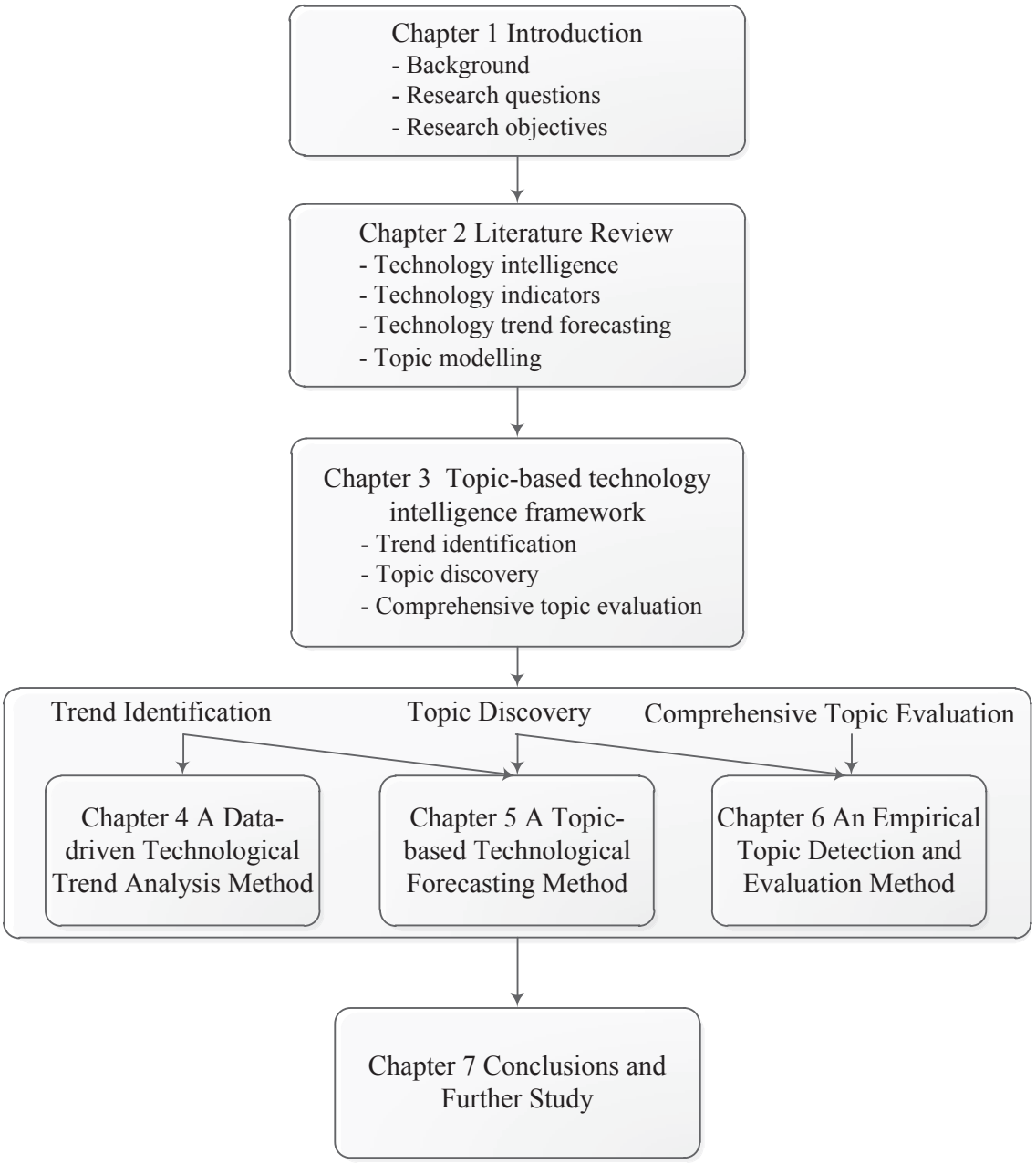


Figure 1-2. Thesis structure

This thesis contains seven chapters. Chapter 1 presents the research background, research questions, objectives, significance, research methodology and process, and the thesis structure. Chapter 2 presents the literature relevant to this study, including the

concepts and techniques of technology intelligence, technology indicators, the existing approaches of empirical technology trend forecasting, the concept and methodologies of tech mining, topic modelling and its applications in the tech mining area. Chapter 3 proposes a framework of topic-based technology intelligence, with three main functionalities. Based on Chapter3, Chapter 4 develops a data-driven empirical technological trend analysis method. Chapter 5 develops a topic-based technological forecasting method. Chapter 6 develops an empirical topic detection and evaluation method. Chapter 7 presents the conclusions and further study recommendations. The structure of the thesis is shown in Figure 1-2.

1.6 PUBLICATIONS RELATED TO THIS THESIS

Below is a list of the refereed international journal and conference papers associated with my PhD research that have been submitted, accepted and published:

- Papers published and accepted:

1. Chen, H., Zhang, G., Zhu, D., and Lu, J. 2015, 'A patent time series processing component for technology intelligence by trend identification functionality', *Neural Computing and Applications,* vol. 26, no. 2, pp. 345-353. (ERA Tier B)

2. Chen, H., Zhang, Y., Zhang, G., Zhu, D., and Lu, J. 2015, 'Modelling Technological Topic Changes in Patent Claims', *The 2015 Portland International Conference on Management of Engineering & Technology Portland*, Portland, USA, 2-6 August, pp. 2049 – 2059. (ERA Tier A)

3. Chen, H., Zhang, G., Lu, J., Zhu, D. 2015, 'A Fuzzy Approach for Measuring Development of Topics in Patents Using Latent Dirichlet Allocation', *The 2015 IEEE International Conference on Fuzzy Systems (IEEE FUZZ)*, Istanbul, Turkey, 2-5 August, in press. (ERA Tier A)

4. Chen, H., Zhang, G., and Lu, J. 2013, 'A Time-Series-Based Technology Intelligence Framework by Trend Prediction Functionality', *The 2013 IEEE*

International Conference on Systems, Man, and Cybernetics (SMC), Manchester, 13-16 October, pp.3477-3482. (ERA Tier B)

5. Chen, H., Zhang, G., Lu, J., and Zhu, D. 2014, 'A Two-Step Agglomerative Hierarchical Clustering Method for Patent Time-Dependent Data', *Proceeding of the seventh International Conference on Intelligent Systems and Knowledge Engineering,* Springer, pp.111-121. (ERA Tier B)

6. Chen, H., Zhang, Y. & Zhu, D. 2015, 'Identifying Technological Topic Changes in Patent Claims Using Topic Modeling', in P.A. Daim T, Chiavetta D, Saritas O (ed.), Anticipating Future Innovation Pathways through Large Data Analytics, USA. (Acceptance date: 25th November 2015)

7. Zhang, Y., Zhang, G., Chen, H., Porter, A., Zhu, D., Lu, J. 2016, 'Topical Analysis and Forecasting for Science, Technology and Innovation: Methodology and a Case Study focusing on Big Data Research', *Technological Forecasting and Social Change*, DOI: 10.1016/j.techfore.2016.01.015. (ERA Tier A)

8. Zhang, Y., Chen, H., Zhang, G., Zhu, D., and Lu, J. 'Multiple Science Data-Oriented Technology Roadmapping Method', *The 2015 Portland International Conference on Management of Engineering & Technology Portland*, Portland, USA, 2-6 August, pp. 2278 – 2287. (ERA Tier A)

9. Zhang, Y., Chen, H. & Zhu, D. 2015, 'Semi-automatic Technology Roadmapping Composing Method for Multiple Science, Technology, and Innovation Data Incorporation', in P.A. Daim T, Chiavetta D, Saritas O (ed.), Anticipating Future Innovation Pathways through Large Data Analytics, USA. (Acceptance date: 6th October 2015)

- Papers under review:

10. Chen, H., Zhang, G., Zhang, Y., Zhu, D., and Lu, J. 2014, Empirical Topic Detection and Evaluation Approach for Scientific Literature', submitted to *Journal of Informetrics*. (ERA Tier A)

11. Chen, H., Zhang, G., Lu, J., Zhu, D. 2015, Topic-based Technological Forecasting Base on Patent Data: Methodology and a Case Study', submitted to *Technological Forecasting & Social Change*. (ERA Tier A)

CHAPTER 2
LITERATURE REVIEW

This chapter presents a discussion of relevant concepts and work in connection with topics of this thesis. In Section 2.1, the concepts and techniques of technology intelligence are expatiated. Section 2.2 explains technology indicators, which represent the concepts "Technology" in this research. Section 2.3 reviews the existing approaches of empirical technology trend forecasting. Section 2.4 introduces existing research on topic modelling and its applications in tech mining area. Last but not least, Section 2.5 provides a summary of Chapter 2.

2.1 TECHNOLOGY INTELLIGENCE

The concept of technology intelligence was first systematically mentioned in supplier management research for understanding and monitoring of new technology worldwide (Shapiro 1985). The most effective manufacturers in this study pursued the use of technology intelligence to stay abreast of relevant technological development and understand likely future substitutes for current products and processes.

2.1.1 CONCEPT AND FRAMEWORK OF TECHNOLOGY INTELLIGENCE

Summarizing existing technology intelligence concept research, it can be viewed as an 'activity' to be conducted by a set of agents, or as a knowledge management 'product' with consumers, and it provides an organization with the capability to capture and deliver information in order to develop an awareness of technology threats and opportunities

(Kerr et al. 2006). Compare to the traditional expert-based approach, technology intelligence uses objective data, not expert grading, as the fundamental data source. It is capable to deal with larger quantities of information that are not be able to analyse by humans alone, and also has ability to generate knowledge by integrating resources from different sources to visualize the outcomes (Mortara et al. 2009a, 2009b; Yoon & Kim 2012; Safdari Ranjbar & Tavakoli 2015). Technology intelligence makes it possible to monitor the direction of technology growth and plan for technology R&D for both government and private companies.

More specifically, Kerr and his colleagues (2006) gave a definition of technology intelligence as follow, "technology intelligence is the capture and delivery of technological information as part of the process whereby an organisation develops an awareness of technology threats and opportunities". Technology intelligence tools have several advantages including, the ability of analysing large quantities of data, generating useful information which humans cannot produce, and supporting decision making with knowledge by integrating resources from different sources to visualize the outcomes (Yoon 2008; Yoon & Kim 2012).

In existing theory research, as shown in Figure 2-1, technology intelligence as a decision making service concept with fuzzy boundaries in different points of view, covers a combined area of artificial intelligence, strategy formulation, technology evaluation, knowledge discovery, innovation management, and even how to train employees to get used to changes of technology impact, and some previous studies considered the intersection of these fields as technology intelligence service areas (Savioz 2004).

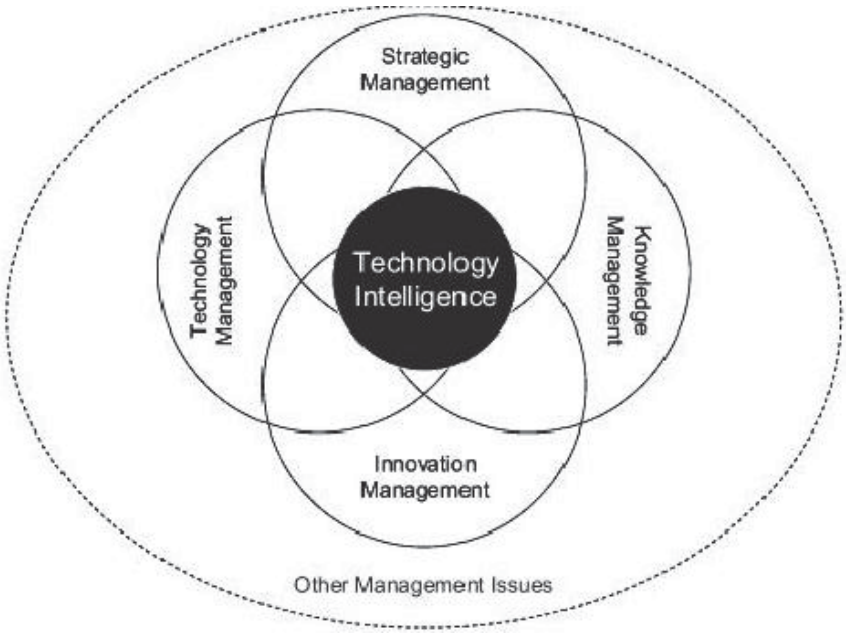


Figure 2-1. Technology intelligence service define by Savioz (2004)

However, this scope is way too large for building technology intelligence systems in practice. It can be imaged that if all the business processes are regarded, the structure will be more directed to process management, the detailed analysis and design of intelligent systems will be simplified accordingly. To solve this problem, a conceptual model built by Kerr (2006) presented four main functions including Trawl, Scan, Mine and Target for technology intelligence, that narrowed down technology intelligence systems with only data from external and internal sources as input and intelligence information for decision makers as output. Figure 2-2 shows the process of the concept framework they designed. As mentioned, this framework described technology intelligence as an activity to be conducted by a set of agents, at same time as a knowledge management outcome with consumers. It provides an organization with the capability to capture and deliver information in order to develop an awareness of technology threats and opportunities. But as it described the model only in the concept level, a framework for more detailed and functional technology intelligence systems is still needed.

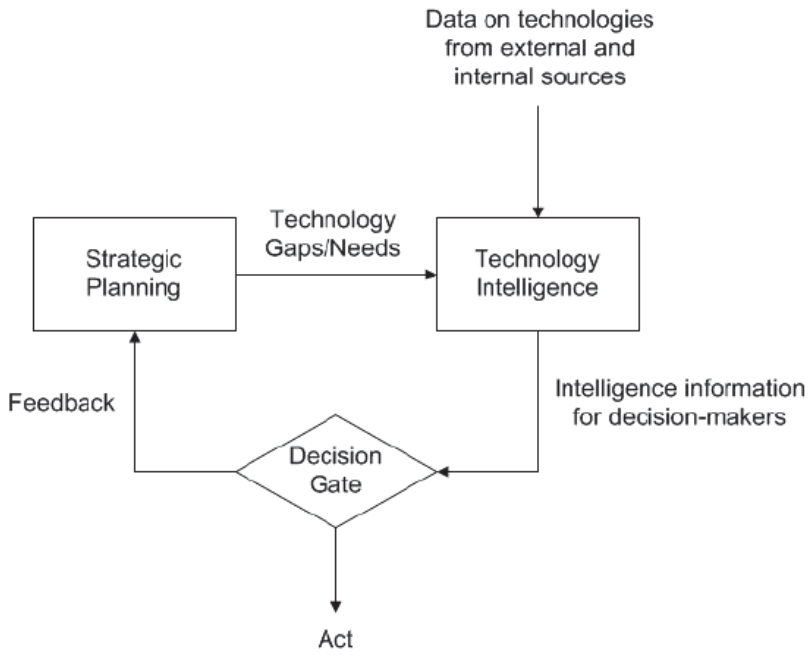


Figure 2-2. The technology intelligence process defined by Kerr at al. (2006)

2.1.2 TECHNOLOGY INTELLIGENCE IN PRACTICE

Research shows that the most fundamental topic changes of a recent Portland International Conference on Management of Engineering and Technology (PICMET), which has the theme of Technology Management in the Age of Fundamental Change, is that an increasing number of researchers focused on use of extremely powerful information technologies and a vast amount of available data that digitally provide us with technology intelligence (Daim, Kocaoglu & Anderson 2011). As a systematic method to improve the effectiveness of R&D activities, technology intelligence has been introduced to help with technology development planning and technology strategies formulation in practice (Yoon & Kim 2012).

In previous study, many application of technology intelligence concept have focused on revealing the hidden knowledge of various technology indicators. Researchers mainly focused on using technology intelligence to carry out patent semantic analysis (Tseng et al. 2007; Wang & Cheung 2011; Yoon & Kim 2012) and technology trend forecasting (Lee, Jeon & Park 2011). *TrendPerceptor* (Yoon & Kim 2012), is proposed to identify trends in invention concept by using a property-function based approach. *Techpioneer*

(Yoon 2008), uses text mining and morphology analysis, is an expert system for excavating potential technology opportunities. *VantagePoint* and *Aureka* are another two intelligent systems, which can support users to analyse trends or relationships in technology by providing clustering, mapping and searching techniques (Zhu & Porter 2002; Trippe 2003). Besides, a visualization system called *Diva*, aims at helping users to perform bibliometric analysis of scientific literature and patents for trend forecasting (Morris et al. 2002). In addition, *PatentMiner*, a topic-driven patent analysis and mining system, focused on studying the heterogeneous patent network derived from the patent database, and derive several patent analytics tools that can be directly used for IP and R&D strategy planning (Tang et al. 2012).

While reviewing the applied research of technology intelligence, it can be noticed that since this concept was presented, research on technology intelligence application gradually expanded from experience-based to data mining-based and became more intelligent in recent years. Researchers have been working on recognising various reasons of organizations' failures in confrontation with radical technological changes (Safdari Ranjbar & Tavakoli 2015). One of the most important reasons has been the partial process of technology intelligence (Lichtenthaler 2007), which means the research on systematically structuring the amount of information that is available these days has to be done (Schuh, Brakling & Apfel 2014). Specifically, the existing design and tools of technology intelligence are mainly constructed on the fundamentals of semantic properties of patent documents or scientific literatures. Time-related property of patents or scientific literature publication/application activities are seldom taken into consideration while processing data, the outcome trends/relations/changes are mainly text-based. In other words, without considering the properties of target technology indicators in a comprehensive perspective, only the semantic attributes mainly being discussed. Thus more work on quantitatively linking time-related characteristic and semantic property is still needed to be done.

2.1.3 TECH MINING

As mentioned in sub-section 2.1.2, the design and tools of technology intelligence are in existing research focus on analysing semantic properties of technical documents and files mainly. Actually, approaches and techniques of tech mining play a very important role of implementing content analysis in technology intelligence research.

Manually conducting content analysis and evaluation on massive technical textual data of scientific publications is very time consuming and laborious, hence, automatic approaches to empirically reveal topics and trends hidden in science and technology information, also known as tech mining, attracts an increasing interest in the past decade (Porter & Cunningham 2004; Cunningham, Porter & Newman 2006; Tseng, Lin & Lin 2007). More specifically, tech mining is short for "text mining of science & technology information resources" (Porter 2009). It has a key premise, that is, intelligence is a prime requirement for efficient and effective technology management. Under such circumstance, tech mining is more than just one technique, but covers approaches of text cleaning, text clustering, topic detection and so forth.

Tech Mining can also be considered a form of "Content analysis" (Porter 2009). It has blossomed with the availability of electronic datasets and analytical software. As mentioned, classic technology forecasting approaches seek to extract useful trends from numerical data. Tech Mining, aims not only to include numerical data sources analysis, but also to exploit textual data of different technology indicators. In particular, watch the distinction between unstructured text sources and structured text sources (Tseng, Lin & Lin 2007; Porter 2009). In the perspective of technology intelligence, tech mining plays an important role of understand the semantic knowledge from semi-structured files and documents, with techniques of Computational Linguistics (Artstein & Poesio 2008), Natural Language Processing (Manning & Schütze 1999), Knowledge Discovery in Databases (Piateski & Frawley 1991), and Machine Learning (Bishop 2006).

2.2 TECHNOLOGY INDICATORS

One of the most important issues to utilize technology intelligence in practice is to find a representative database with sufficient data supply first. That is, an indicator for

the concepts “Technology” and also “Technology Innovation” needs to be determined (Trajtenberg 1990). In existing studies, the researches addressing Science, Technology, & Innovation (ST&I) activities are widening into multiple perspectives (Bengisu 2003). Technology indicators, involving with academic publications, patents, academic proposals, etc., provide possibilities for describing previous scientific dynamics and efforts, discovering innovation capabilities, and forecasting probable evolution trends in the near future (Porter & Detampel 1995). Considering the convenience of collecting and organizing data, summarizing existing research, the most frequently used and well accepted technology indicators are scientific literature and patent files (Basberg 1983; Garfield, Malin & Small 1983; Griliches 1990; OECD 2005; Radicchi & Castellano 2012; Mohammadi et al. 2015). The following sub-sections will review the study on using these technology indicators to technological decision making support research and practice.

2.2.1 PATENT

Patent is the main manifestations of intellectual property for a company or a government sector, also a very important resource for the continuous development of technology. They contain the entire information of developed technologies and provide an inventor with exclusive rights on special technical knowledge by limiting the access possibilities of other companies to this special technology (Herce 2001). The study of using patents as indicators for technology analysis has begun from early 1980s, since then patent analysis became a very useful tool that utilized to support technology research and development (R&D) planning, competition analyses, and analytic studies of how technologies emerge, mature and disappear(Basberg 1983; Campbell 1983; Griliches 1990).

As an indicator of technology changes, patent publication and application behaviour can be seen as dynamic system that is affected by factors, such as the technology innovation and upgrading, political environment, economic situation, intellectual property rights infringement and protection and so forth. Under recent circumstances of fast and complex technological advances, data mining for patent analysis has become increasingly important to decision making of both private companies and governments

for continuous technology development and risk reducing. In general, previous study of patent analysis focuses on monitoring technological horizon, forecasting the future trend (Cozzens et al. 2010), identifying emerging technology (Bengisu & Nekhili 2006; Robinson et al. 2011), clustering and classifying technologies (Chen et al. 2014) and constructing technology intelligence systems (Yoon & Kim 2012).

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 (Lai & Wu 2005; Sheikh et al. 2011; Nishijima, Anzai & Sengoku 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, in recent years, there has been a lot of interest in applying text mining techniques to unstructured patent data to set domain analysts free from studying and understanding massive amounts of technological content (Camus & Brancaleon 2003; Tseng, Lin & Lin 2007).

Patents are the ideal data source for technology intelligence study. On one hand, since a positive relationship between R&D activities and subsequent patenting activities has been found (Griliches 1990), the value of utilizing patent data for empirical research has been emphasized increasingly in recent years. On the other hand, it is not difficult to obtain patent data from public patent offices of many different countries for academic study or commercial business purposes. This makes patent analysis a useful and convenient method to support technology R&D planning, competition analyses, and analytic studies of how technologies emerge, mature and disappear (Campbell 1983; Faust & Schedl 1983; Ernst 1997).

When doing research, the patent database of United States Patent and Trademark Office (USPTO) is mainly used for standardization reason. This is because patents submitted in other countries are often simultaneously submitted in United States, which making USPTO database the most representative and standard system for technology

analysing (Lee et al. 2012). UTSPTO patent database contains all US patents from 1790 to today, which are classified by both International Patent Classification (IPC) and United States Patent Classification (USPC) (USPTO 2012a). One technological field usually covers several categories of patents, which makes the classification and clustering the foundation of technology analysis.

2.2.2 PATENT CLAIMS

Patent claims, as an important part of unstructured segments of a patent document, embody all the important technical features of an invention with the most essential technological terms to define the protection (Tong & Frame 1994; Yang & Soo 2012). They reflect to the core inventive idea of a patent, at the same time they provide a direct technological scope in which patent examiners classify the patent to different technological classes (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 (Sheldon 2001; USPTO 2012c; Yang & Soo 2012). For example, "a method of producing a soya bean product, the method including the step of exposing soya beans to an acidic aqueous solution", is a typical claim, in which "a method of producing a soya bean product" is the preamble; "including" serves as the transition phrase; "the step of exposing soya beans to an acidic aqueous solution" is the 'body' that contains the elements of the invention.

In short, patent claims, in technical terms, define the scope of protection (Mailänder 2013). These claims must meet relevant patentability requirements, such as novelty and non-obviousness. On one hand, they reveal the core inventive topics and the major technological scope of a patent; on the other hand, they are written in concise, but precise language, which make them the best resource for identifying technological topics and facilitating patent document analysis (WIPO 2002; Yang & Soo 2012; Xie & Miyazaki 2013; Novelli 2014). Research showed that even using only the first 300 words from the abstract, claims, and description sections, the performance is better than those using the

full texts regardless of which classifiers are used (Tseng, Lin & Lin 2007; Tseng et al. 2007).

2.2.3 SCIENTIFIC LITERATURE

Scientific literature presents original work with theoretical and experimental thinking, plays an important role in profiling R&D and estimates scientific trends for potential innovation assistance. It contains plenty of technical terms and academic-related words that describe research outcomes. Since the abstract of a scientific publication usually embodies the technological scope and the essential terms to define the main theme of the paper, they are the ideal resource to identify and analyse scientific topics. To achieve the aim of understanding valuable thematic knowledge from massive scientific literature, and improve the approaches for doing so, there are two main phases need to be considered. The first is automatically detecting latent knowledge from textual data. The second is assisting further thematic evaluation using discovered topics or emblematic keywords.

Much effort has already been devoted to identifying valuable terms and phrases from scientific publications and patent documents, for the first phase, using computer-aided text mining, keyword-based morphology, term clumping and so forth (Watts & Porter 2003; Yoon & Park 2005; Cascini & Russo 2007; Zhang et al. 2014). Nevertheless, the outcomes of text mining-based techniques applied to assist topic extraction are mostly keywords with a ranking. These words alone, however, are usually too general or misleading to indicate a concept, especially when there are polysemous words actually describing different topics (Tseng, Lin & Lin 2007). Thus in past five years, topic model-based approaches that provide more managerially utilizable extraction and informative representations of technological concepts have attracted increasing research interest. In particular, Latent Dirichlet Allocation (LDA), one of well-known probabilistic topic models, has provided aid in analyzing citation networks, time gaps, content comparison and scientific maps of publications in various areas (Ding 2011; Jeong & Song 2014; De Battisti, Ferrara & Salini 2015; Suominen & Toivanen 2015).

2.3 EMPIRICAL TECHNOLOGY TREND ANALYSIS AND FORECASTING

In technology management issues, technology trend analysis and forecasting are the most studied topics in the context of technology research and development (R&D) management, in both public and private domains. It can be traced back to the 1950s when technology forecasting was first studied as a mere sub-task of project arranging (Gerybadze 1994). With the impact of new technologies and the accelerated accumulation of scientific and technological records, it soon drew much greater attention as a core process of technical strategy formulation, implementation to gain competitive advantage, and most importantly, the construction of technology intelligence, for private companies and governments (Coates et al. 2001; Zhu & Porter 2002; Kerr et al. 2006; Lichtenthaler 2007).

Specifically, technology trend analysis has been employed in prioritizing R&D projects, and creating strategic alliances such as licensing and joint ventures due to the intensive economic competition between businesses (Barker & Smith 1995). From another level, governments also require technology trend forecasting to advance public agendas in the face of increasing rates of technology change and budgetary constrain (Zhu & Porter 2002). The trend forecasting on technologies, bases on the understanding of the technical horizon, in previous study, researchers summarized its role as two main parts. First of all, technology forecasting makes it possible to monitor the direction of technology growth and plan for technology R&D for both government and private companies (Yoon & Park 2007). Second, by exploring new technology alternatives and technical developing trends, technology forecasting recognizes emerging technology that is attracting increasing attention nowadays by introducing more opportunities and threats. What's more, it can also evaluate the potential and significance of a specific technology. Technology trend forecasting has been applied to several technology intelligence cases (Morris et al. 2002; Yoon & Kim 2012), and in this research, as a main function of learning-based technology intelligence systems, technology forecasting will be discussed in the form of modules combination.

In practice, forecasting the future trends of a technology draws on both empirical evidence and expert evaluation (Zhu & Porter 2002). Compared with expert-experiences dominant approaches which suffer from a relatively expensive collection procedure, on the one hand, empirical technology trend forecasting sets panels free from analysing large volumes of data that increases and accumulates every moment; on the other hand, it builds a bridge between trend patterns and the observations derived from technology indicators such as patents, scientific literatures and R&D expenditure. Thus the study and application of this topic have attracted increasing attention within the past decade.

2.3.1 CURVE FITTING BASED APPROACHES

Most existing studies of empirical trend identification and forecasting based on various approaches of curve fitting (Gupta et al. 2002; Baskurt 2010; Rao & Srivastava 2010; Lv et al. 2011). Among them the most accepted and adopted empirical technology forecasting is the growth curves method, which presents a causal analogy between technology performance advance and a living organism growth (Martino 1993). The term S-curves usually depicts a similar modality with growth curves, so as the term extrapolation in the context of technology forecasting. Typically, a sigmoidal model will be fitted to the observing historical data in the general procedure of growth curves (Carrillo & González 2002). The logistic curve and Gompertz curve are the two most frequently used growth curves by technological forecasters.

The logistic curve is also known as the Pearl Curve, which has the formula as Equation 2-1. The expression of the Gompertz curve is presented in Equation 2-2 (Martino 1993; Bengisu & Nekhili 2006).

$$y = \frac{L}{1+ae^{-bt}} \tag{2-1}$$

$$y = Le^{-ae^{-bt}} \tag{2-2}$$

where y is the variable representing performance for both curves, L is the saturation value, showing the upper limit of variable y, e is the base of the natural logarithms, t is time, a and b are the coefficients obtained by fitting the curve to the observation. As t varies, both the logistic curve and the Gompertz curve range from 0 to L. The logistic

curve is a symmetrical curve, which has the inflection point at $t = (lna)/b$, where $y = L/2$. However, the Gompertz curve is not symmetrical. The inflection point of this curve occurs at $t = (lna)/b$, where $y = L/e$. Figure 2-3 provides an example of using curve fitting approaches in technology forecasting.

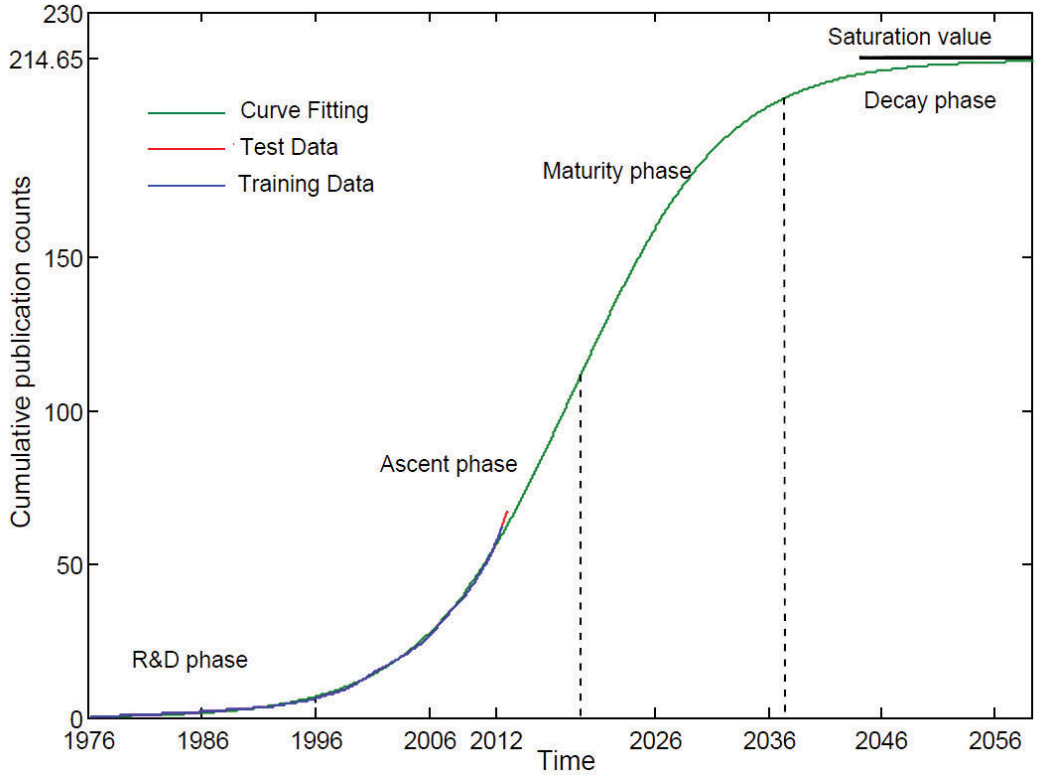


Figure 2-3. An example of curve fitting approaches in technology forecasting

There are three important hypothesises for growth curve method. These are the upper bound of the growth curve is known in advance; a correct curve is chosen to be fitted to the observation data; and the historical data is able to give the coefficients for the chosen model correctly (Martino 1993). 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. Thus although the simple computation and straightforward presentation that the growth curve provides quite workable for general trend identification, when it comes to capture and forecast the periodical trend patterns and shifts, a more data-based approximation needs to be produced.

2.3.2 OTHER APPROACHES

Besides the growth curves, it should also be mentioned that, Jabło´nska-Sabuka et al. (2014) proposed an trend identification and prediction approach based on population dynamics with Burgers'type global interaction, allowing a detailed analysis. In addition, Hidden Markov Model (HMM) approach was used to focus on modelling patterns of innovation and clustering technologies with similar patterns (Lee, Lee & Yoon 2011; Lee, Lee & Yoon 2012), which has brought machine learning theory into the technology trend analysis area, but future trend forecast work is still needed. Heuristics methods are also used to deal with technology clustering problem by using patents count on base of ICP (Dereli et al. 2011). Yoon and Kim (2012) mentioned they will use machine learning approach to improve the measurement and classification function of their system, including support vector machines and k-nearest neighbours. Inspired by the principle of machine learning, this research will propose the innovative concept and framework of learning-based technology-intelligence systems, which is based on explicit patent properties and unknown patterns learning from the patent data.

Time series analysis, as "a most direct" way to forecast the parameters that measure technologies" (Porter et al. 1991; Hossain et al. 2011), is brought into the research because of technology indicators actually have "time-related" feature. Time series has been successfully used into science, engineering and business for years, and it is now still of interest and actual research tool for researchers. A time series is a sequence of data points, measured typically at time instants spaced at uniform time intervals (Palit & Popovic 2005). Time series forecasting is the use of a model to predict future values based on previous observations. Jun and Uhm (2010) considered frequent time series model based on patent data. In their study, they forecasted a trend of technology using Moving Average Method. However, technology forecasting usually expects meaningful trend or future patterns of technological development as outcomes, rather than precisely predicted values of the observation. In other words, in the context of technology trend forecasting, an abstract quantitative representation of real-world dynamics is necessary, that is, an approximation of original observation needs to be presented. Therefore, although classical time series analysis approaches can deliver much more accurate result

than other empirical methods in time series prediction, growth curves models are more widely accepted when dealing with technology forecasting tasks, e.g., when searching "growth curve", "S-curve" or "extrapolation" in the Journal of Technological Forecasting and Social Change, the number of publications during the past 20 years is 360, and for keywords "ARIMA", "ARMA" or "moving averages" the number is 59 in the same time interval.

Except above approaches, bibliometric analysis is also a widely used method for technology forecasting since 1980s. It uses counts of publications, patents, or citations to measure and interpret technological advances (Watts & Porter 2003). Historically, bibliometric methods have been used to trace back academic journal citations and then researchers find out that it is useful to help understanding the past and even potentially forecasting the future. Three major forms of bibliometric analysis have emerged citation analysis, patent analysis, and publication analyses (Garfield, Malin & Small 1983). This method could provide a nicely accessible and cost-effective data or information. It helps to explore, organize and analyse amounts of historical data helping researchers to identify “hidden patterns” that may help researchers in the technology forecasting and decision making process (Chen, Chen & Lee 2011). In addition, with the help of Biblimetric analysis, system dynamic method was also used in this area, to model the dynamic ecosystem of the technologies and their diffusion (Daim et al. 2006).

However, although bibliometrics can help with figuring out the hidden pattern of a particular technology, it focuses on the historical evaluation result more and will not be able to show the prediction directly. Quantitative measures have very seldom been used to forecast technical development trend, even though patents data and literatures data have been used a lot in technology management and evaluation research. Actually, the derived technological impacts tend to be higher when the patent is cited more frequently, more recently, and more intensively in the time units of the past. Future technological impacts can be directly assessed by using estimated future quantity or citation counts, instead of using data in the past (Lee et al. 2011).

2.3.3 PIECEWISE LINEAR REPRESENTATION

The concept of piecewise approximation is brought into technology trend analysis heuristically by Philips (1999). He proposed a piecewise linear regression method to capture trend changes of the polyvinyl chloride price. In 2001, Keogh and his colleagues proposed piecewise linear representation (PLR), which is suitable for capturing short-term tendency and trend turning point, since trend patterns hidden in massive time series are easier to observe when data is simplified (Keogh & Pazzani 1998; Keogh et al. 2001; Keogh & Kasetty 2003; Keogh et al. 2004). Owing to its ability to decompose a time series into several compressed segments, PLR approach has been applied to time series mining in stock prediction (Chang, Fan & Liu 2009; Luo & Chen 2013) and audio signal analysis (Kimura et al. 2008) in recent studies. Generally, PLR refers to the approximation of a time series P, of length n, with k straight lines. It represents a time series with a number of line segments end to end.

There are many kinds of time series segmentation algorithms which appear under different names in the research; however, their implementation has slight differences. Most approximation algorithms can be summarized into one of following three types(Luo & Chen 2013):

- **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.

In this research, a bottom-up algorithm is used to segment the patent time series into a number of straight lines. The algorithm begins by creating the finest possible approximation of the time series, which makes n/2 segments to approximate the n-length time series. Then it calculates the cost of merging each pair of adjacent segments and

starts to iteratively merge the lowest cost pair, until a stopping criterion is met (Luo & Chen 2013). Patterns of the trend become easier to obtain after the segmentation, and the straight lines produced by the PLR algorithm are transformed into a new sequence of trend signals for future forecasting work.

Summary the existing concepts and approaches used for empirical technology trend analysis. It is not hard to see that an approximation of the real-world data is needed for trend pattern recognition. There are various ways to transform the original data to the trend representative data. The more the original data are approximated, the less accurate the result will be. Under such circumstance, the level of the approximation needs to be controlled to obtain a result which is as accurate as possible. A visualized comparison among the well-accepted S-curve model, PLR-based approaches and classical time series analysis methods in the context of technology forecasting is shown in Figure 2-4.

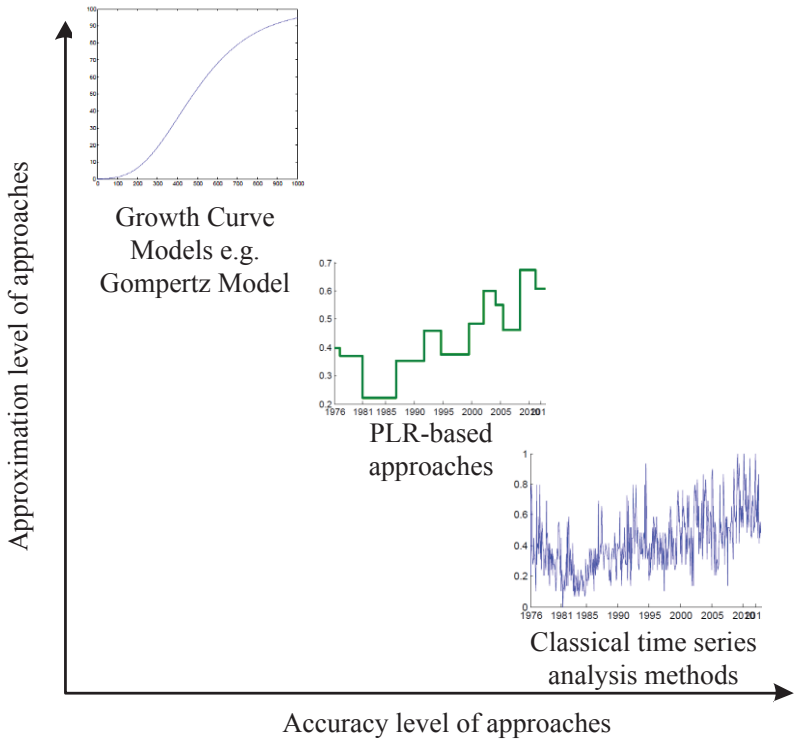


Figure 2-4. Comparison of growth curves, PLR-based approaches and time series analysis

2.4 TOPIC MODELLING

A topic model, also called statistical topic model, is a method to extract topics automatically from a certain corpus of text documents (Steyvers & Griffiths 2007; Srivastava & Sahami 2009; Blei 2012). 'Topics' in topic modelling mean distributions of words compose a collection of documents. A word may occur in many topics with a different probability, however, with a different typical set of neighbouring words in each topic. Because of the nature of language usage, the vocabularies show higher frequency when constitute a topic, or a 'word soft cluster', are often semantically related. Topic models originally were developed to automatically index, search, cluster, and structure massive unlabelled textual documents. Due to its ability of efficiently calculating the 'co-occurrence' of items, topic modelling is now widely used in natural language processing, information retrieval, and machine learning.

The earliest concept related to topic modelling was mentioned in 1970s, by Salton and his colleagues (1975), in their research of vector space model. Vector space model, or term vector model, presents text documents mathematically, as vectors of identifiers. However, one of the biggest limitations of the vector space model is that long documents are poorly represented since the dimensionality can be very large. In 1988, Salton and Bucley (1988) proposed the Term Frequency-Inverse Document Frequency (TF-IDF) method to reflect how important a word is to a document in a collection or corpus. Now TF-IDF has been used as an important weighting factor in text mining and information retrieval.

In 1990，Deerwester et al. (1990) innovatively brought the concept of 'semantic structure' to the research, proposed latent semantic analysis approach to take advantage of implicit higher-order structure in the association of terms with documents in order to improve the detection of relevant documents. Since latent semantic analysis approach is very time consuming, in 1999, Hofmann (1999) proposed the probabilistic latent semantic analysis model, using expectation-maximization algorithm to learn all the parameters needed, brought machine learning techniques to the research of text mining. In 2003, based on the research of probabilistic latent semantic analysis, Blei et al. (2003)

proposed LDA, which was first presented as a graphical model for topic discovery. The LDA model is similar to probabilistic latent semantic analysis, except that in LDA the topic distribution is assumed to have a Dirichlet prior. In practice, this results in more reasonable mixtures of topics in a document (Blei 2012).

2.4.1 SEMANTIC SPACE

In recent years, statistical topic models have been applied as an efficient tool to discover underlying but potentially useful content in large volumes of textual data (Blei 2012). The reason why topic models are being emphasized is that, instead of simply explaining concepts in textual data with 'keywords', topic models measure the probability of words 'co-occurrence' in a corpus. In such perspective, the semantic meaning of a concept can be better presented, and at the same time, it also opens a door to the possibility of better evaluating the outcome of after extracting latent topics (Chen, Zhang, Zhang, et al. 2015). Figure 2-5 gives a brief introduction of the 'semantic space' of a topic model. It follows the assumption of 'bag-of-words', which means text (a sentence or a document) is represented as the bag (multi-set) of its words, disregarding grammar and even word order but keeping multiplicity (Blei, Ng & Jordan 2003). Figure 2-5 present a brief introduction of the semantic space of topic modelling.

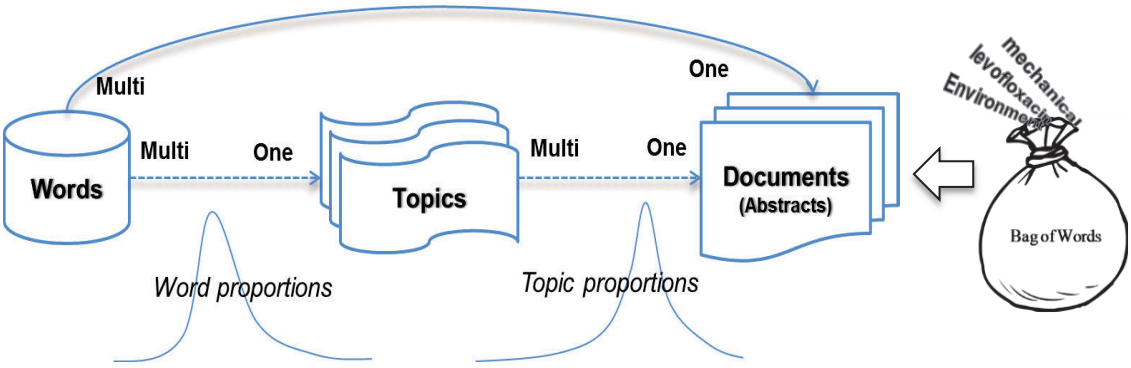


Figure 2-5. Brief introduction of the 'semantic space' of topic modelling

Traditionally, the terms and documents construct a semantic space. Topic modelling actually performs dimensionality reduction, relating each document to a position in low-dimensional 'topic' space. Each document is viewed as a mixture of various topics and

they are assumed to be characterized by a particular set of topics. In addition, each topic is presented with a word distribution; the vocabularies got top highest possibility that belonging to the topic can be used to present its semantic meaning. For example, a topic has probabilities of generating various words such as apple, fruit, agriculture, grow, red, which can be interpreted by the viewer as 'fruit apple', while another topic has probabilities of generating words such as apple, company, smart, technology, mac, can be classified as 'apple company'.

2.4.2 LATENT DIRICHLET ALLOCATION

LDA (Blei, Ng & Jordan 2003) is a probabilistic model that aims to estimate the properties of multinomial observations by unsupervised learning. So far it is one of the most well-known and widely-accepted topic models. It gives an estimation of the latent semantic topics hidden in large archives of documents, and indicates the probabilities of how various documents belong to different topics.

LDA has been used as a very efficient tool to assist topic discovery and analysis, in practice. For example, Ding (2011) introduced topic-dependent ranks based on the combination of a topic model and a weighted PageRank algorithm; Griffiths and Steyvers (2004) applied LDA-based topic modelling 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 generative process of LDA can be denoted by the joint distribution of the random variables as Formula 2-3,

$$p(\vec{w}_d, \vec{z}_d, \vec{\vartheta}_d, \phi|\vec{\alpha}, \vec{\beta}) = \prod_{n=1}^{N_d} p(w_{d,n}|\vec{\varphi}_{z_{d,n}}) p(z_{d,n}|\vec{\vartheta}_d) p(\vec{\vartheta}_d|\vec{\alpha})\, p(\phi|\vec{\beta}), \quad (2\text{-}3)$$

where graphical model of LDA is presented in Figure 2-6, 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 d^{th} document in document collection D. Each node in Figure 2-4 stands for a random variable in the generative process of LDA and all the three plates indicate replication. On the left of the figure, $\vec{\vartheta}_d$ stands for the topic proportions for the d^{th} document. 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. 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. The shaded circles are observable nodes, where $W_{d,n}$ stands for the n^{th} word in document d. All of the unshaded circles indicate hidden nodes. 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, Ng & Jordan 2003; Heinrich 2005; Steyvers & Griffiths 2007; Jeong & Song 2014; Koltcov, Koltsova & Nikolenko 2014).

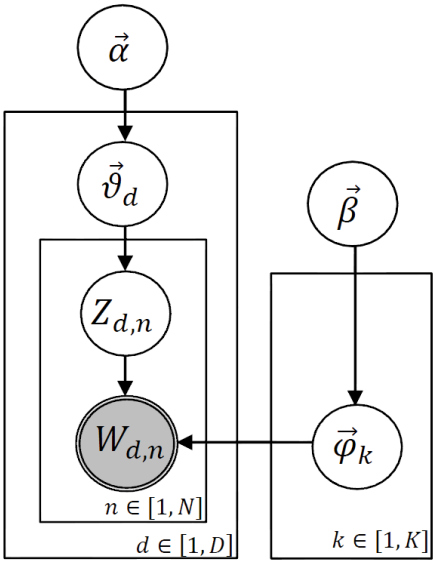


Figure 2-6. The graphical model of Latent Dirichlet Allocation

The parameters of LDA need to be estimated by 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) and widely used to estimate the assignment of words to topics by observed data (Griffiths & Steyvers 2004; Lukins, Kraft & Etzkorn 2010; Noel & Peterson 2014). The randomness introduced by the initiation of the sampling affects the estimation of

probabilities in LDA, so that the result is slightly different even with exactly the same setting of input and parameters; yet on the whole, the results of different experiments won't change much.

After proposing LDA, Blei and his team members have improved LDA facing various application needs. For example, in 2004, Griffiths et al. (2004) presented a generative model that uses both kinds of dependencies based on hidden Markov model, and can be used to simultaneously find syntactic classes and semantic topics; in 2006, Teh et al. (2006) proposed hierarchical Dirichlet process facing the limitation of determining topics number; in the same year, Blei and Lafferty (2006) presented correlated topic model to analyse the relationships between estimated topics; in 2007, Mcauliffe and Blei (2008) proposed supervised topic model to use labels to improve the accuracy of classification; in 2010, Blei et al. (2010) presented the hierarchical latent Dirichlet allocation to analyse the hierarchy of the topic structure.

2.4.3 Topic Modelling in Tech Mining

Tech mining extracts intelligence from Science, Technology and Innovation records by using text mining techniques and text analytics to examine topical content and relationships (Porter & Cunningham 2004; Zhang et al. 2014). Recently, statistical topic models represented by LDA have been applied to the tech mining area as an efficient tool for discovering the underlying, and potentially useful, content in massive textual data. Ding (2011) introduced topic-dependent ranks based on the combination of a topic model and a weighted PageRank algorithm; Liu and Chen (2013) compared latent topics identified from citing abstracts versus citing sentences to the target reference using LDA; Jeong and Song (2014) proposed a temporal analysis approach based on LDA to use heterogeneous resources in an integrated manner; Battisti et al. (2015) presented an LDA-based post-processing approach to describe a field of research over time; Suominen and Toivanen (2015) validated an unsupervised learning-based map of science with the assistance of LDA.

It is believed that topic models are, and will continue to be, emphasized in this area, is that instead of simply modelling massive textual data based on 'keywords and

frequency', they measure the 'co-occurrence' of words and use their distributions to present concepts. With such perspective, the semantic meaning of a concept can be better delivered. At the same time, the importance of a concept is no longer derived solely from the frequency of a certain single word, and this opens the door to the possibility of better evaluating the outcome of tech mining. However, although LDA can serve as an efficient tool for estimating latent knowledge, it is usually difficult for researchers to come up with a specific number of topics without a very thorough understanding of the observation data. In existing research, the majority of experimental case studies, that apply LDA, guess the topic number K from its thematic structure or for the convenience of presenting results. A perplexity value has already been used to evaluate the performance of an LDA model with different topic numbers, which improved the setting of t K (De Battisti, Ferrara & Salini 2015). Ways to reduce the random effects, introduced by Gibbs sampling, in the final topic set is still a research issue of interest.

2.5 SUMMARY

In summary, there are the following research gaps in previous research:

1) Previous technology intelligence framework and methods have not shown the ability to handle the temporal characteristics and semantic properties of semi-structured technology indicators.

2) Previous technology intelligence framework and applications have not shown the ability to analyse technological publication count sequence, textual data and metadata of target technology indicators comprehensively.

3) Even though technology intelligence has been effectively used as a tool of latent knowledge discovery, it still lacks effective post-topic-modelling evaluation for a comprehensive overview of the technological landscape and an in-depth insight of scientific details.

Facing the limitations and challenges in existing research, this research aims to develop a topic-based technology intelligence framework and a set of data-driven

methods to improve the design and applications of existing research, especially for comprehensive analysis on semi-structured technology indicators.

CHAPTER 3
FRAMEWORK OF TOPIC-BASED TECHNOLOGY INTELLIGENCE

3.1 INTRODUCTION

While reviewing the theoretical and applied research of technology intelligence, it can be noticed that since this concept was presented, research on technology intelligence application gradually expanded from experience-based to data mining-based and became more intelligent in recent years. However, the existing design and tools of technology intelligence are mainly constructed on the fundamentals of analysing semantic properties of scientific literatures or patent documents. Temporal characteristic underlying in publication activities of these technology indicators are seldom taken into consideration while processing data.

In order to improve the design and application of technology intelligence, this chapter proposes a framework of *Topic-based Technology Intelligence*, with three main functionalities: (1) trend identification functionality; (2) topic discovery functionality; (3) comprehensive topic evaluation functionality. These functionalities will be accessed in the following chapters with three main methods.

The contributions of this chapter include:

(1) Defining the temporal characteristic and semantic property of typical technology indicators, after summarizing the limitations of existing technology intelligence research;

(2) Proposing a topic-based technology intelligence framework to process and analyse technological publication count sequence, textual data and metadata synthetically;

(3) Proposing a trend identification functionality in the framework with a piecewise linear representation (PLR) step, to process raw literature/patent count sequence and help with producing a more reasonable temporal measure for further topic analysis;

(4) Proposing a topic discovery functionality in the framework with a Latent Dirichlet Allocation (LDA) step, to discover and characterize the semantic knowledge in target technology indicators by topics, meanwhile overcoming the limitation of assuming topic number and determining the final topic set while topic modelling;

(5) Proposing a comprehensive topic evaluation functionality in the framework, which applies metadata of technology indicators, to provide researchers and analysts with a more complete overview of the technological landscape and insight.

The remainder of this chapter is organized as follows. Section 3.2 introduces the limitations of existing technology intelligence frameworks. Section 3.3 defines the temporal characteristic and semantic properties of technology indicators. Section 3.4 presents a full framework of the proposed topic-based technology intelligence by illustrating the input, output, and the conceptual model step by step. Section 3.5 explains the main steps in three functionalities of the topic-based technology intelligence in detail. Finally, the summary is given in Section 3.6.

3.2 The Limitations of Existing Technology Intelligence Frameworks

Technology intelligence indicates the concept and applications that transform data hidden in patents or scientific literature into technical insight for technology development planning and strategies formulation (Savioz 2004; Kerr et al. 2006; Hongshu Chen 2013). Since this concept was first systematically mentioned (Shapiro 1985), research in technology intelligence has gradually expanded from experience-based to data mining-based, and has become more intelligent in recent years. That is, an increasing number of

researchers focused on the use of extremely powerful information technologies and a vast amount of available data that digitally provides us with technology intelligence (Daim, Kocaoglu & Anderson 2011).

On the theory side, in existing research, as a decision making service concept with fuzzy boundaries, technology intelligence covers a combined area of artificial intelligence, strategy formulation, technology evaluation, knowledge discovery, innovation management, and even how to train employees to get used to changes of technology impact. Thus, some previous studies considered the intersection of these fields as technology intelligence service areas (Savioz 2004). However, it is a way too large scope for researchers to design and construct technology intelligence in practice. Kerr and his colleagues (2006), facing this problem, proposed a conceptual model which contained four main functions including Trawl, Scan, Mine and Target for technology intelligence, which narrowed down technology intelligence systems with only data from external and internal sources as input and intelligence knowledge for decision makers as output. As it described the model only in the concept level, a framework for more detailed and functional technology intelligence systems is still needed.

On the application side, in existing practice research, researchers mainly focused on using technology intelligence to carry out text mining-based content analysis (Yoon 2008; Wang & Cheung 2011; Yoon & Kim 2012) and technology trend forecasting (Morris et al. 2002; Daim & Suntharasaj 2009). Comparing with the traditional expert-based approach, the existing framework can provide effective applications to deal with large volumes of data that are not be able to analysed by humans alone (Yoon & Kim 2012); however, one of the most important issues in practice is that, semantic content analysis and temporal trend estimation were conducted separately, lacking a comprehensive perspective. More specifically, existing frameworks and applications of technology intelligence, represented by Techpioneer (Yoon 2008), TrendPerceptor (Yoon & Kim 2012), VantagePoint (Zhu & Porter 2002) and Aureka (Trippe 2003), are mainly constructed on the fundamentals of semantic properties of patent documents or scientific literatures. Text mining techniques such as information extraction, text cleaning, text clustering are successfully used to access knowledge hidden in massive documents. Yet it

is usually very difficult to dynamically provide a corresponding temporal measurement to the text mining outcomes.

In summary, the existing design and tools of technology intelligence are mainly constructed on the fundamentals of semantic-related feature of technology indicators. Although 'technology forecasting' was mentioned frequently in many previous studies, temporal trend analysis was seldom involved when a technology intelligence structure was considered. Thus, how to analyse the links between text-mining results and their temporal evolution is still an important research question.

3.3 Temporal Characteristics and Semantic Properties of Technology Indicators

This research selects scientific literature and patents as the main data sources of topic-based technology intelligence study, for the following reasons:

(1) As well-accepted indicators, scientific literature and patents are the ideal technology descriptors to support technology R&D planning, competition analyses, and analytic studies of how technologies emerge, mature and disappear (Campbell 1983; Faust & Schedl 1983; Ernst 1997). Specifically, scientific literature explains original work with theoretical and experimental thinking, plays a crucial role in profiling R&D and estimates technological trends for potential innovation assistance; patents, representing the intellectual properties of organizations ranging from individual ones to the multinational level, have been proven to have a positive relationship between R&D activities and their issuing activities (Griliches 1990).

(2) The other reason is that, it is not difficult to obtain scientific literature or patent data from public databases for academic study and commercial business purposes. For standardization reason, the primary source of scientific literature employed in a majority of studies is Web of Science (WoS) database, a premier source of information on

fundamental research (Woon, Zeineldin & Madnick 2011; Zhang et al. 2014; Ciarli, Coad & Rafols 2015). Among patent databases from different countries, the United States Patent and Trademark Office (USPTO) database which contains all US patents from 1790 to today is mostly used since it is the largest commercial market in the world (USPTO ; Lee et al. 2012). This research selects WoS and USPTO as primary data sources for technology intelligence study. Hereinafter, they are named collectively as 'target technology indicators'.

Target technology indicators within a particular scope, hold explicit technical information and hidden knowledge that indicates technological concepts, themes, trend and related R&D activities, which can be used as decision making supports, early warning signals or trend pointers (Campbell 1983; Griliches 1990; Ernst 1997; WIPO 2004). After reviewing the limitations of the existing technology intelligence framework, it is not hard to summarize that the previous work on target technology indicators can be seen in two different perspectives. From a semantic perspective, many studies have been attempted to identify valuable terms and phrases from scientific publications and patent documents, as mentioned in Chapter 2, using computer-aided text mining, keyword-based morphology, term clumping and so forth (Watts & Porter 2003; Yoon & Park 2005; Cascini & Russo 2007; Zhang et al. 2014). From a temporal perspective, much effort has also been devoted to the study of empirical technology trend analysis, focusing on applying restrained fitting models on publication counts to provide a tendency estimation of the corresponding industry or technical area (Porter et al. 1991; Lee et al. 2012; Lee, Lee & Yoon 2012).

Summarizing existing research, users of technology intelligence expect to perceive not only the semantic knowledge hidden in textual data, but also the corresponding trend of that knowledge, which underlies the publication activities and metadata of target technology indicators. Corresponding to the above two perspectives, target technology indicators actually have both semantic properties and temporal characteristics, where the former defines the understandable content of the concerned technologies and the latter shows the time-related trend of this content. Figure 3-1 provides a three-dimensional schematic structure of the semantic property and temporal characteristic of target

technology indicators. If the entire knowledge presentation is considered as having a semantic space and a temporal dimension, concepts in the semantic space actually have different levels of development. Compared to considering the whole framework with only one perspective, the combination of trend underlying in publication activities of target technology indicators and knowledge hidden in their corresponding documents can provide decision makers with more comprehensive awareness of technological advances.

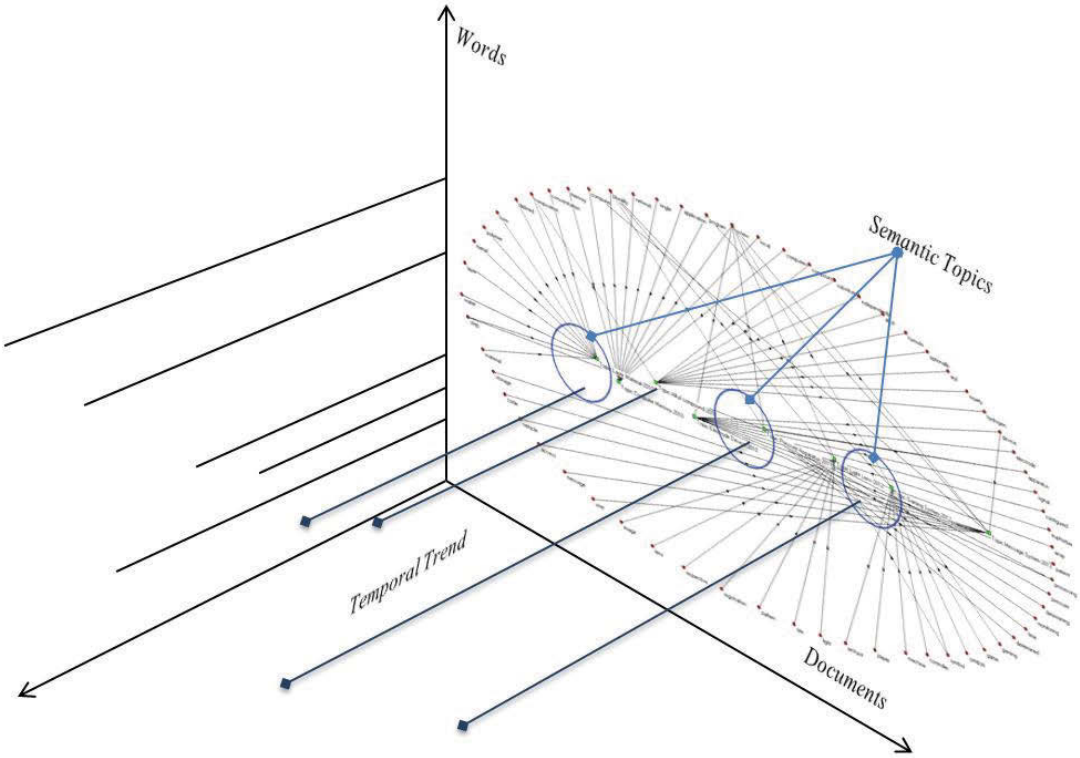


Figure 3-1. The three-dimensional schematic structure of the semantic property and temporal characteristic

The most direct method in existing research to bring the semantic property and temporal characteristic together is to split the original document collection into several sub-collections according to their publication year, season or month, then analyse the time-related semantic evolution between time intervals with equal length (Jeong & Song 2014). However, it is very difficult to quantitatively link the semantic property and temporal characteristic in a real sense, other than just considering thematic changes from year to year. This is because the existing temporal trend measurement is usually based on the assumption of technology life cycle and carried out by curve fitting. The fitting

curves, however, do not have any noticeable trend turning points to correlate semantic extraction with a specific time interval.

To solve this problem, a number of studies have investigated the possible trend turning points on a Growth curve. For example, Trappey and his colleagues (2011) selected 10%, 50% and 90% of the entire growth curve and divided it into four partitions, as four stages include Introduction Stage, Growth Stage, Maturity State and Saturation Stage. Although in such a way, researchers are able to identify several trend turning points on the smooth curve, how to determine the proportion scale is a fuzzy issue; moreover, it is still hard to get more specific trend turning activities using historical publication records.

To face the above difficulties and integrate the semantic property and temporal characteristic in later studies, before building the framework of topic-based technology intelligence, there is first a need to define and clarify how to quantitatively capture and represent the two properties. In this research, as shown in Figure 3-2, 'topics' are used to represent the semantic property of all target technology indicators, and use 'trend segments' and 'trend turning points' to define the temporal characteristic. The concept 'topic' here is represented as a group of soft-clustered words that frequently show up together in a collection of documents, while 'trend turning points' and 'trend segment' indicate the occurrence time of significant changes of the publication activities and how long the changes last. An ideal technology intelligence system needs to have functionalities to process textual data, publication counts and even metadata of the target corpus. Automatic identification and quantitative analysis of topics, trend turning points and trend segments lay the important foundations of this thesis.

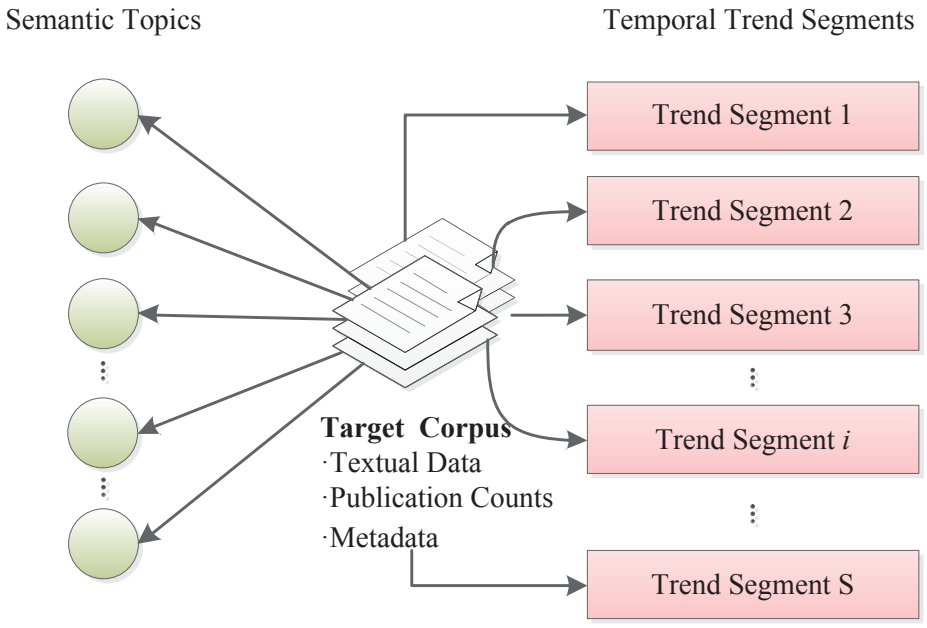


Figure 3-2. 'Topics' and 'Trend Segments'

3.4 Framework Description

After defining the semantic properties and temporal characteristics of target technology indicators, in this section, a topic-based technology intelligence framework is developed to face the challenges of processing not only textual data, but also publication count sequence and metadata, for technological decision making support or opportunity discovery. This section clarifies the input and output of the proposed framework, and explains the concept model and its main components hierarchically.

3.4.1 Input and Output

As mentioned, this research has summarized scientific literature and patents collectively as target technology indicators. Generally speaking, target technology indicators are the input of the proposed topic-based technology intelligence. Actually, although scientific literature and patents are two kinds of documentation with different styles, they share very similar semi-structured organization. Scientific literature comprises structured items such as ISI article number, publication year, subject category and unstructured items like title, abstract, funding information; patents contain structured

items such as patent number, issue date, international patent classification (IPC), United States patent classification (USPC), and unstructured items like title, abstract, claims. Their similar structure makes it is possible to process them with one single technology intelligence framework.

Since theoretically, this study has generated topics to define the semantic property of the target document collection, and has represented the temporal characteristic with trend segments and trend turning points, detailed data input needs to be specified. As shown in Figure 3-3, in this research, there have been three types of inputs for the proposed topic-based technology intelligence framework. Publication count sequence has been used as the observation of historical publication activities. Because topics embody the semantic knowledge underlying in text of technology indicators, specifically, paper titles and abstracts have been selected as one part of the input of scientific literature, and at the same time patent titles and claims have been chosen as the input of patents, as these texts cover the very significant technological features of a research or an invention with concise but precise language. Moreover, to further evaluate the outcome after identifying the semantic topics and temporal trend, the metadata of target papers and patents has been set as a part of the final input, which involves publication years, authors, affiliations or inventors, assignees, IPC and so forth. Through the estimation of topics, trend turning points and trend segments, this research has been able to eventually provide a user of topic-based technology intelligence with output of topic trend detection, topic contribution estimation, and topic evaluation. The three parts of the outputs offer a comprehensive overview of technological landscape in a semantic-temporal-space; that is, the latent semantic structure of massive textual data in a technological area of interest, the detailed developing trend of each detected topic, and the evaluation of each prominent topic.

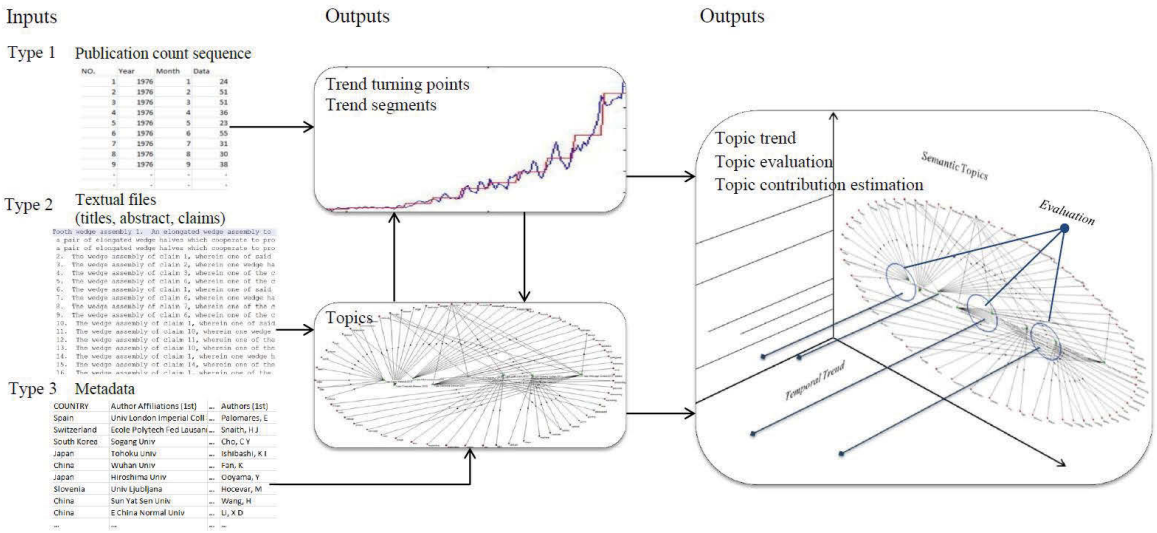


Figure 3-3. Schematic diagram of the input and output of the topic-based technology intelligence

3.4.2 TOPIC-BASED TECHNOLOGY INTELLIGENCE

The existing concept of technology intelligence is built fundamentally on an intersection of artificial intelligence and technology strategy decision support. To satisfy the demands of exploring knowledge from target technology indicators as much as possible, after the very first conceptual framework was designed, it then united with text mining techniques, focusing on content analysis. In this research, in order to provide more integrated knowledge of the target technology domain more in depth, topic modelling, time series analysis, metadata analysis and fuzzy number-based approaches are applied to enrich the existing framework. Specifically, this research proposes a topic-based technology intelligence framework, including three major functionalities, trend identification functionality, topic discovery functionality and comprehensive topic evaluation functionality, as shown in Figure 3-4.

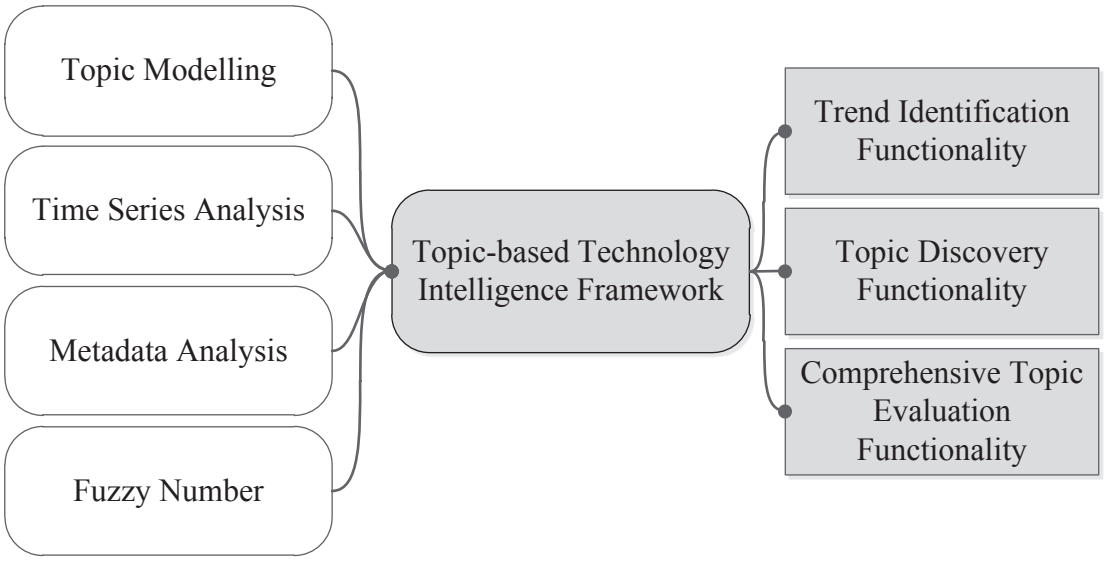


Figure 3-4. Brief introduction of topic-based technology intelligence

The three functionalities, presented with shaded rectangles in Figure 3-4, are designed to enrich the performance and overcome the limitations of tech mining-based technology intelligence. The proposed framework with the three components promises to make it possible to process textual data, publication count sequence and metadata of the target technology indicators, to present decision makers with a comprehensive awareness of technological landscape and advances, in their area of interest. To sum up, topic-based technology intelligence is an intelligent system which applies statistical topic modelling, time series analysis and metadata analysis techniques synthetically, to set decision makers, researchers and analysts free from reading, understanding and analysing massive technical documents and records, thus providing more efficient and effective decision support.

3.4.3 Conceptual Model of Topic-based Technology Intelligence

After giving a brief introduction of topic-based technology intelligence, this subsection presents and describes the conceptual model of the framework and its specific components. The whole framework of topic-based technology intelligence builds on the fundamentals of understanding, extracting and utilizing both semantic property and

temporal characteristic of historical observation of technology indicators (Chen, Zhang & Lu 2013).

As shown in Figure 3-5, to learn the current and historical development of target technologies, users of topic-based technology intelligence need to first initiate their technology area of interests as system input. Here users indicate researchers, analysts in universities, technology R&D managers of companies, technology planning officers in government sectors or other decision makers who need to gain technical insight from massive technical files. To gain insight of technologies that they are interested in, they all need to assess external technological developments to determine how they can gain from technology changes, avoid potential risk and plan their future R&D activities (Watts & Porter 2003). Their technology scope determination will then be transformed into one or several search statements for the WoS or USPTO under expert supervision. This will provide a large volume of scientific papers or patents that conform to the queries, such as IPC related to the target area, selected keywords, a specific subject category and so forth. Although experts will still participate in system procedures, their effort is confined to supervise the input and provide appropriate advice on search strategy.

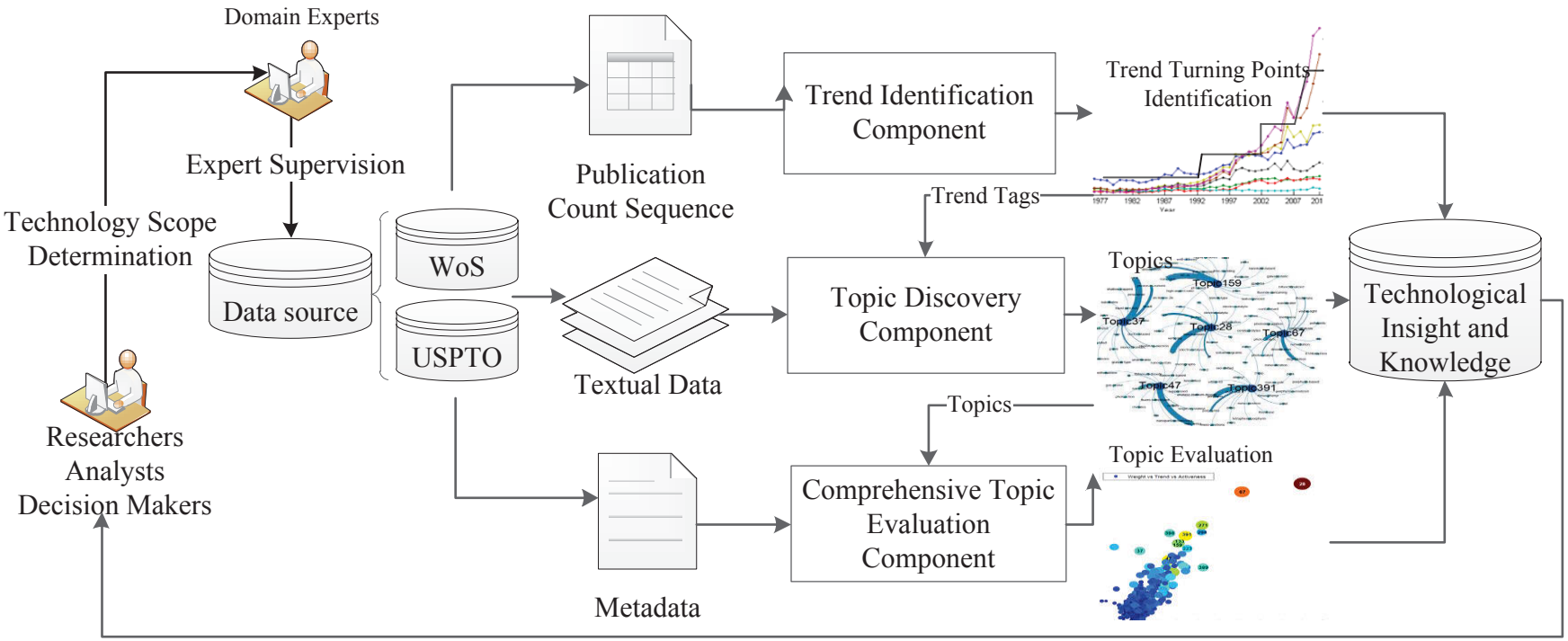


Figure 3-5. The Conceptual model of topic-based technology intelligence

In the next phase, publication count sequences, textual data and metadata are extracted separately from the selected scientific literature or patents, and the data passed to the three major functionalities, trend identification component, topic discovery component and comprehensive topic evaluation component. Trend identification component identifies turning points and trend segments. The trend tags are passed to the topic discovery component, which is in charge of detecting semantic topics. The temporal trend of each topic can then be estimated using the outcomes of the two components, thus quantitatively linking the semantic property and temporal characteristic in a real sense, without splitting the target corpus. Subsequently, all the discovered topics are passed to the component of comprehensive topic evaluation, where metadata will be connected with topics. This then provides the ability to identify the most representative topic set to explain a target corpus, and provide valuable evaluation on various themes. On the one hand, it provides a better and more comprehensive overview of a technological landscape; on the other hand, it can help researchers and analysts identify the most prominent topics and papers in their area of interest.

The outcome of the system integrates the results from text mining, time series analysis and metadata analysis. That is, the output includes identified trend turning points, trend segments and their corresponding text-based knowledge, topics, as well as, indices values of topic evaluation. Eventually, the output of topic trend detection, and how each topic contributes to publication activities, and also detailed evaluation on all themes will be collected, visualized and presented as technology insight and knowledge, to serve the needs of users of topic-based technology intelligence.

3.4.4 Overall Framework Description

The proposed overall framework of the topic-based technology intelligence is described in Figure 3-6. After a target technological area has been determined, search statements targeting analytic requirements are passed to WoS or USPTO. The corresponding corpus is then built, where all related web pages that belong to the scope are crawled to a corpus waiting for further analysis. The target scientific literature corpus consists of the title and abstracts of all the WoS papers within the target technological

scope, plus the subject category and metadata of those publications, including ISI unique article identifier, publication year, country, authors, affiliation and so forth. The target patent corpus, analogously, is composed of the title and claims of all the USPTO patents within the target technological scope, plus their metadata, such as IPC, USPC, inventor, issue year, assignee and so on.

Publication counts of each month, textual data and metadata are then extracted separately. The title and abstract for a paper, or the title and claims for a patent, constitute one textual document in a corpus, while the metadata of all documents comprises a single document, and publication counts are presented as a sequence of data points. The publication count sequence is first normalized for viewing convenience and then passed to a PLR step, where original observation is decomposed into a number of straight lines, strengthening and emphasizing the trend patterns underlying the patenting activities of the target area. Trend turning points and trend segments are generated in the next step, waiting for further processing. In Figure 3-6, all the detailed steps referring to publication counts processing, from step 5 to step 7, are marked in green.

The textual data, meanwhile, is passed to the Segmentation & Cleaning Step which removes all punctuation and meaningless symbols, in preparation for topic modelling. Stop words and common vocabularies used in target technical area are removed using a series of subsequent steps. The LDA then generates latent topics and a distribution of topics of the document collection, which is the observation of the model. To overcome the limitation of assuming topic number and determining the final topic set, the next step is to apply data likelihood and metadata, as shown in the figure, to decide which model fits the given data best. All the detailed steps related to textual data topic modelling, including steps 8 to 14, are marked in blue. The round-cornered outline indicates replication.

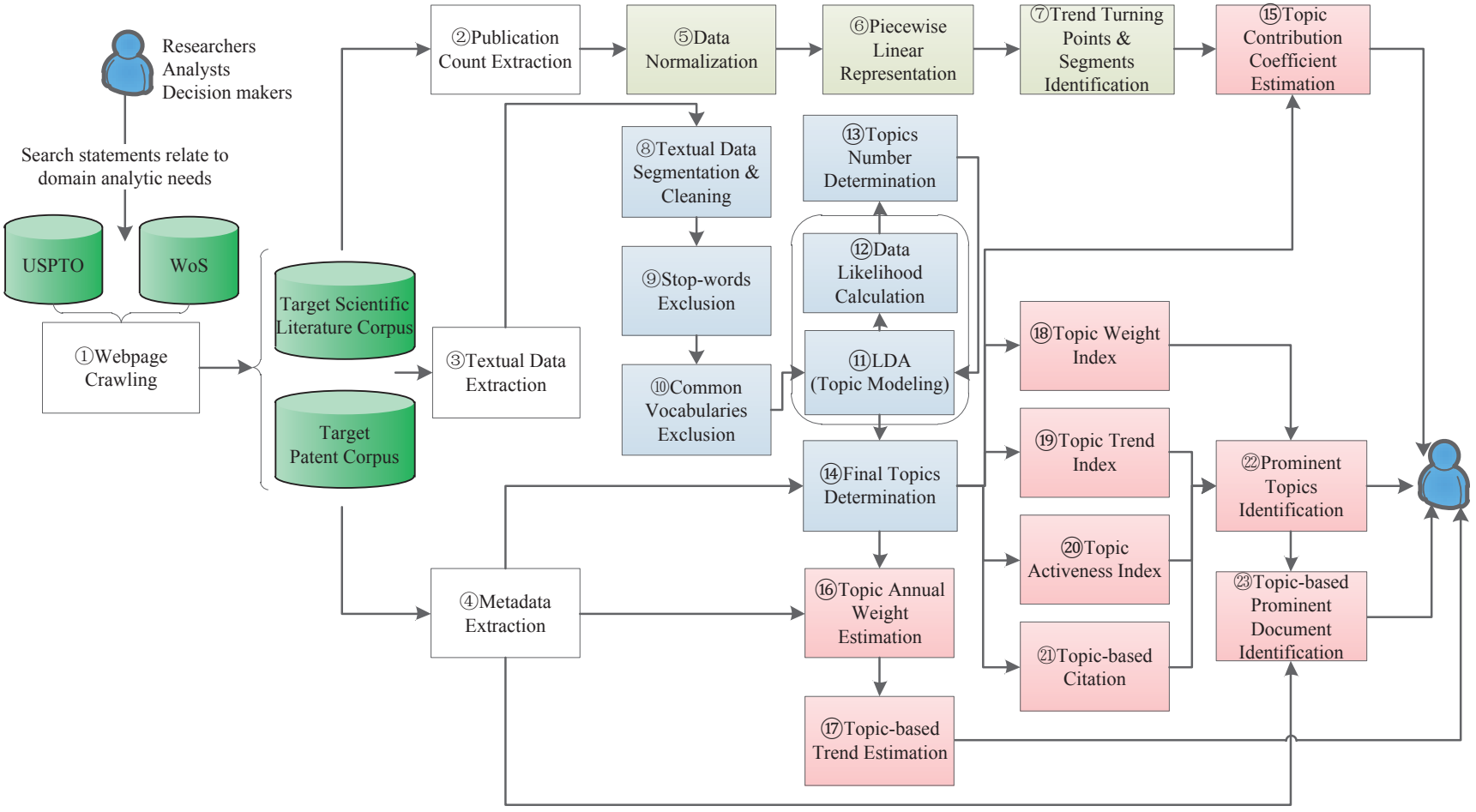


Figure 3-6. Framework of topic-based technology intelligence

Once the final topic set has been confirmed, the results of topic modelling are linked with metadata to compute the topic annual weight matrix. At the same time for each topic, trend turning points and trend segment are joined up with topics, where a contribution coefficient sequence is calculated to illustrate to what degree the topic has contributed to the publication activities of the whole area. The topic-based trend estimation by combining weight variation and contribution coefficient changes is then conducted. In addition, the weight, developing trends and activeness of all the topics are evaluated using the metadata and estimated topic distribution. Meanwhile, topic-based citations and their corresponding distributions are also calculated to quantitatively characterize the influence and contribution these estimated topics have made to the target area. Evaluation of the three indices and the topic-based citations provides the final number of topics and the most significant papers supporting them, to assist researchers and analysts in understanding the landscape and analysing the interesting themes of technological investigation in the area. All steps associated with comprehensive topic evaluation, from step 15 to step 23, are marked in red. Finally, the above outcome, include topic contribution coefficients, topic trend presentation, prominent topics and topic-based prominent documents will be eventually passed to the users of topic-based technology intelligence.

3.5 Description of the Framework Components

Users of technology intelligence expect to perceive from target technology indicators not only concepts and knowledge hidden in textual data, but also what is the corresponding technological trend of this knowledge and which topics play more important roles currently and in the future. In other words, the combination of trend in publication activities and themes hidden in corresponding documents, have the ability to provide users of topic-based technology intelligence with comprehensive awareness of technological advances in two different dimensions. This requirement needs to be satisfied on the fundamental of utilizing steps of text mining techniques, time series

analysis and even metadata analysis in technology intelligence. In this session, the main steps in three components of the topic-based technology intelligence will be introduced in details.

3.5.1 TREND IDENTIFICATION COMPONENT

There are two main purposes of the trend identification component: (1) to identify technology trend turning points and trend segments; (2) to provide the topic discovery component and the comprehensive topic evaluation component with temporal trend tags that are indicating changes of publication activities. In order to emphasize the function and steps of the component, here the two other components can be seen as black boxes. Figure 3-7 describes the detailed steps of the trend identification component.

After users of topic-based technology intelligence define the technology scope of their concern under expert supervision, all the documents that conform to the query statement are collected into a target corpus, where publication counts, textual data and metadata of the document collection are extracted separately. Specifically, the component of trend identification receives the raw data from the publication counts extraction step and passes it to the data normalization step. The normalized count sequence which has values between 0.0 and 1.0 is then transferred to outliers exclusion step to eliminate interferences, for identifying the main trend. The prepared sequence is then transferred to a PLR step, where the data is simplified and decomposed to several segments showing the trend movements. The output of the PRL step is then used to generate temporal trend turning points, trend segments and trend tags. The trend tags here are presented as a matrix, which contains all the important time points at which a trend shift occurred. Finally, the trend tags are passed to the components of topic discovery and comprehensive topic evaluation, for further calculation. The estimated trend turning points and trend segments are stored in a technology intelligence knowledgebase and delivered to researchers, analysts or decision makers who use the technology intelligence framework.

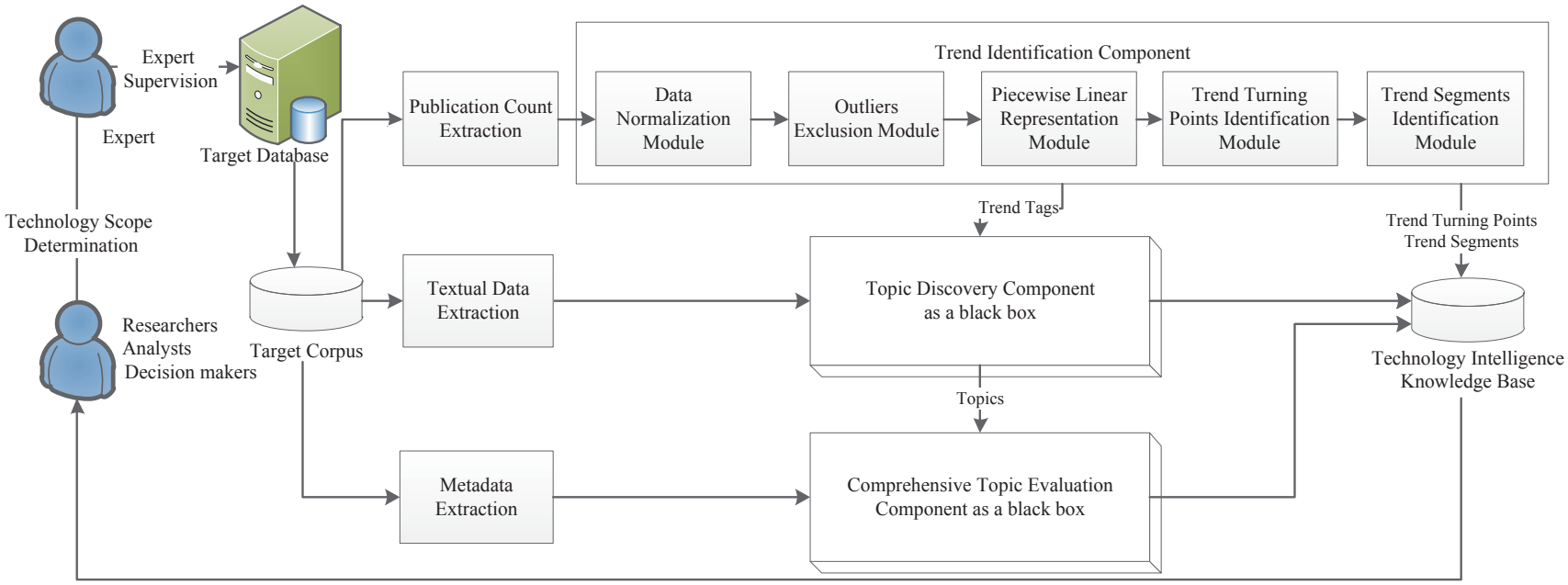


Figure 3-7. Details of the trend identification component

3.5.2 TOPIC DISCOVERY COMPONENT

As mentioned, much effort has already been devoted to identifying valuable terms and phrases from scientific publications and patent documents. Nevertheless, the outcomes of text mining-based techniques applied to assist topic extraction are mostly keywords with a ranking. These words alone, however, are usually too general or misleading to indicate a concept, especially when there are polysemous words actually describing different topics (Tseng, Lin & Lin 2007). Thus in the past five years, topic model-based approaches that provide more managerially utilizable extraction and informative representations of technological concepts have attracted increasing research interest. However, to detect a topic set that best explain the observation, there are two main limitations that LDA implementation suffers from, which are topic number pre-setting and random results brought by sampling.

To address the above issues, other than presenting concepts and themes with keywords, a topic discovery component has been constructed in the topic-based technology intelligence framework, to discover and characterize the semantic knowledge in target technology indicators with topics. As shown in Figure 3-8, after textual data extraction, segmentation and cleaning, all target documents are processed with a series of words exclusion steps to filter out stop words, high frequency words that commonly appeared in scientific literature or patent claims, and common vocabulary used in target technical area. Then, the prepared texts are passed to the topic modelling step. To overcome the limitation of assuming topic number and determining the final topic set, data likelihood and metadata are then applied, as shown in the Figure 3-8, to decide which trial fits the given data best. The round-cornered outline indicates replication. After the final topic set is determined, the contribution coefficients for each topic are then able to be computed. Finally, all the discovered topics are passed to the components of comprehensive topic evaluation, for further calculation. In addition, these topics and their estimated contribution coefficients are stored in a technology intelligence knowledgebase and delivered to users of topic-based technology intelligence.

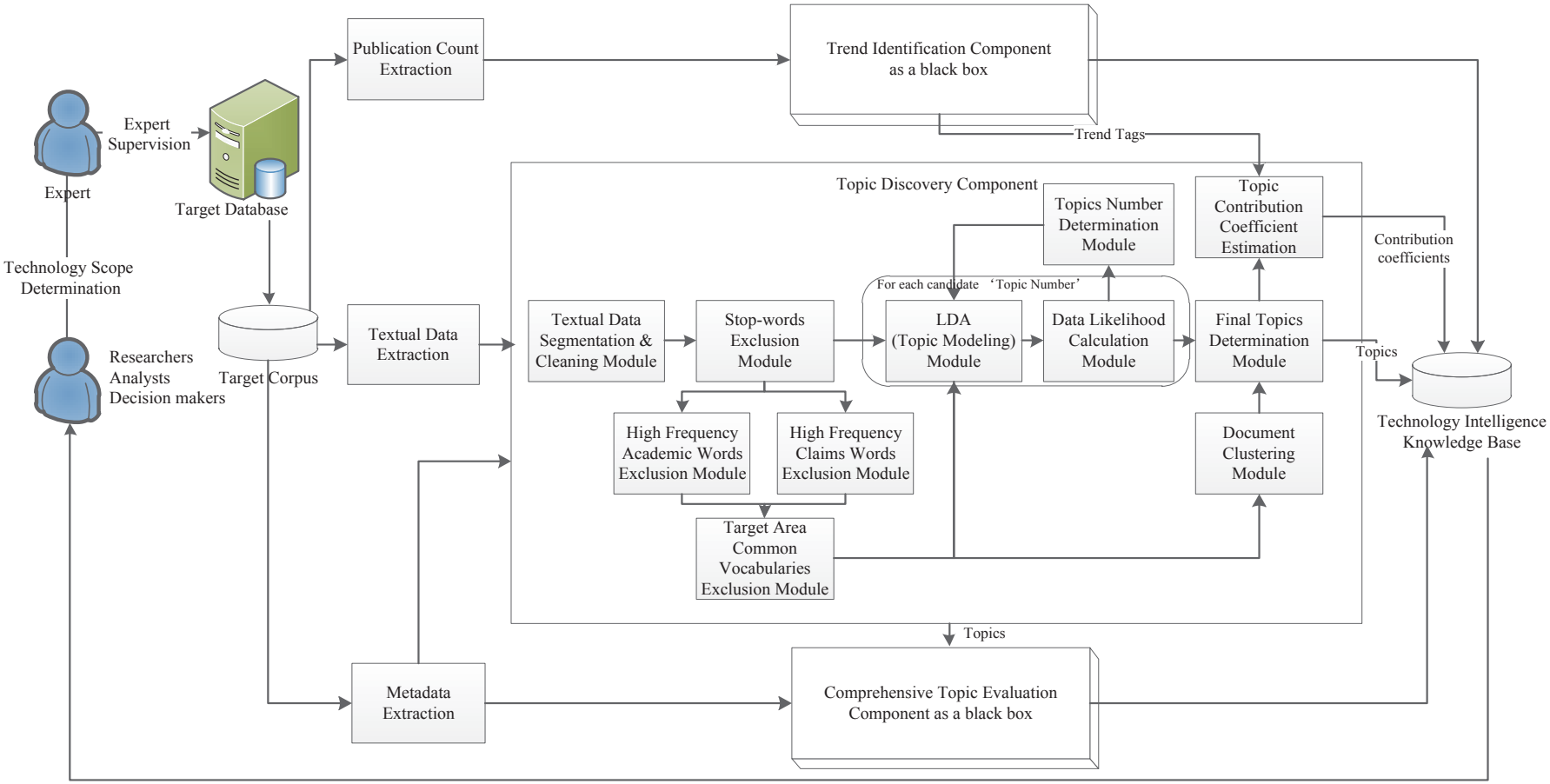


Figure 3-8. Details of the topic discovery component

The outcome of the above two components provides researchers, analysts or decision makers with patent knowledge that indicates: (1) what the main trend in the technological area of interest is; (2) where the states turning points of the technology developing trend are; (3) what are the main topics of the document collection; and (4) how do these text-based knowledge evolve from one trend segment to another. This outcome will help users of topic-based technology intelligence to gain a better awareness of technology development in target area over time and will provide a prospect of the future trend state in their technical areas of interest.

3.5.3 Comprehensive Topic Evaluation Component

In real content, to achieve the aim of understanding valuable thematic knowledge from massive technological documents, and improve the approaches for doing so, there are two main phases which need to be considered. The first is automatically detecting latent knowledge from textual data. The second is assisting further thematic evaluation using discovered topics or emblematic keywords.

As shown in Figure 3-9, this study proposes a comprehensive topic evaluation component to comprehensively consider metadata of technology indicators and the result of the other two components. Once the final topic set has been confirmed, trend tags and discovered topic are applied to construct an annual weight matrix and then quantitatively measure the topic-based trend, also the most contributive trend segment for each prominent topic. In addition, the weight, developing trends and activeness of all the topics are evaluated using the metadata and estimated topic distribution. Meanwhile, topic-based citations and their corresponding distribution are also calculated to quantitatively characterize the influence and contribution these estimated topics have made to the target area. Evaluation of the three indices and the topic-based citations provides the final number of topics and the most significant documents supporting them, to assist researchers, analysts or decision makers in understanding the landscape and analysing the interesting themes.

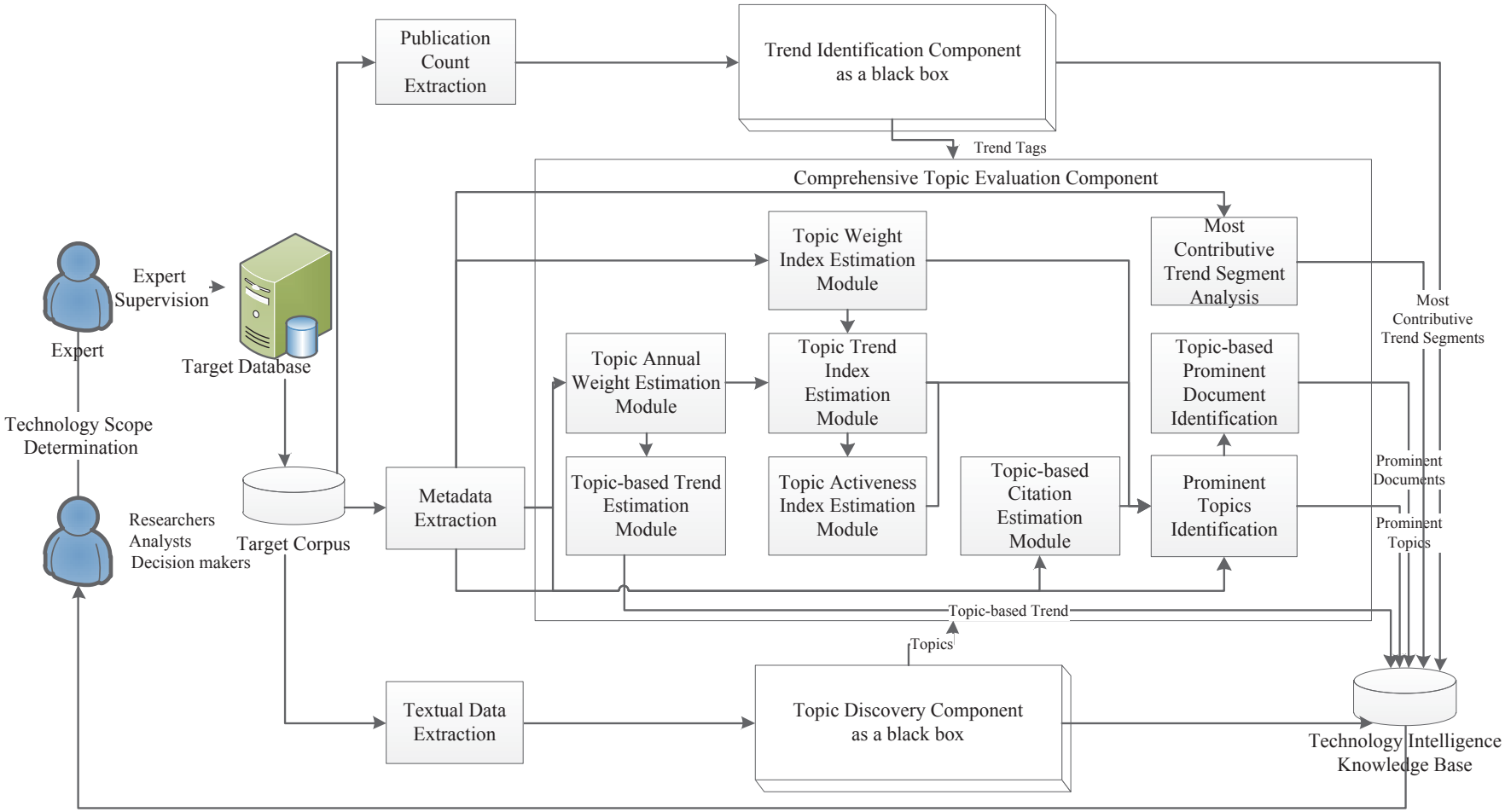


Figure 3-9. Details of the comprehensive topic evaluation component

Finally, the result of topic-based trend estimation, a group of prominent topics, the most contributive trend segment and topic-based prominent document for each of them, are passed to the technology intelligence knowledgebase and delivered to users of topic-based technology intelligence. The aim of this component is to link metadata with the estimated topics and trend segments to provide valuable evaluation on various themes. The three indices presented in this component will quantitatively characterize the weight, developing trend and activeness of all discovered topics after executing the LDA. This research can provide a better and more comprehensive overview of the technological landscape, and at the same time help users of topic-based technology intelligence to identify the most prominent topics and papers in their area of interest.

3.6 SUMMARY

The theoretical and applied research on technology intelligence will continue to be emphasized to assist decision makers learn technical knowledge in massive data effectively and efficiently. In this chapter, by first recognising and measuring the semantic property and temporal characteristic of typical technology indicators, a framework of Topic-based Technology Intelligence, with trend identification functionality, topic discovery functionality and comprehensive topic evaluation functionality, has been designed and constructed to improve the framework and applications of existing technology intelligence.

The proposed topic-based technology intelligence can be used to systematically extract technological knowledge underlying in large volumes of scientific literature, patents or other semi-structured technology indicators, with much less human intervention. Hereinafter, this thesis will define and explain detailed methods used in this framework, and provide case studies in each chapter to demonstrate these methods, using both scientific literature and patent data.

CHAPTER 4
EMPIRICAL TECHNOLOGY TREND ANALYSIS METHOD

4.1 INTRODUCTION

Technological trend analysis is one of the most concerned research topics in the context of R&D management, in both public and private domains. However, existing model-driven empirical technological trend analysis usually suffers from model choosing and limitations that a certain model may bring. While designing and implementing technology intelligence frameworks, it is quite difficult to reveal all the trend patterns by only restrained models. Under conditions of rapid technologies development, budgetary constraint and time limitation, an increasing number of demands have created new requirements for the technological trend analysis process especially in technology intelligence research, that is, to be more effective and less time consuming.

From a temporal perspective, to capture detailed trend pattern of publication activities and proceed with future trend estimation, at the same time overcome the limitations of using only restrained models, this chapter proposes a data-driven empirical technological trend analysis (TTA) method, which mainly explains the trend identification functionality of this proposed technology intelligence framework. As mentioned in Chapter 3, publication count sequences serve as the main input of technological trend analysis research, since they are the observation of historical publication activities of technology indicators.

TTA method is based on the concept of piecewise approximation, which is brought into technology trend analysis heuristically by Philips (1999), in his work of using a piecewise linear regression method to capture trend changes of the polyvinyl chloride price. This sub-section improves the usage of the concept of piecewise approximation by identifying and depicting the movements of observed publication records with a number of trend turning points and trend segments, to strengthen and highlight trend changes and possible sudden shifts. These segments are transformed into a stage-wise signal to quantitatively illustrate trend movements of a target technology. Trend movement intensity and future trend are then estimated based on identified trend turning points and trend segments.

To better explain the superiority of this proposed approach in capturing detailed shifts and forecasting further trend, this chapter compares it with one of the most popular and accepted empirical technology trend forecasting approaches, the growth curves. Specifically, the logistic curves model and the Gompertz curves model are selected to conduct the comparison. Experimental comparisons show that the proposed approach has a better performance when depicting and forecasting the detailed trend movements and shifts. In addition, it is more effective when forecasting short-term tendency without specific model selection or upper limits estimation. Two case studies, using IP Australia patents and USPTO patents respectively, are given to demonstrate the performance of this proposed approach.

The remainder of the chapter is organised as follows. Section 4.2 describes data preparation for the proposed TTA method. The stepwise explanation of the method full process is presented subsequently in Section 4.3, followed by Section 4.4, which describes a case study using IP Australia patents to conduct an examination of the approach and explain how to use it in a real technological trend analysis context. In order to compare the proposed TTA method with model-driven empirical technological trend analysis approaches, Section 4.5 continues to provide a case study using USPTO patents and to discuss the output of the proposed approach and growth curves models. Finally, a summary of this chapter is given in Section 4.6.

4.2 DATA PREPARATION FOR TTA METHOD

Before conducting technological trend analysis, trend patterns of the target technology indicators need to be quantitatively represented first. As described in Chapter 3, either scientific literature or patents within a particular industry, the quantity of publications under a certain search statement during a period of time, can be presented as a publication count sequence, $P = \{p_1, p_2, \dots, p_i, \dots, p_r\}$, where p_i represent the counts of the i^{th} time intervals (months, seasons or years) and r indicates the total number of these time intervals.

4.2.1 OUTSIDER EXCLUSION

Since the range of raw data of publication counts varies widely, this research first normalizes the publication count sequence $P = \{p_1, p_2, \dots, p_i, \dots, p_r\}$ using feature scaling. P is first normalized into $\bar{P} = \{\bar{p}_1, \bar{p}_2, \dots, \bar{p}_i, \dots, \bar{p}_r\}$ where $\bar{P}$ has a value between 0.0 and 1.0, by

$$\bar{P} = (P_{max} - p_i)/(P_{max} - P_{min}). \quad (4\text{-}1)$$

The normalized publication count sequence $\tilde{P}$ is then transferred to an outliers exclusion step to eliminate interference of identifying the main trend. If the data is normally distributed, then this research utilizes the three sigma rule on the difference of normalized data and its polynomial fit to check if there is any outlier; the algorithm of publication count sequence outsider exclusion is presented in Algorithm 4-1. After outliers removing, the prepared literature/patent count sequence$\tilde{P} = \{\tilde{p}_1, \tilde{p}_2, \dots, \tilde{p}_i, \dots, \tilde{p}_r\}$ is obtained. If the data does not follow a normal distribution, then the outlier exclusion step can be skipped.

Algorithm 4-1. Publication count sequence outsider exclusion algorithm

Input: the normalized patent count sequence $\bar{P} = \{\bar{p}_1, \bar{p}_2, \dots, \bar{p}_i, \dots, \bar{p}_r\}$ (if normally distributed), time T

Output: the prepared patent count sequence without outliers, $\tilde{P} = \{\tilde{p}_1, \tilde{p}_2, \dots, \tilde{p}_i, \dots, \tilde{p}_r\}$

1 **polyfit** $(T, P, 2) = P_1$

2 $\mathbf{abs}(P_1 - P) = P_2$

// calculate the absolute difference between original values and the Polynomial fit

3 $\mathbf{std}(P_2) = sigma$

//calculate the standard deviation of the difference

4 $\mathbf{abs}(P_2 - mean(P_2)) > 3 * sigma = outliers$

// use the three sigma rule to identify the outliers

5 **find** (outliers) = position

//find the positions of outliers

6 $P(position) = (P(position - 1) + P(position + 1))/2$

//use the mean value of the two data points near the outlier to replace it

7 $\tilde{P} = P$

8 **end**

4.2.2 Parameter Setup for Piecewise Linear Representation

In order to learn valuable trend turning points in historical trend observation and represent variations in trend states over time, this research extracts temporal trend patterns based on the approach of piecewise linear representation (PLR)(Keogh et al. 2001; Keogh & Kasetty 2003; Keogh et al. 2004). In this case, the PLR result presents several trend segments of the corresponding patent counts record. The more obvious the patterns obtained from the data, the easier the understanding of trend changes will be.

Here, segment number s is a threshold of the PLR. It is important to set the value suitably, since it directly affects the sensitivity of the trend pattern extraction. A comparatively smaller value of segment number produces larger trend segments that present the trend more explicitly, despite slight fluctuations; conversely, a larger value of segment number makes it more sensitive when trend segments are determined. In the research of stock trading points prediction using PLR, genetic algorithm is often used to select the model's parameters. Several threshold values for PLR are set and the one that generates the most profit will ultimately be chosen (Chang, Fan & Liu 2009; Chang et al. 2011). However, in the context of technological publication activities trend identification,

genetic algorithm is not suitable since this research does not have the "evaluation criteria", like stock benefit, to evaluate patenting or scientific literature publishing records.

In a real case, from the prospective of observing detailed trend shifts, a smaller value of s is preferred. However, a small threshold will provide a quite large residual sum of squares (RSS) value between $\tilde{P}_{PLR}$ and $\tilde{P}$, which means the PLR model is less representative. Experiments show that the discrete data of $\tilde{P}_{PLR}$ RSS is gradually declining while s is rising, which can be fitted to an exponential curve. To maintain the balance of producing the number of segments as small as possible at the same time keeping a comparatively lower RSS of $\tilde{P}_{PLR}$, a cost function is built as shown in Formula 4-2:

$$f(normalized\ s, normalized\ RSS) = w_1 \times normalized\ s + w_2 \times normalized\ RSS, \quad (4\text{-}2)$$

where w_1 and w_2 are two weight coefficients adjusting the importance of $normalized\ s$ and $normalized\ RSS$. It is assumed that getting a smaller value for m is equally important as keeping a lower RSS; thus here $w_1 = w_2 = 1$is set. When the cost function intersects with the fitted curve, it is expected to have the minimum value. That means, the value of s at the point where the cost function and the RSS fitted curve has a point of tangency, is the preferable one.

Figure 4-1 provides an example of PLR threshold setting, showing that the RSS values of $\tilde{P}_{PLR}$ with different threshold s are fitted to an exponential function. At the very beginning, the decline speed of RSS values is much faster than the growth of threshold s; then it gradually decreases. The green solid line is the cost function which has a point of tangency with the fitted curve. As shown in the figure, the point of contact appears when the value of s is 22, thus here in the example, the selected threshold for PLR model is 22 (for readers' convenience the x-axis is presented as segments number instead of normalized m values).

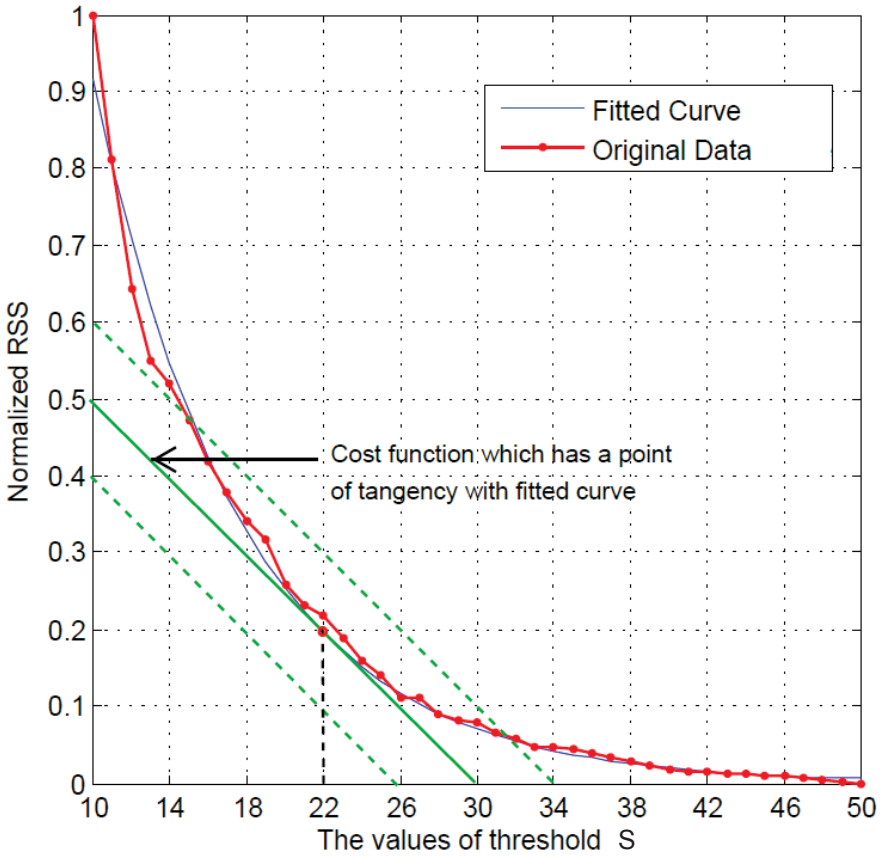


Figure 4-1. An example of PLR threshold setup

In other words, experiments show that the discrete data of the RSS value between $\tilde{P}_{PLR}$ and $\tilde{P}$ is gradually declining, while s is rising, fast to slow. To balance the explicitness of trend shifts and the representability of the model, s is selected where the declining rate of RSS starts to obviously slow down, as the preferable one. After solving Formula 4-2, it can be noticed that the approximate derivative (AD) of a series of RSS produced by their corresponding s, is a more simple way to calculate for threshold determination, as show in Formula 4-3,

$$AD_{s_{preferable}} = max \left| \frac{\Delta RSS}{\Delta s} \right|, \tag{4-3}$$

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

4.3 TTA METHOD AND AFTERWARDS TREND FORECASTING

As discussed in Chapter 3, to further link the semantic property and temporal characteristic in technology intelligence research, instead of using fitting curves, there is a need to identify noticeable trend turning points to correlate semantic extraction with a specific time interval. Facing the limitations of applying model-based approaches, the proposed TTA method uses 'trend segments' and 'trend turning points' to define the temporal characteristic. Here, the trend turning points and trend segment indicate the occurrence time of significant changes of the publication activities and how long the changes last.

4.3.1 TREND TURNING POINTS IDENTIFICATION

After the parameter of s is determined, the prepared technological publication count sequence $\tilde{P}$ is decomposed by PLR into s segments, as shown in Formula 4-4,

$$\tilde{P}_{PLR} = \left\{ \begin{array}{c} L_1(\tilde{p}_1, \tilde{p}_2, \dots, \tilde{p}_{t_1}), L_2(\tilde{p}_{t_1+1}, \tilde{p}_{t_1+2}, \dots, \tilde{p}_{t_2}), \dots, \\ L_i(\tilde{p}_{t_{i-1}+1}, \tilde{p}_{t_{i-1}+2}, \dots, \tilde{p}_{t_i}), \dots, L_s(\tilde{p}_{t_{s-1}+1}, \tilde{p}_{t_{s-1}+2}, \dots, \tilde{p}_r) \end{array} \right\}, \quad (4\text{-}4)$$

where $\tilde{P}_{PLR}$ denotes the combination of s segments and $L_i(\tilde{p}_{t_{i-1}+1}, \tilde{p}_{t_{i-1}+2}, \dots, \tilde{p}_{t_i})$ indicates the i^{th} $(1 < i < s)$ segment of $\tilde{P}_{PLR}$ (Keogh et al. 2001; Keogh & Kasetty 2003; Keogh et al. 2004). In the same way, $\tilde{P}_{PLR}$ 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. The calculation of trend turning points is presented as a matrix in Formula 4-5,

$$TP = \begin{bmatrix} 1, & t_1 \\ t_1 + 1, & t_2 \\ \vdots & \vdots \\ t_{i-1} + 1, & t_i \\ \\ \vdots & \vdots \\ t_{s-1} + 1, & r \end{bmatrix}, \quad (4\text{-}5)$$

where each row of matrix TP indicates a start and an end of a trend state (Chen, Zhang, Zhu, et al. 2015).

4.3.2 TREND SEGMENTS IDENTIFICATION

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. Then $\tilde{P}_{PLR}$ is converted into corresponding trend segments $TS = \{ts_1, ts_2, \dots, ts_s\}$, 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,

$$TS_i = \left(ts_{t_{i-1}+1}, ts_{t_{i-1}+2}, \dots, ts_{t_i}\right), \tag{4-5}$$

$$ts_{t_{i-1}+1} = ts_{t_{i-1}+2} = \cdots = ts_{t_i} = mean\ L_i\left(\tilde{p}_{t_{i-1}+1}, \tilde{p}_{t_{i-1}+2}, \dots, \tilde{p}_{t_i}\right), \tag{4-6}$$

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 Neural Networks i^{th} segment of $\tilde{P}_{PLR}$, $L_i\left(\tilde{p}_{t_{i-1}+1}, \tilde{p}_{t_{i-1}+2}, \dots, \tilde{p}_{t_i}\right)$. Figure 4-2 explains the process of transforming original data to trend segments step by step. The transformation between $\tilde{P}_{PLR}$ 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.

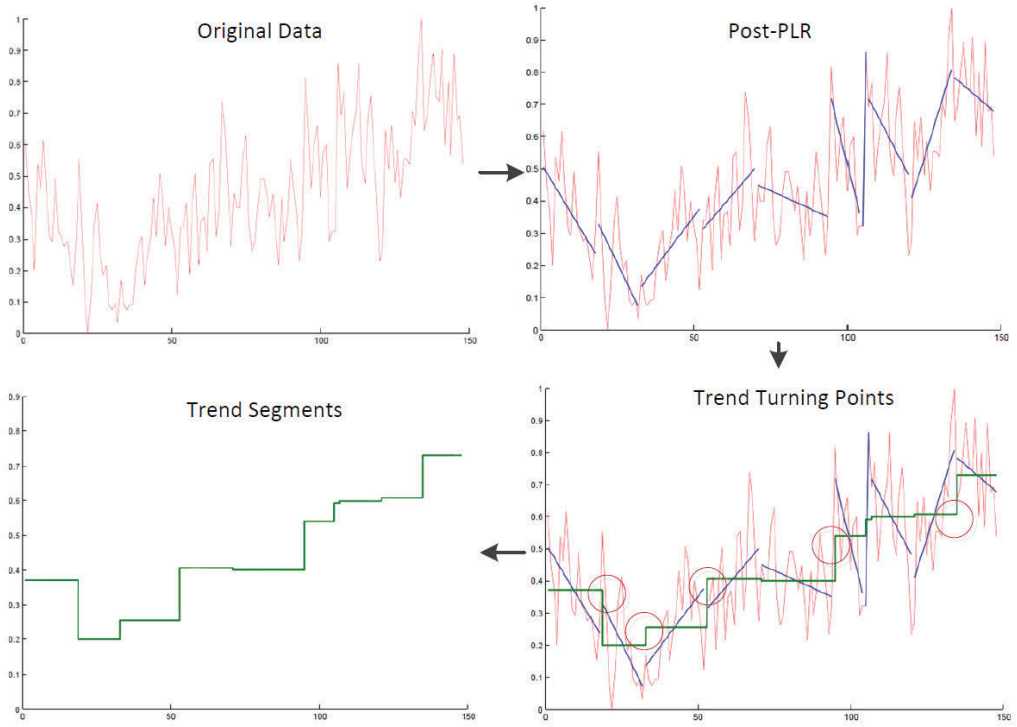


Figure 4-2. An example of transforming original data to trend segments step by step

4.3.3 TREND MOVEMENT INTENSITY

After the trend segments has been generated, a group of intensity coefficients indicating different levels of trend changing from segment to segment are calculated with the Formula 4-7,

$$intensity\ i_k = \frac{mean\{L_k\} - mean\{L_{k-1}\}}{(t_{k+1} - t_k + 1)/n}, \tag{4-7}$$

where i_k indicates the intensity coefficient of the k^{th} segment, L_k and L_{k-1} stand for the k^{th} and $(k-1)^{th}$ pieces of $\tilde{P}_{PLR}$; in the denominator, ($t_{k+1} - t_k + 1$) is the time that the k^{th} segment lasts, $(t_{k+1} - t_k + 1)/n$ presents the proportion the k^{th} segment takes from the whole time length. Since there is no trend changing for the first segment, i_1 is set as 0. The larger the intensity coefficient is, the stronger changing a segment has compared to its previous one.

4.3.4 TTA-BASED TREND FORECASTING

After quantitatively identifying the temporal trend patterns from technological publication count sequences, further trend forecasting can proceed by using the proposed TTA method. This research applies the future trend signal with the nonlinear autoregressive neural networks (NARNNs) model to estimate the future developing trend of the target technologies.

Neural Networks have been proved that can provide preferable performance over time series forecasting by many researchers. A NARNN is a recurrent dynamic network that can be trained to predict future values for univariate data series forecasting. It estimates the future values based on its historical output by utilizing a neural network nonlinear structure to estimate the model's parameters (Safavieh, Andalib & Andalib 2007). A NARNN model has endogenous inputs only; input values are taken from the memorization of earlier sequences, that is, the output is fed back to the input of the neural network. As a recurrent dynamic network, it can be mathematically represented as follows:

$$y(t) = f(y(t-1), y(t-2), \dots, y(t-d)), \tag{4-8}$$

where f is the nonlinear function, and $y(t-1), y(t-2), \dots, y(t-d)$ are the input of a number of multilayer perceptrons which are also the feedback of previous outputs, and d indicates the number of delays (Mandal & Prabaharan 2006). Thanks to the ability of NNs to perform arbitrary nonlinear mapping of input-output patterns, they are suitable for the prediction of data-rich and theory-poor domains (Florita & Henze 2009). This research thus utilizes NARNNs to predict the future trend signal of technologies which are of interest.

A neural network is trained by the historical trend signal and utilized to predict the future trend movements. While forecasting, the output of the neural network is fed back to the input, future values of the trend signal can be predicted from past values. The recurrent dynamic network can be mathematically represented as follows:

$$ts_{s+1} = f(ts_1, ts_2, \dots, ts_s)\ , \tag{4-9}$$

where f is the nonlinear function, and $ts_1, ts_2, \dots, ts_s$ are the input of a number of multilayer perceptrons, and d indicates the number of delays. Figure 4-3 shows the detailed process of the PLR-based technology trend forecasting approach using patent counts record in summary.

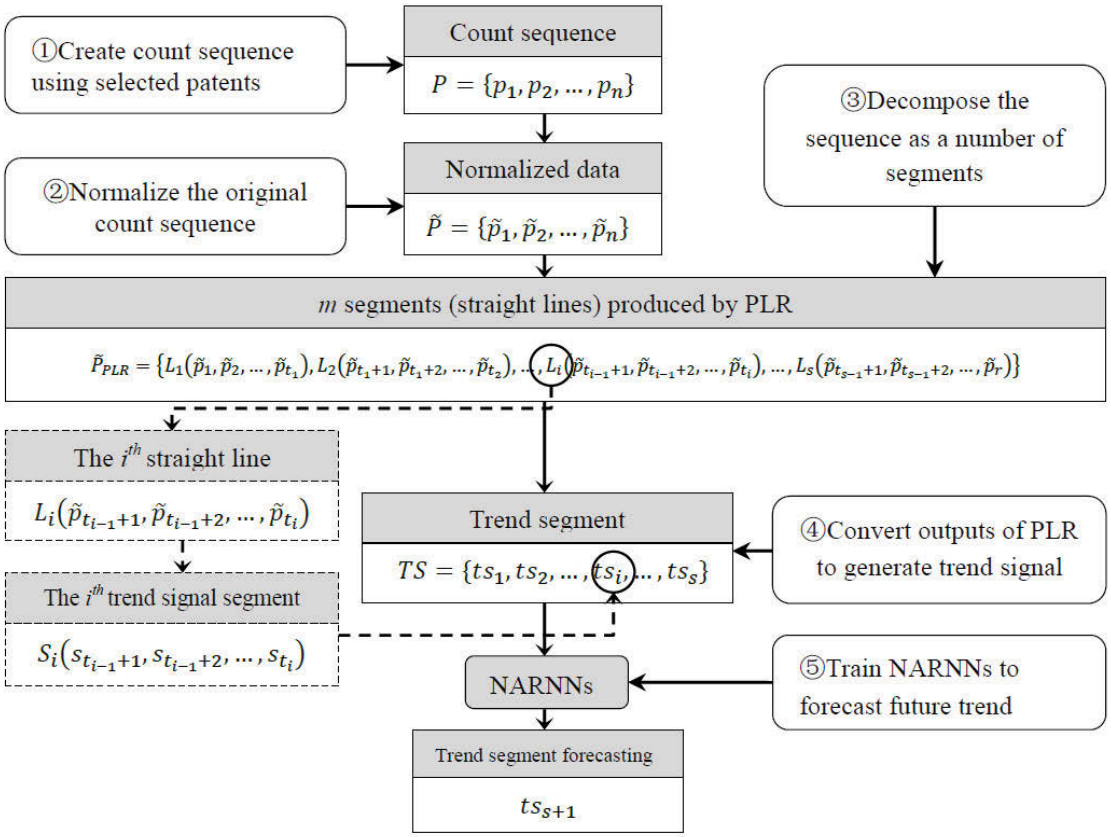


Figure 4-3. The detailed process of the TTA-based trend forecasting

4.4 Case Study 1: Patent Data in IP Australia

In this sub-section, the proposed empirical TTA method is applied in a real patent analysis context. Patents from the Australian government intellectual property department patent database (IP Australia) are collected and utilized to demonstrate the

validity of trend identification functionality, when dealing with real-world tasks. The result of the experiment shows that the new component learns valuable trend turning points in historical patent count sequence.

4.4.1 DATA SETS

The data used in this research comes from IP Australia (AusPat). The industry background of the patent data chosen is Information and Communications Technologies (ICT). ICT-related technologies have attracted attention increasingly in the area of industrial globalization for their rapid growth in recent years (OECD 2005; Lee, Kim & Park 2009). According to the Organization for Economic Cooperation and Development (OECD), how to seize benefits and opportunities of ICT for economic growth and development has become an important concern to OECD governments (Publishing 2010), including Australia.

The data in the case study is collected based on the search statement for patents with IPC indicating ICT technologies, published by OECD (OECD 2008), which splits the ICT sector into telecommunications, consumer electronics, computers and office machinery and other ICT. The time interval unit here is set as month. This case study collects the quantity of issued patents in every month during 1983-2012 to create a literature/patent count sequence, which makes 360 months in total. That is, n in $P_{case1} = \{p_1, p_2, \dots, p_n\}$ equals to 360, p_i shows the number of issued patents in each corresponding month. The raw data used in this case study follow a normal distribution.

4.4.2 OUTLIERS EXCLUSION

After data normalization, the original patent publication count sequence is converted to a new one without the outliers. As shown in Figure 4-4, the ICT patent count sequence was fitted to a quadratic polynomial. By using the three sigma rule, this case study locates the position of outliers showing as red points, and replaces them with the mean value of the data points in the front and at the back of each outlier for trend maintaining. The detailed values of the outliers and replacements are shown in Table 4-1.

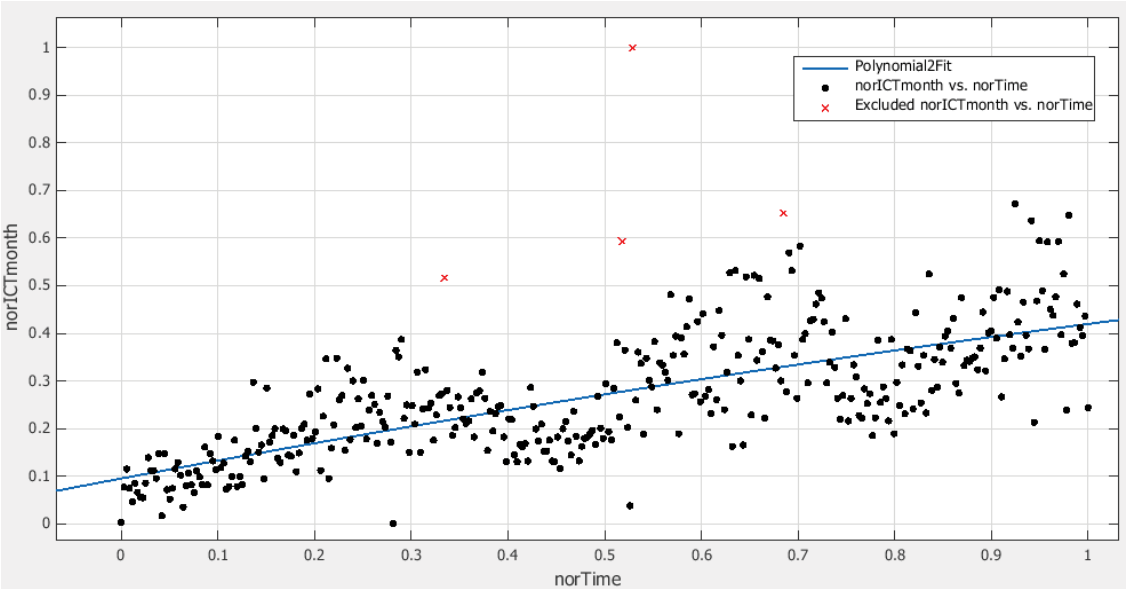


Figure 4-4. The outliers exclusion for ICT patent count sequence

Table 4-1. The detailed outliers' values of ICT patent count sequence

Outliers position	Normalized time	Value of outliers	Value of outliers before normalization	Replacement value
121	0.334262	0.516129	395	0.276999
187	0.518106	0.593268	450	0.29453
191	0.529248	1	740	0.148668
247	0.685237	0.652174	492	0.28892

4.4.3 Trend States and Trend Turning Points Identification

After excluding the outliers, the prepared patent count sequence is processed and decomposed by PLR step. In this case study, $s = 9$ was chosen as PLR threshold, as it maintains the relative balance of least segments and lowest RSS (RSS will reduce while s will rise. At the same time, the mean value of each segment is used to generate a new series showing trend states. As shown in Figure 4-5, the original data is presented with a blue line and represented as nine straight red lines by PLR to retain the main tendency. The final trend states are illustrated by the green line in the figure. It can be observed that

the trend changing points are July of 1991, January 1998, March 1999, November 2011, September 2003, May 2006, September 2010 and May 2011.

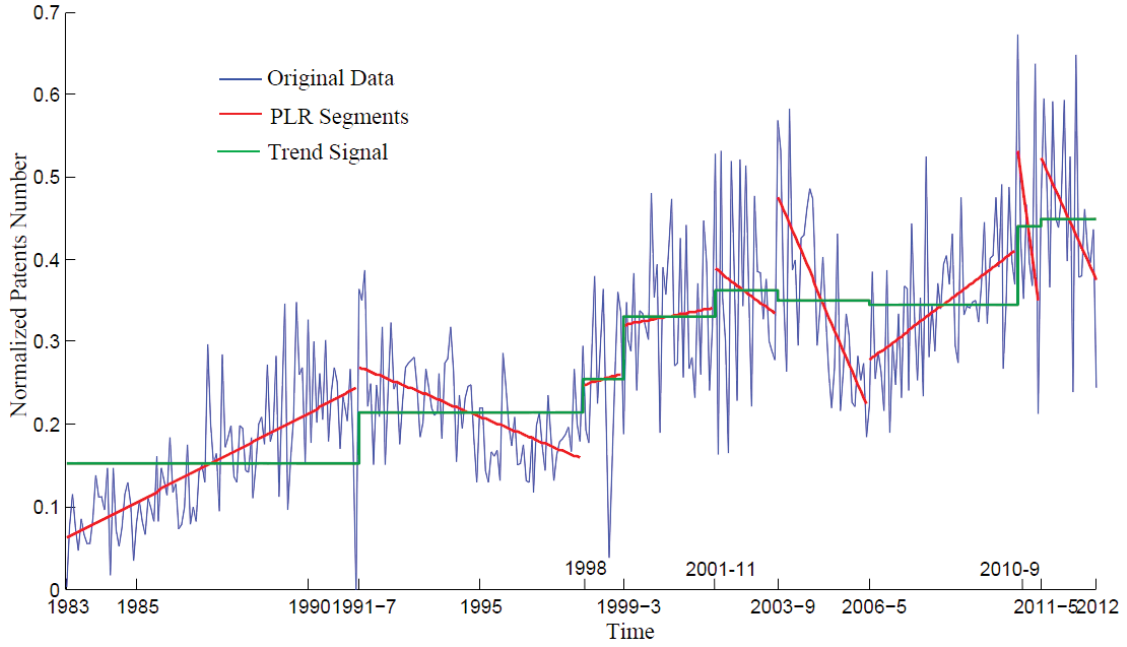


Figure 4-5. Original data, PLR segmentation and trend segments of technologies in ICT industry

On the whole, the ICT development in Australia experienced fluctuant trend rise during the past 30 years. The growth trend descended slightly twice between years 1991-1998 and years 2003-2010. In the recent three years, the development of ICT is fluctuating and descending from 2011-2012, yet the main trend is escalating compared with previous years. The detailed trend segment values and trend tags are showing in Table 4-2. The tags will be provided to topic-based technological forecasting method in Chapter 5, to assist future semantic analysis on target technologies. The corresponding text-based knowledge variation of the trend changing can be identified as well.

Table 4-2. Trend signal value and trend tags of ICT technologies in Australia

Number of segments	Tag start	Tag end	Trend Tags	Trend signal value
1	1	102	JAN 1983 to JUN 1991	0.153081429
2	103	180	JUL 1991 to DEC 1997	0.214253605
3	181	194	JAN 1998 to FEB 1999	0.254157483
4	195	226	MAR 1999 to OCT 2001	0.330382188
5	227	248	NOV 2001 to AUG 2003	0.361596328
6	249	280	SEP 2003 to APR 2006	0.350017532
7	281	332	MAY 2006 to AUG2010	0.344265832
8	333	340	SEP 2010 to APR 2011	0.440743338
9	341	360	MAY 2011 to DEC 2012	0.449228612

4.5 Case Study 2: Patent Data in USPTO

In this sub-section, the generalization of the prosed TTA method and its afterwards trend forecasting validity are continued to be demonstrated by USPTO patents. Specifically, patents of three industries that have drawn wide attention, Telecommunication, Solar Cell and Radar, are collected from USPTO online database to conduct empirical trend analysis and forecasting. In addition, one of the most studied and accepted methods in technology trend forecasting area, the growth curves model, specifically applied the Pearl Curve and the Gompertz Curve model to the same data set, is chosen as a comparison to the proposed method.

The results of experiments demonstrate that the proposed approach learns valuable trend patterns and shifts in historical data and predicts future technology trends reasonably well. Compared with the growth curves model, it has a better performance with capturing and depicting the detailed movements of technology trends, and at the same time, delivers comparatively more accurate future trend estimation than a growth curve does. All the case studies are implemented and run in MATLAB.

4.5.1 Data Collection and Parameter Setting

Three search statements, as shown in Table 4-3, are utilized to identify the patents in the fields of telecommunication, solar cell and radar. Following each search requirement, the number of patents published monthly is collected to create corresponding patent

counts sequences. Specifically, the data employed in the three case studies are respectively extracted from patents in USPC 455, which stand for telecommunication technology; patents with the keywords 'Solar Cell' in their abstracts, and patents with the keyword 'Radar' in their abstracts which also belong to USPC 342 (Communications: Directive Radio Wave Systems and Devices) (USPTO 2012b). The detailed information of patent counts collection is shown in Table 4-3 as well. For each case, the number of patents published in every month from 1976 to 2012 is collected. Figure 4-6 presents the original and cumulative patent counts after normalization.

Table 4-3. Description of the data sets for case study 2

NO.	Search Statement	Description	Month NO.	Total Counts
1	CCL/455/$	Patents in the Telecommunication technology Class	444	88382
2	ABST/"solar cell "	Patents with "solar cell" in their abstract	444	2699
3	ABST/"radar"AND CCL/342/$	Patents with "radar" in their abstract and at the same time in the Communications: Directive Radio Wave Systems and Devices Class	444	5423

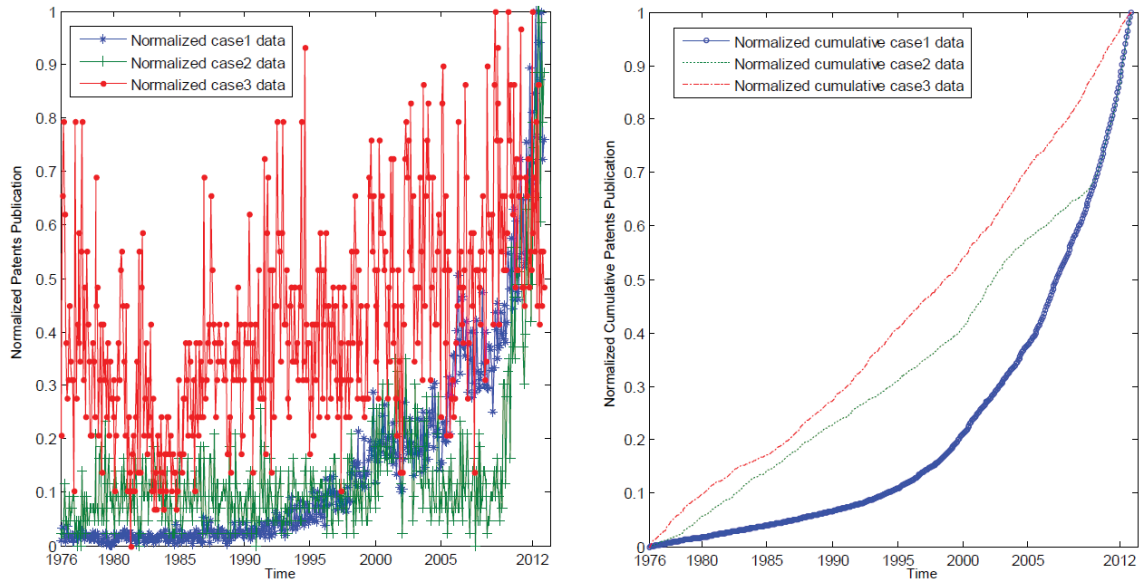


Figure 4-6. Normalized original and cumulative patent data in the case study

For each technology, an optimal threshold *s* needs to be selected to maintain the balance of producing a comparatively smaller number of PLR pieces and a lower RSS before piecewise representation. The RSS of each PLR model is collected with 20

different threshold s from 5 to 24, and the RSS values is denoted for the three technologies with publication count sequences R_1, R_2, and R_3. Then each normalized RSS series is fitted to a corresponding exponential curve, as shown in Figure 4-7. The minimized cost functions for three cases are also presented in the figure below, as red solid lines.

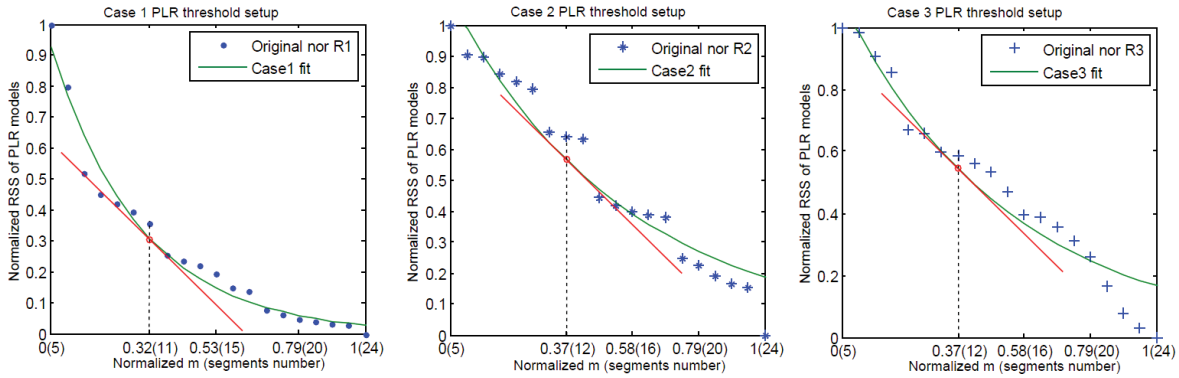


Figure 4-7. PLR threshold setup for the three technologies in the case study

Figure 4-6.

The preferred values for threshold m, where the cost functions and RSS fitted curves have points of tangency in the three cases, are 11, 12, and 12 respectively. The detailed information of every normalized RSS series fits to the corresponding curve is shown in Table 4-4. Before curve fitting, values of x-axis and y-axis are both normalized.

Table 4-4. Curve fitting information for PLR threshold determination

Case NO.	RSS	SSE	RMSE	R-square	Adj R-sq	s
Case 1	R_1	0.0389	0.0465	0.9715	0.9699	11
Case 2	R_2	0.1046	0.0762	0.9383	0.9349	12
Case 3	R_3	0.1117	0.0788	0.9364	0.9329	12

4.5.2 TREND FORECASTING FOR TELECOMMUNICATIONS TECHNOLOGIES

USPC 455, which represents telecommunication technologies, is an important patent class of the Information and Communications Technology industry that has attracted attention globally because of its quick growth in recent years (OECD 2005; Lee, Kim &

Park 2009). This research applied the proposed TTA method using the calculated threshold $s = 11$ to the data of Telecommunication technologies. The piecewise linear representation result is illustrated in Figure 4-8, where the original data is presented in blue, the experimental result of PLR segmentation is highlighted with 11 straight red lines, and the generated trend signal is marked in green. As can be observed from the figure, the green stage-wise line are the trend segments that retains the detailed trend state variation; all the trend turning points are marked in the figure as well.

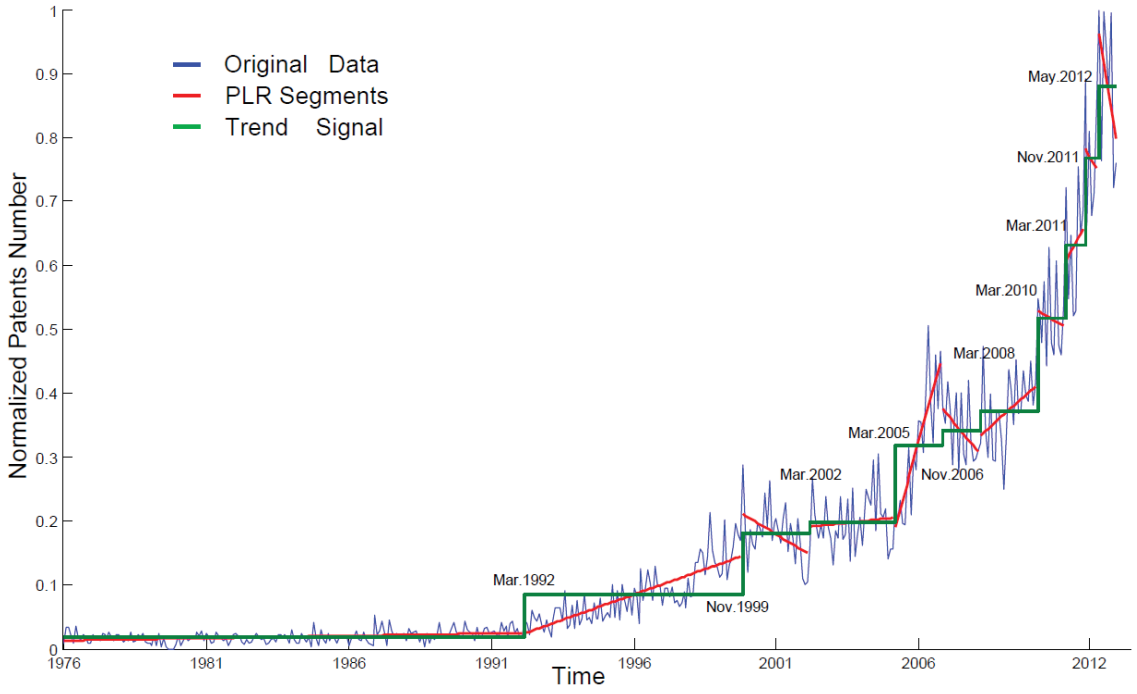


Figure 4-8. Original data, PLR segmentation result and trend segments of telecommunication technologies

The final trend segments presented as a green line in Figure 4-8 illustrate that the telecommunication technologies on the whole experiences a ladder-type growth during the past 3 decades. The gradual upward transitions occurs one step at a time, from slow to fast. Specifically, the trend segments maintain a low and stable status from year 1976 to the beginning of 1992, increases at a generally steady rate from year 1992 to 2010 and finally grows dramatically from the trend turning point at March of 2010. Table 4-5 provides detailed information about all the trend turning points, value of each trend segments and intensity measure of trend state changing. As can be directly observed from Figure 4-8, a very rapid growth starts in March 2010, the coefficient of trend

segments changing intensity is 5.368, much higher than the previous one 0.561, which indicates a high speed rise of the technology trend. In addition, the three trend turning points after March 2010 also lead strong positive movements of the trend, reflecting telecommunication technology entered a period of rapid growth.

Table 4-5. The trend segments information of Telecommunication technologies

Seg.NO.	Turning point start	Start time	Turning point end	End time	Signal value	Intensity
1	1	Jan. 1976	194	Feb.1992	0.0187	0
2	195	Mar.1992	286	Oct.1999	0.0842	0.3158
3	287	Nov.1999	314	Feb.2002	0.1799	1.5188
4	315	Mar.2002	350	Feb.2005	0.1978	0.2206
5	351	Mar.2005	370	Oct.2006	0.3185	2.6782
6	371	Nov.2006	386	Feb.2008	0.3418	0.6464
7	387	Mar.2008	410	Feb.2010	0.3721	0.5612
8	411	Mar.2010	422	Feb.2011	0.5172	5.3680
9	423	Mar.2011	430	Oct.2011	0.6323	6.3891
10	431	Nov.2011	436	Apr.2012	0.7673	9.9917
11	437	May.2012	444	Dec.2012	0.8811	6.3157

After the trend segments were generated, NARNNs was utilized to forecast the future trend. In order to demonstrate the validity of the proposed TTA-based trend forecasting approach, the data of the first 438 months is used as the training set of the NARNNs to predict the trend segments of the last 6 months of 2012. Figure 4-9 shows the empirical forecasting results, where the actual trend segments are presented as a green line, the NARNNs estimated values are illustrated as a blue line. While training the NARNNs model for telecommunication technologies trend segments, 80% of the data is used as training data, 10% data as validation and the last 10% as testing. The number of hidden neurons is set as 20, the number of delays is set as 3. The mean square error (MSE) of the network is $2.082e^{-4}$, while its regression R value is 0.9965. Then the network is used to predict the trend segments for the next 6 months. The forecasting results are marked by red points in Figure 4-9. The MSE of the forecasting results is 0.0049, the mean absolute error (MAE) is 0.0637 and the mean absolute percentage error (MAPE) is 6.6479. The results of experiments demonstrate that this trained NARNNs model predicts the future technology signal reasonably well.

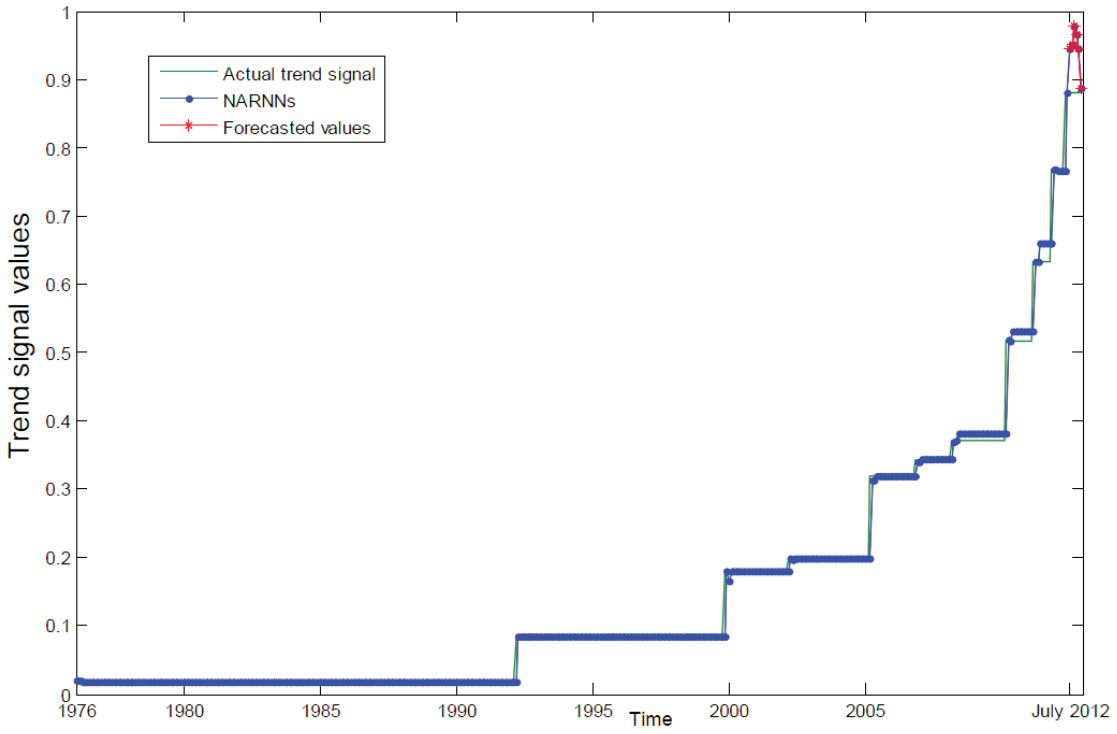


Figure 4-9. The forecasting result of the trend segments of Telecommunication technologies

4.5.3 TREND FORECASTING FOR SOLAR CELL TECHNOLOGIES

As a green energy, the solar cell has experienced very vigorous growth during the past decade. The consumption of solar cell technologies continues to rise from the market's perspective (Tseng et al. 2011). The number of patents for solar cell technologies that published every month from years 1976 to 2012 are collected and a calculated threshold $s = 12$ are adopted as mentioned in sub-section 4.5.1. Figure 4-10 shows the original data, the experimental result of the PLR segmentation, and the corresponding trend segments. PLR-processed data is presented as 12 straight red lines and the trend segments is presented as a green stage-wise line to highlight detailed trend movements, where all the trend turning points are marked in the figure as well.

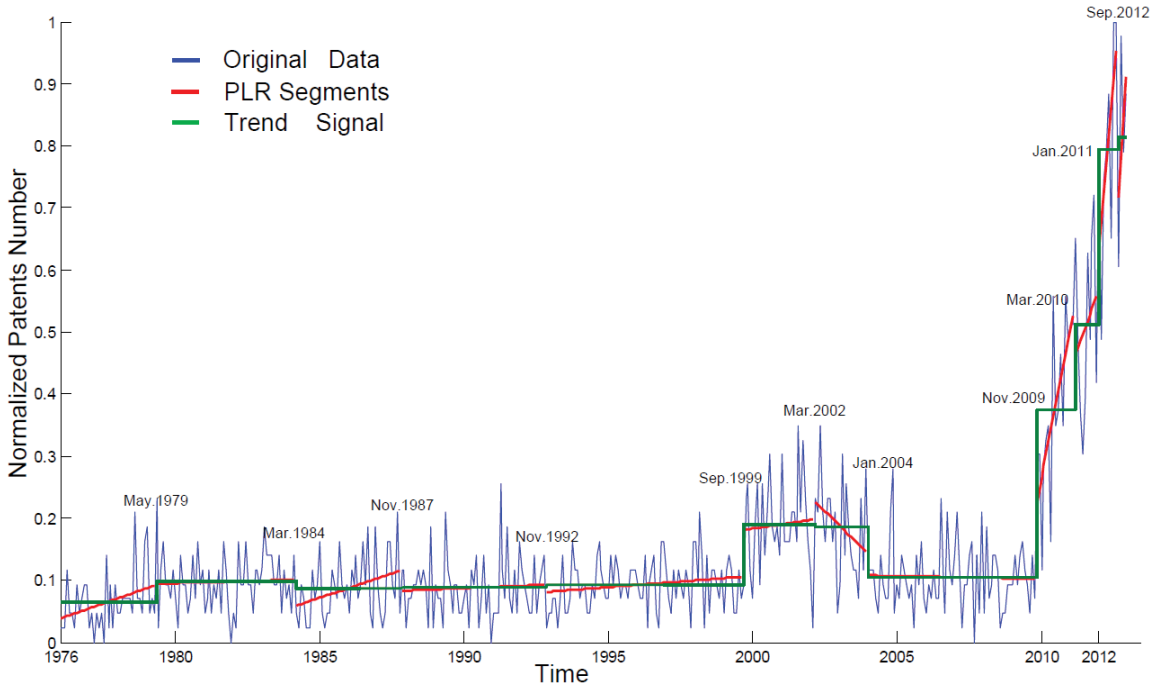


Figure 4-10. Original data, PLR segmentation result and trend segments of solar cell technologies solar cell technologies

Compared with telecommunication, the technology of the solar cell is a relatively emerging one. Its trend exhibits quite different patterns, with less gradual growth but more sudden upward transitions. The final trend segments, as shown with a green line in Figure 4-10, exhibits trend movements of solar cell technologies in the past 3 decades. It remains in a steady state before year 2000, then goes through a slight fluctuation between year 2000 and 2004. Most importantly, from the trend turning point, November of 2009, the trend of solar cell technology shows a quite rapid growth suddenly, which may indicate an important breakthrough in its R&D activities.

Table 4-6 illustrates all the trend turning points, the values of corresponding signal and the intensity measure of trend segments changing. The above mentioned sudden trend shift can also be observed in the table. The intensity coefficient of the 9th trend segment that starts from the turning point in November 2009 and ends in February 2010, is 7.49, which is much higher than the previous one -0.51 and at the same time brings the trend state from downward to upward.

Table 4-6. The trend segments information of solar cell technologies

Seg.NO.	Turning point start	Start time	Turning point end	End time	Signal value	Intensity
1	1	Jan. 1976	40	Apr.1979	0.0645	0
2	41	May.1979	98	Feb.1984	0.0974	0.2518
3	99	Mar.1984	142	Oct.1987	0.0867	-0.1085
4	143	Nov.1987	202	Oct.1992	0.0876	0.0068
5	203	Nov.1992	284	Aug.1999	0.0930	0.0294
6	285	Sep.1999	314	Feb.2002	0.1891	1.4226
7	315	Mar.2002	336	Dec.2003	0.1860	-0.0626
8	337	Jan.2004	406	Oct.2009	0.1050	-0.5142
9	407	Nov.2009	422	Feb.2010	0.3750	7.4930
10	423	Mar.2010	432	Dec.2010	0.5116	6.0662
11	433	Jan.2011	440	Aug.2012	0.7936	15.6497
12	441	Sep.2012	444	Dec.2012	0.8139	2.2587

In the forecasting step, the data of the first 438 months are used as the training set of the NARNNs then the trend segments of the last 6 months of 2012 are predicted. The blue line in Figure 4-11 shows the NARNNs estimated values, while the actual trend segments are presented as the green line. In the same way, as sub-section 4.5.2, 80% of the data is used as training data, 10% as validation and the last 10% is for model testing. The number of hidden neurons is 20, the number of delays d is set as 3. For solar cell technology, the MSE of the network is $4.2857e^{-4}$, while its regression R value is 0.9839. Then this research uses this network to predict the trend segments for the next 6 months. The forecasting results are marked by red points and illustrated in Figure 4-10 as well. The MSE of the forecasting results is $2.7594e^{-4}$, the MAE is 0.0136 and the corresponding MAPE is 1.7092.

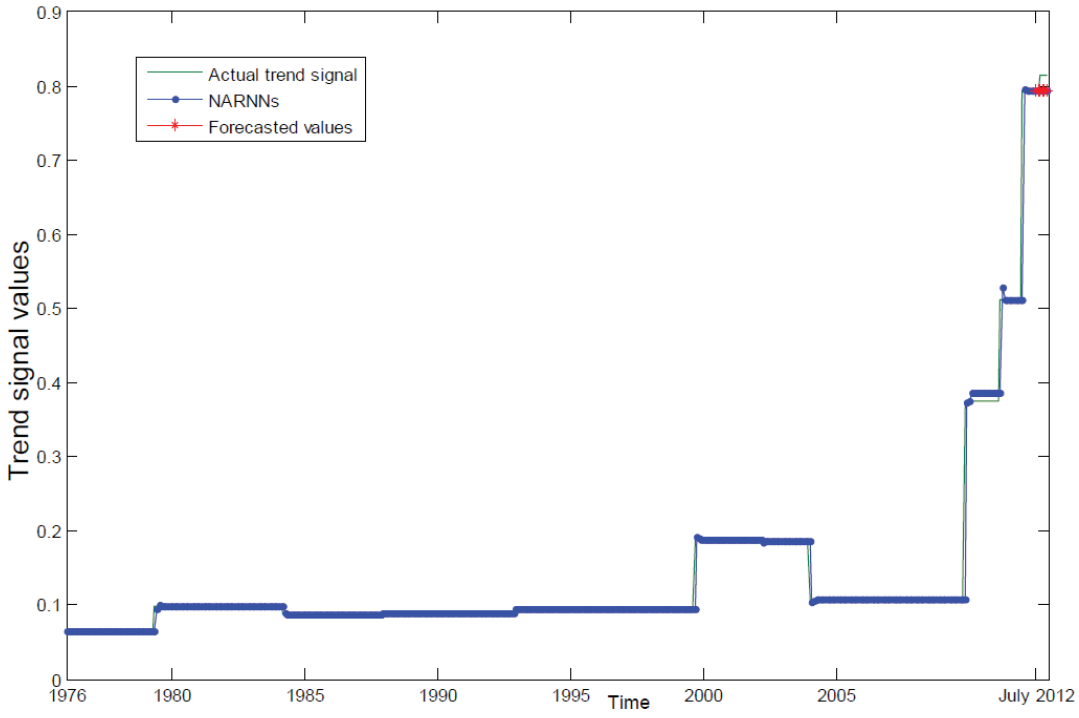


Figure 4-11. The forecasting result of the trend segments of solar cell technologies

4.5.4 Trend Forecasting for Radar-related Technologies in USPC 342

The target technology area of this sub-section is radar-related technologies belonging to USPC 342, which is an important and continuously innovative area in the RFID (Radio Frequency Identification) industry. To improve the efficiency of inventory tracking and management, RFID is increasingly used in enterprise logistics and supply chain management (Ngai et al. 2007). The count of patents published every month, from 1976 to 2012, is collected as raw data. Then the proposed method is utilized with the calculated threshold $s = 12$. Figure 4-12 illustrates the normalized original data, the experimental result of PLR segmentation, and the final trend segments. As can be observed from the figure, PLR-processed data is shown as 12 straight red lines and the trend segments are presented as a green stage-wise line to highlight the captured trend state shifts.

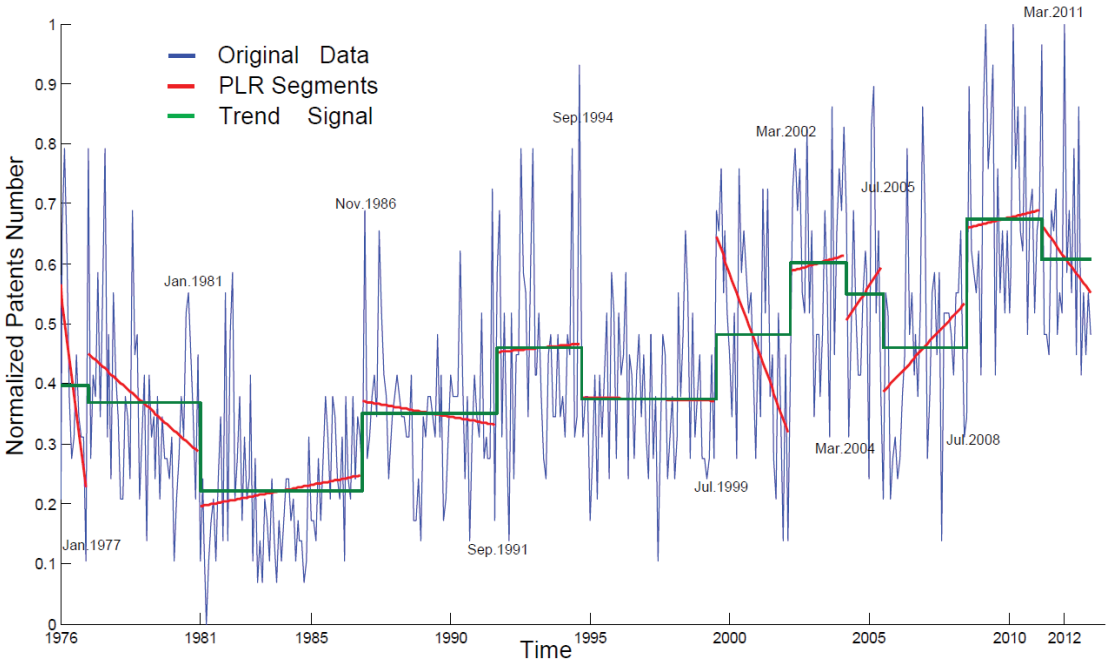


Figure 4-12. Original data, PLR segmentation result and trend segments of radar-related technologies

Different with the two case studies before, radar-related technologies go through several fluctuations and eventually exhibit a slight growth on the whole. As shown in Figure 4-12, the detailed trend of the technology drops down in the first decade, from 1976 to 1986. Then it starts to fluctuate but slowly rise at the same time. However, every time the trend shows an upward movement, it will then go downward in the next trend slice period. The possible explanation for this result is that there may be no significant technological progress in this area and the existing technology is basically stable and sufficient to support industrial activities and development. Although a number of trend segments show upward transitions, the technology improvement may be restricted to fine adjustment or temporary policy/market encouragement. Thus the main tendency of the technology presents only a slow and slight growth during the past 30 years. Table 4-7 provides the detailed information about all the trend turning points, the value of each trend segments and intensity measure of trend segments changing. Corresponding with the fluctuating movements shown in Figure 4-12, for different trend segments, the intensity coefficients alternate between positive and negative repeatedly. The two most obvious trend growths occur at trend turning points March 2002 and July 2008; both bring the trend state from downward to upward.

Table 4-7. The trend segments information of radar-related technologies

Seg.NO.	Trend Start point	Start time	Trend End point	End time	Signal value	Intensity
1	1	Jan. 1976	12	Dec.1976	0.3966	0
2	13	Jan.1977	60	Dec.1980	0.3685	-0.2592
3	61	Jan.1981	130	Oct.1986	0.2212	-0.9346
4	131	Nov.1986	188	Aug.1991	0.3514	0.9966
5	189	Sep.1991	224	Aug.1994	0.4598	1.3370
6	225	Sep.1994	282	Jun.1999	0.3746	-0.6523
7	283	Jul.1999	314	Feb.2002	0.4828	1.5013
8	315	Mar.2002	338	Feb.2004	0.6006	2.1796
9	339	Mar.2004	354	Jun.2005	0.5496	-1.4154
10	355	Jul.2005	390	Jun.2008	0.4607	-1.0957
11	391	Jul.2008	422	Feb.2011	0.6746	2.9670
12	423	Mar.2011	444	Dec.2012	0.6082	-1.3404

Then the data set of the first 438 months are applied as the training set to train a NARNN, which is used to forecast the trend segments of the next 6 months. The blue line in Figure 4-13 shows the NARNNs estimated values, while the actual trend segments are presented as the green line. As in sub-section 4.5.2 and 4.5.3, 80% of the data is used for model training, another 10% for validation and the last 10% for testing. The number of hidden neurons is 20, the number of delays d is set as 3. For radar-related technologies, the MSE of the network is $3.4752e^{-4}$, while its regression R value is 0.9895. Then the trained network is employed to predict the trend segments for the next 6 months. The forecasting results are marked by red points and illustrated in Figure 4-13. The MSE of the forecasting results is $3.5545e^{-5}$, the MAE is 0.0053 and the corresponding MAPE is 0.8877.

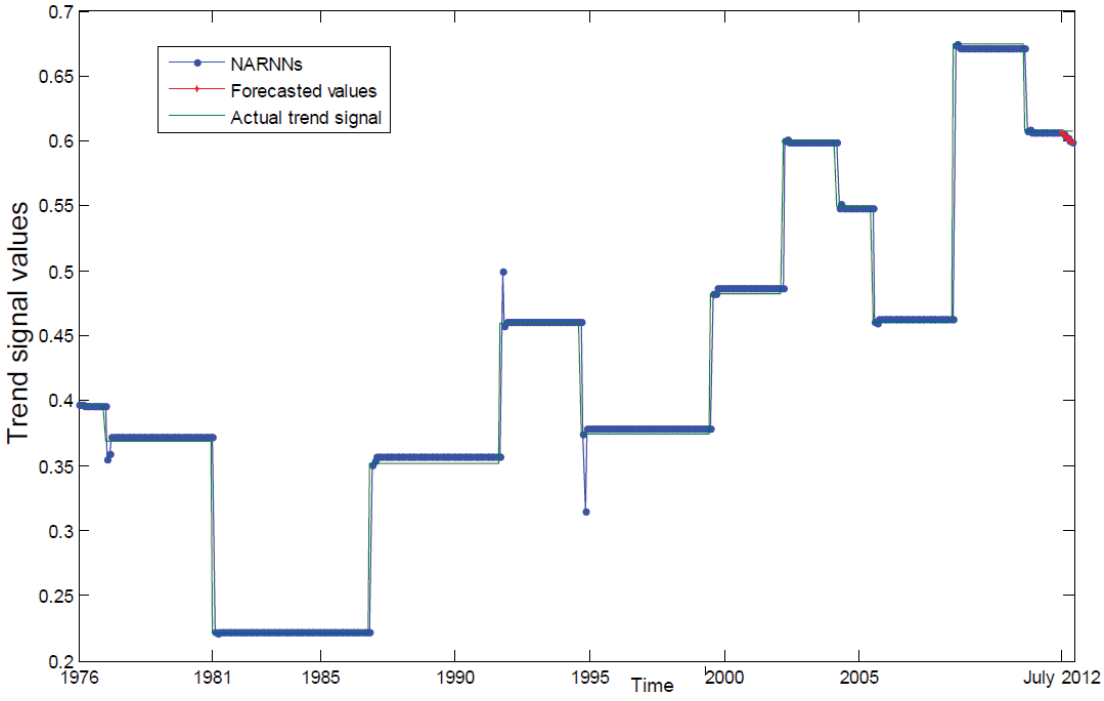


Figure 4-13. The forecasting result of the trend segments of radar-related technologies

4.5.5 Comparison and Discussion

In this sub-section, to demonstrate the superiority of the proposed TTA-based approach in capturing and forecasting detailed shifts, it is compared with one of the most studied and accepted empirical technology trend forecasting approaches, the growth curves. Specifically, this research applies the logistic curves and the Gompertz curves methods to model the cumulative data of the same patent counts record when conducting the comparison.

According to Martino (1993), the upper bound of a growth curve model should be based on the physical and chemical limits that are imposed by nature on the particular technology. However, it is quite rare that the true saturation stage of a technology is known in advance. Thus here an upper limit is estimated by searching a minimized mean square error between actual data and fitting result. Data of the first 438 months of the three case studies are used in growth curve fitting, which is the same data set that employed in the neural networks training. The detail curve selection and their coefficients are presented in Table 4-8. For comparison convenience, this research uses the normalized data to calculate their corresponding cumulative series; in this way, after

curves fitting, the cumulative series can be transformed back to a comparable data which has the same scale as the forecasting result by the PLR-based approach in this study.

Table 4-8. Growth curves selection for comparison experiments

NO.	Growth Curve	a	b	Upper Limit	MSE
1	**Logistic**	612.17	0.0125	214.65	**0.5293**
1	Gompertz	NAN	NAN	NAN	NAN
2	Logistic	33.29	0.0116	61	2.6629
2	**Gompertz**	4.19	0.0041	98.5	**1.7546**
3	Logistic	21.71	0.0103	219	5.0267
3	**Gompertz**	3.81	0.0037	377	**2.7682**

For telecommunication technologies, the logistic curve with upper limit 214.65 is chosen as the final forecasting model, since it provides a relatively smaller fitting MSE under a reasonable upper bound value. The Gompertz curves model is also fitted to cumulative patent counts in telecommunication technologies; however, there is no minimized MSE under a reasonable saturation value when fitting the Gompertz curve. For solar cell and radar-related technologies, the Gompertz curves models are chosen; as shown in Table 4-8, they provide better fitting results in both cases than the logistic curve. The specific fitting results of the three growth curves are shown in Figure 4-14.

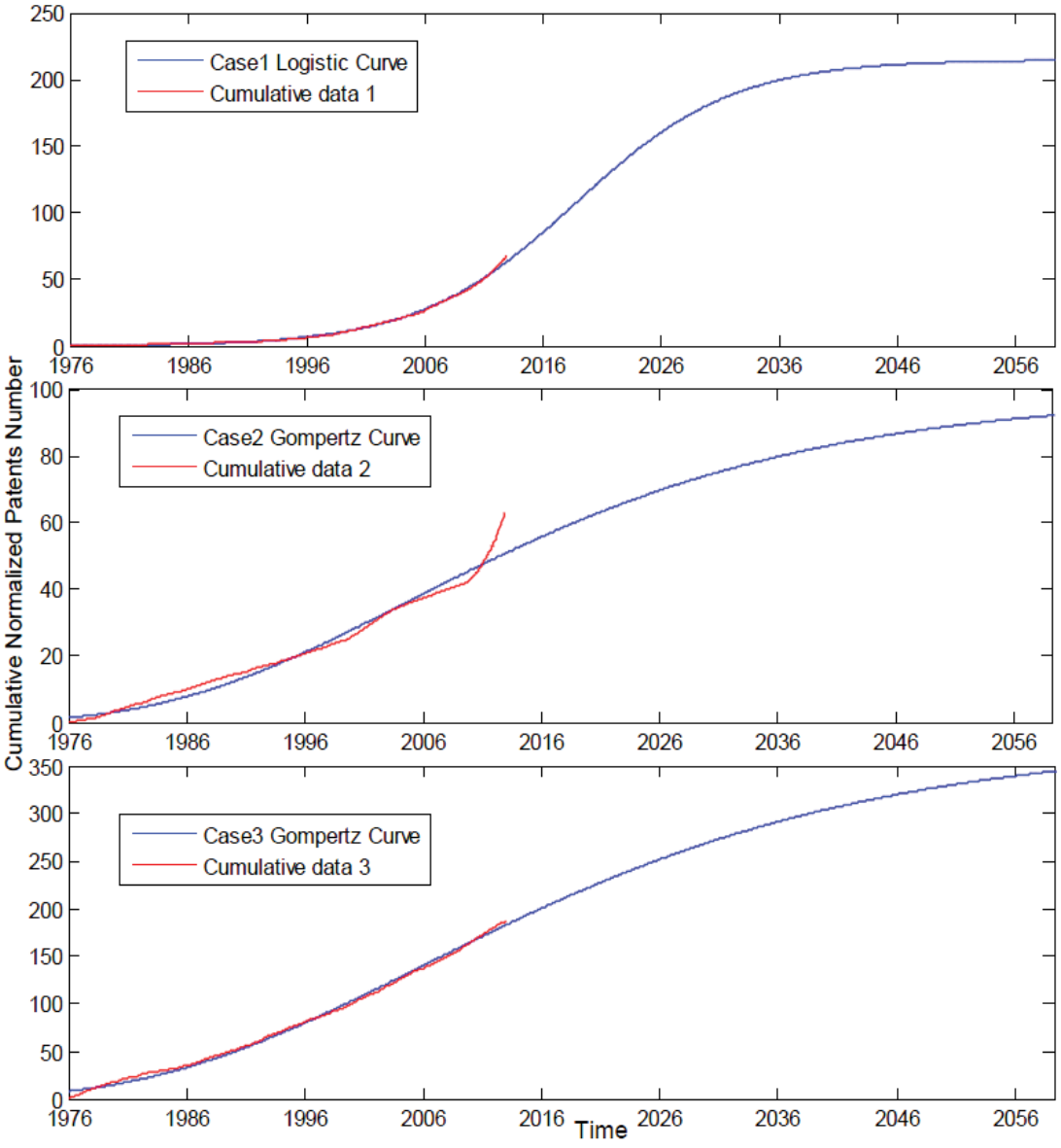


Figure 4-14. The growth curves fitting result for the three technologies

What is obtained from growth curves fitting are the cumulative values of the estimated patent counts series. In order to compare the forecasting result by the curve fitting with the prediction of the proposed approach, this research then conducts difference disposal to the curve fitting results since it is the inverse operation of cumulative sum. The experimental comparisons between this proposed approach and the growth curves method of the three case studies are illustrated in Figure 4-15 below. It can be observed from the figure directly, that the approximation the growth curves provided can generally reveal the main trend of a technology, yet when the observation exhibits

any sudden shifts, it is quite hard to detect the trend movement by growth curves. The proposed TTA-based trend forecasting approach on the contrary, has better performance when dealing with sudden trend shifts and detailed trend movements. It approximates the original observation into stage-wise trend segments, maintaining only the valuable trend turning points of the historical trend, which is quite useful for more accurate trend forecasting.

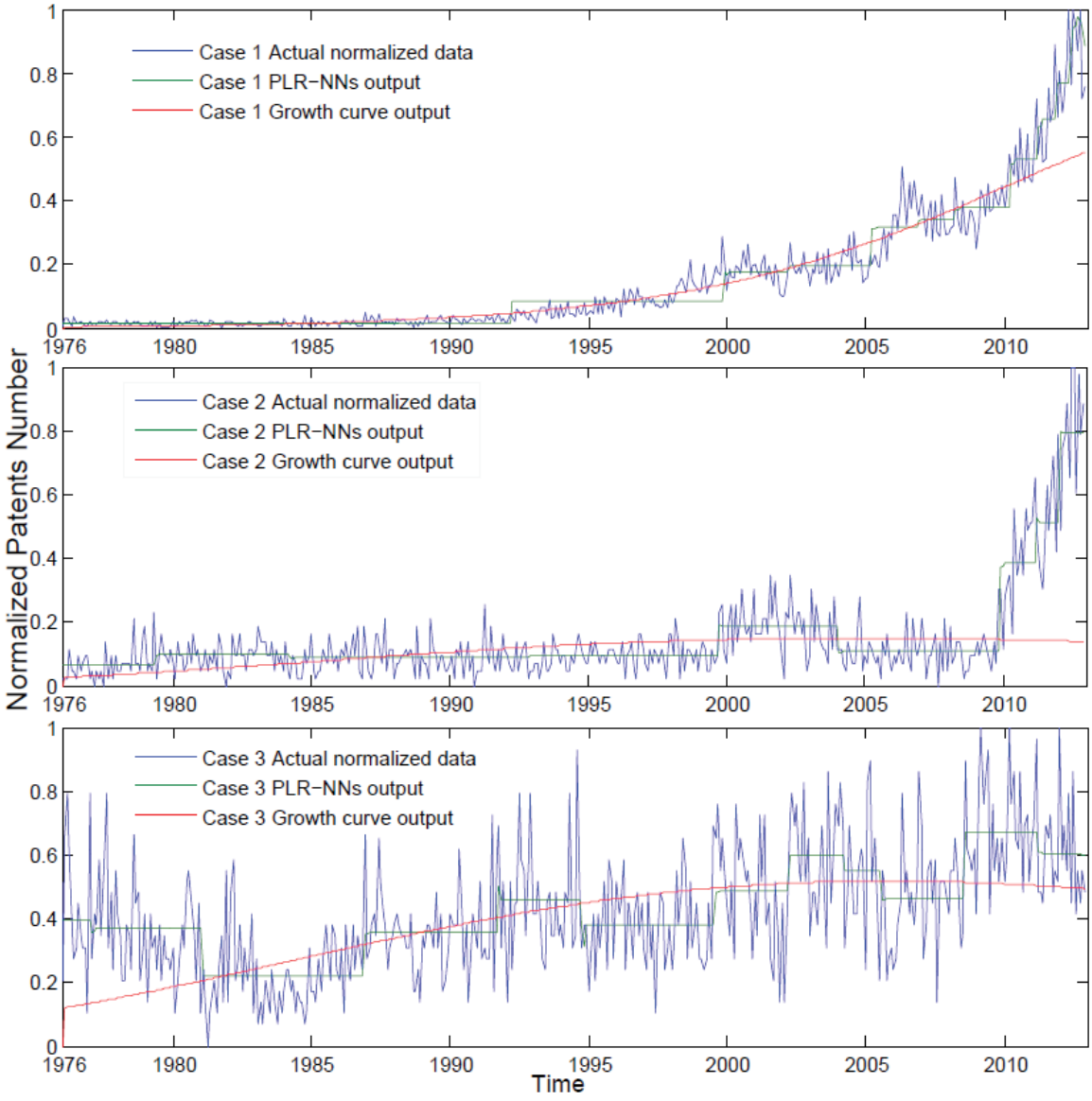


Figure 4-15. Experimental comparison between the proposed approach and the growth curves model

To explain further and to demonstrate the superiority of this proposed approach in trend forecasting, the estimated values of the last 6 months by the proposed approach and by growth curves are listed in Table 4-9. In the trend forecasting, although estimating trend presenting values is useful enough comparing with predicting how many patents will be published exactly in future, the distances between estimated trend presenting values and the original data can still be used to judge the performance of different trend forecasting models. Thus, in Table 4-9, the MSE values of the two models are given to measure the forecasting performances of the two approaches. As can be seen from the table, owing to the more precise trend identification and depiction, the proposed approach has smaller MSE in all three cases, which indicate it has better performance when dealing with technology trend forecasting tasks, irrespective of the growth mode that a technology may have.

Table 4-9. Forecasting result comparison

Telecommunication	Jul-12	Aug-12	Sep-12	Oct-12	Nov-12	Dec-12	MSE
TTA	0.9446	0.9501	0.9775	0.9650	0.9458	0.8857	0.0135
Logistic Curve	0.5373	0.5403	0.5432	0.5462	0.5491	0.5520	0.1255
Original data	0.9966	0.9399	0.8711	0.9957	0.7216	0.7595	
Solar Cell	Jul-12	Aug-12	Sep-12	Oct-12	Nov-12	Dec-12	MSE
TTA	0.7936	0.7936	0.7936	0.7936	0.7936	0.7936	0.0271
Gompertz Curve	0.1406	0.1404	0.1402	0.1400	0.1398	0.1396	0.5618
Original data	1	1	0.6047	0.9767	0.7907	0.8837	
Radar-related	Jul-12	Aug-12	Sep-12	Oct-12	Nov-12	Dec-12	MSE
TTA	0.6066	0.6051	0.6037	0.6021	0.6005	0.5988	0.0240
Gompertz Curve	0.4975	0.4970	0.4965	0.4960	0.4955	0.4950	0.0247
Original data	0.8621	0.4138	0.5517	0.4483	0.5517	0.4828	

The experimental results of the above case studies illustrate that a technology may have different growth patterns during its development. Besides the type of showing partly sigmoidal shape after steady growth as telecommunication technologies, it could have a quite rapid growth without an obvious transitional period just like the trend shown in solar cell technologies, or it may keep fluctuating and showing a slow rise tendency, as shown in the trend of radar-related technologies. It is quite hard to reveal all types of the

patterns and at the same time identify and measure sudden trend shifts or dramatic growth by using only restrained models.

Furthermore, for the afterwards trend analysis and forecasting, compared with the growth curves model, the proposed TTA-proposed trend forecasting approach has better performance while capturing and depicting the detailed movements of technology trends. It is more flexible and data-driven when presenting the concept of "trend". In addition, the proposed approach is effective when it forecasts short-term tendency without specific model selection or upper limits estimation; thus it is less time consuming. It delivers more accurate forecasting results compared to the growth curves.

In summary, the proposed TTA method is a data-driven solution for real-world technological trend analysis tasks. It achieves quantitative trend identification, depiction and prediction by exploring trend segments and a number of trend turning points. The valuable trend patterns learned by the proposed approach in historical patent counts records can be used to predict future technology trends reasonably well. Since the method is data-driven, time and effort can be saved from model choosing and saturation stage estimating in real-world forecasting tasks.

4.6 SUMMARY

In order to capture the hidden trend turning points of patent publication activities, and at the same time improve the framework and applications of existing technology intelligence, this chapter proposes a TTA method with technological trend identification, analysis and forecasting functionalities. The significant contributions of the TTA method include, first of all, the development of an innovative solution for empirical technology trend patterns identification and future trend forecasting by quantitatively identifying and depicting the concept "trend" with trend segments from a data-driven perspective; in addition, it overcomes the limitations of model choosing and upper limits estimating which is experienced by existing empirical technology forecasting approaches; finally, it learns valuable trend patterns from historical patent counts records, then uses the learned trend turning points and trend segments to predict future technology trend. Since the

approach is data-driven, time and effort can be saved from model choosing and saturation stage estimating in real-world forecasting tasks.

CHAPTER 5
TOPIC-BASED TECHNOLOGICAL FORECASTING METHOD AND AFTERWARDS CONTENT ANALYSIS

5.1 INTRODUCTION

This thesis has identified trend turning points and trend segments to quantitatively capture the trend patterns of patenting activities in Chapter 4. This chapter mainly discusses the topic discovery functionality of the proposed technology intelligence framework, provides a topic-based technological forecasting (TTF) method, which aims to uncover the latent topics and temporal trends underlying massive technical documents, and evaluate to what degree various topics have contributed to the patenting activities of the whole area.

In the existing study of technological forecasting, 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 single keywords are usually too general to represent concepts and reflect their corresponding temporal pattern. In real-life situations, even one patent document may contain a number of different technological topics. To overcome these limitations, and integrate the temporal trend patterns and semantic topics quantitatively, based on the outcomes of Chapter 3 and Chapter 4, the proposed TTF method proposed in this Chapter aims to discover and estimate the trends for specific topics underlying large volumes of patent claims using LDA. A case study, utilizing

USPTO patents, is presented to demonstrate the effectiveness of the proposed approach. The results indicate that this proposed approach can effectively generate latent topics from massive patent claims, as well as estimate their very own trend and different contribution levels to the patenting activities, thus providing valuable topic-based knowledge and corresponding temporal patterns to facilitate further technological decision making. This research then expands afterwards technological content analysis based on TTF, presents a topic change identification approach and a fuzzy number-based technological trend measurement approach. The former one builds a topic change identification model to analyse the thematic evolution in a target technical area; the later one provides a fuzzy linguistic description to better explain the estimated trend status.

The remainder of the chapter is organised as follows. Section 5.2 describes patent crawling and cleaning, as the preparation of topic modelling. The stepwise explanation of the proposed TTF method is presented in Section 5.3, followed by Section 5.4, which describes a case study using USPTO patents to conduct an examination of the approach and then explains how to use it in a real patent analysis context. In Section 5.6 this research starts to conduct the afterwards content analysis using LDA, to illustrate a technological topic change identification approach. Section 5.6 continues to expand the application of TTF by presenting a fuzzy number-based technological trend measurement approach. Finally, a summary of this chapter is given in Section 5.7.

5.2 PATENT CRAWLING AND CLEANING

Patent data crawling is an important but more complicated pre-processing task than scientific literature preparation; since not all the time a researcher or analyst is able to get a ready corpus, webpage information crawling may be needed. In this chapter, this research selects patent as the main technological indicator, and explains the textual data collection and preparation. As one of the most significant semi-structured technology indicators, patents comprise structured items such as patent number, issue date, IPC and unstructured items like title, abstract, claims. While patent crawling， this research mainly focuses on its structured items of patent ID, issue date, assignee, USPC, IPC; for unstructured items, this research targets patent title and patent claims, which embody the

core inventive idea and the most essential technological terms to define the protection of the invention. Table 5-1 explains the source code start and end of the webpage source code. All the structured items are put into one single document, while the claims and title for each patent constitute one document in the two corpuses and each of these files is named with its corresponding patent ID.

Table 5-1. The start and end of webpage source code while patent crawling

Item	Source code start	Source code end	Storage methods
PatentID	<TITLE>	</TITLE>	Centralised
IssueDate	<b>United States Patent	<CENTER><b>Abstract	Centralised
Assignee	Assignee:</TH>	</TR>	Centralised
U.S.Class	<b>Current U.S. Class:	</TR>	Centralised
IPC	<b>Current International Class:	</TR>	Centralised
Title	<FONT size="	</FONT>	Separated
Claims	<CENTER><b><i>Claims	<HR>	Separated

Figure 5-1 provides an example of a webpage crawling result. Totally, this chapter creates two patent corpuses based on the USPTO database, to conduct technological forecasting and content analysis. The first corpus covers patents published during years 2000 to 2014 in USPTO (http://www.uspto.gov) with Australia as their assignee country; these are selected as the target patents (ACN/AU AND ISD/20000101->20141231). Their patent ID, titles, issue time, inventors, assignee, USPC and most importantly, their claims, were crawled from USPTO and placed in a patent database for further processing. 13,910 utility patents covering 374 different main USPC are collected. IDs and the issue time of all the target patents form one single file, while the claims and title for each patent constitute one document in the target corpus, which totals 13,910 documents. Altogether, in the target corpus, this research found 103,935 unique vocabularies containing all essential technological topics of inventions owned by Australian assignees in the past 15 years. The second corpus contains the patents related to solar cell (ABST/"solar cell") and published during years 1985 to 2014 in USPTO. In total, there are 3277 target patents covering 3271 utility patents, 5 reissue patents and 1 statutory invention registration in the solar cell area. All the case studies that are presented afterwards in this Chapter are based on these two corpuses or their sub-corpuses.

```
The file has been catched!!!!!!!!!!!!!!!!!!!!!!!!!!!
allnumberofpatents = 16380
The file has been catched!!!!!!!!!!!!!!!!!!!!!!!!!!!
The file has been catched!!!!!!!!!!!!!!!!!!!!!!!!!!!
The file has been catched!!!!!!!!!!!!!!!!!!!!!!!!!!!
The file has been catched!!!!!!!!!!!!!!!!!!!!!!!!!!!
The file has been catched!!!!!!!!!!!!!!!!!!!!!!!!!!!
The file has been catched!!!!!!!!!!!!!!!!!!!!!!!!!!!
URL:http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-
ID = 8864946
time = October 21, 2014
title = Disposal of electrical and electronic equipment
inventors = Scheirs; John (Edithvale, AU)
assignee = P-Fuel Ltd  (Victoria,  AU)
USClass = 201/19; 110/346; 585/241; 588/301; 588/321
INClass = C10G 1/10 (20060101); C10B 53/07 (20060101); C22B 7/00 (20060101)
<CENTER><b><i>Claims</b></i></CENTER> <HR> <BR><BR>The invention claimed is: <BR><BR> 1.  A method
<BR> 1.  A method for disposing electrical and electronic equipment comprising plastic and metal co
```

Figure 5-1. An example of webpage crawling result

After crawling patent data from USPTO, before topic discovery, this research cleans the target titles and claims to retrieve technical terms by performing words cleaning and consolidation. It is not hard to observe that patent claims are a special kind of textual data, containing many technical terms; they also hold specific words serving as transition phrases and numerous academic-related words that describe invention outcomes. Although claims are written with concise but precise language for legal protection reasons, these textual data still need to be cleaned in order to maintain words that provide the most meaningful information reflecting technological topics only. Therefore, this research utilizes three steps 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 this research 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 (Wikipedia 2014); the general academic words list is provided by the University of Nottingham; this

research selected the top 100 most frequent academic words and removed them from the final corpus (Haywood 2003).

5.3 TOPIC-BASED TECHNOLOGICAL FORECASTING METHOD

As this thesis has discussed in Chapter 3 and Chapter 4, when it comes to trend estimation for detailed technical content hidden in large volumes of documents, text mining techniques are required to uncover the latent trends from a semantic perspective. Zhu and Porter (2002) concluded that a managerially utilizable empirical technological forecasting needs to have the capability to efficiently exploit massive textual data. Under such circumstances, in the past five years, topic model-based approaches that provide efficient extraction and informative representations of technological concepts have attracted increasing research interest. In particular, LDA, one of the well-known probabilistic topic models, has provided aid in analysing citation networks, time gaps, content comparison and scientific maps of publications in various areas (Ding 2011; Jeong & Song 2014; Chen, Zhang, Zhang, et al. 2015; De Battisti, Ferrara & Salini 2015; Suominen & Toivanen 2015). In this sub-section, this research presents the details of the proposed TTF method to discover and estimate the trends for specific topics underlying large volumes of patent claims.

5.3.1 METHOD FRAMEWORK

To forecast and analyse the topic-based trend of content underlying in massive textual technical description, there are mainly two tasks to deal with: (1) identifying temporal trend patterns and semantic topics quantitatively; (2) integrating two features in different dimensions to provide topic-based technological trend forecasting. The overall framework, input and output of the proposed approach is examined in Figure 5-2. 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 as shown in step 1. Then, the

titles and claims of patent documents, their corresponding patent ID, issue dates, USPC, and patent publication counts for each month are extracted separately, as shown in steps 2 to 5. The claims and title for each patent constitute one txt document in the corpus, while the patent ID and Issue Date of all patents compose a single file, USPC information form a single file, and patent counts are presented as a sequence of data points.

Textual data containing all the patent titles and claims are first passed to several cleaning and consolidation steps to remove all the punctuation, meaningless symbols, stop-words, general words used in claims and high frequency academic words, as shown in steps 9 to 12. Subsequently, in step 13, LDA is applied to generate latent topics and topic distribution from the prepared corpus, where the USPC information is used to assist with selecting a suitable topic set that better explains the thematic structure of the corpus from multiple experiments, which are presented in steps 14 to 16. Meanwhile, the patent counts sequence is transformed into trend turning points and trend segments. The temporal trend pattern identification task has been finished in Chapter 4, as shown in steps 6 to 8, generated out a number of trend turning points and trend segments waiting for further analysis. As shown in step 18, for each topic, trend turning points and discovered topics are then integrated to calculate a topic-based trend coefficient sequence, which illustrates to what degree the topic has contributed to the patenting activities of the whole area. Meanwhile, step 17 uses the extracted patent issue date information and results from topic modelling to compute a topic annual weight matrix. Finally, the topic-based technological forecasting is conducted by analysing the annual weight variation and trend coefficient changes of a number of prominent topics in step 19.

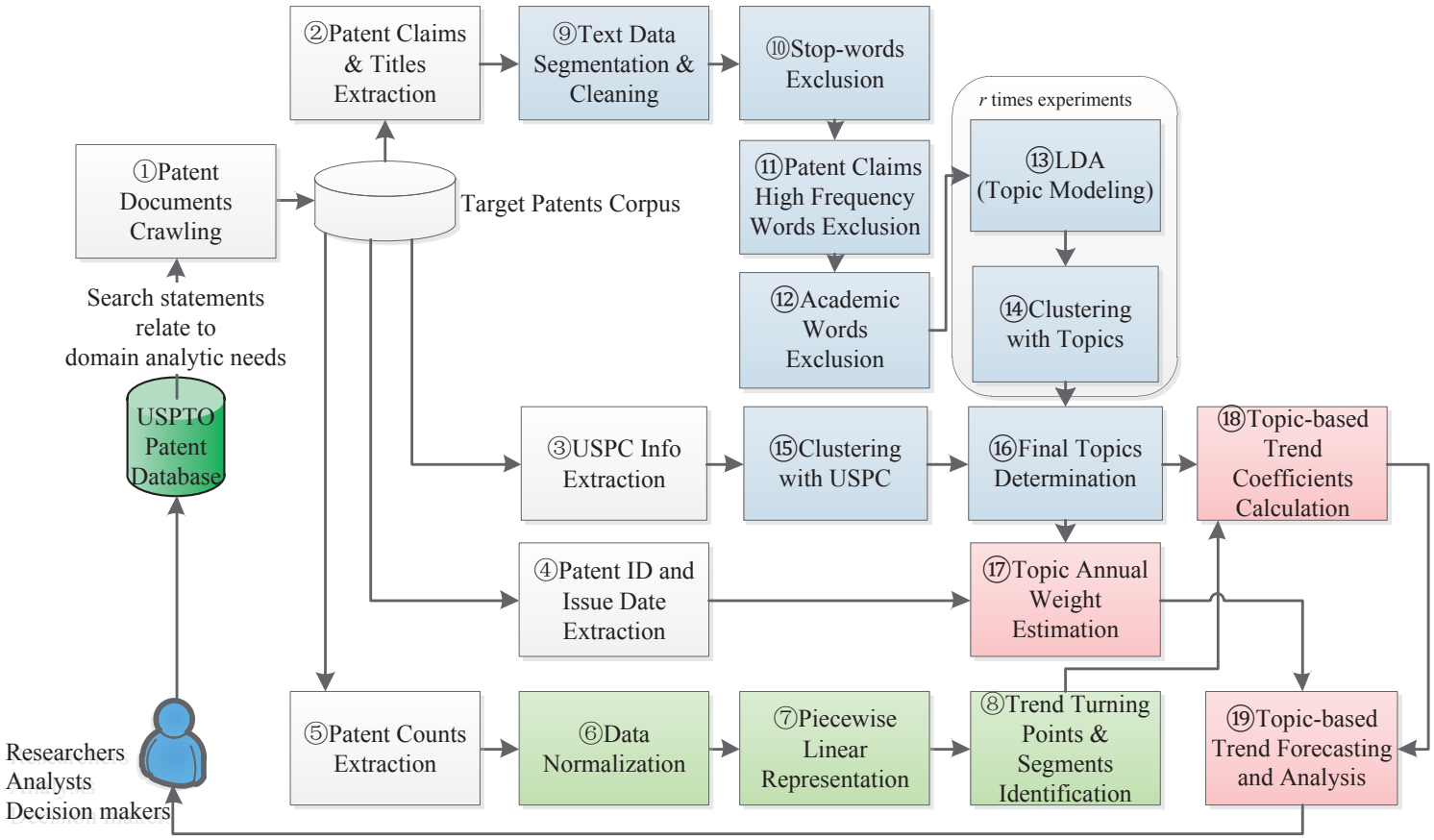


Figure 5-2. Framework for the topic-based technological forecasting method

5.3.2 TOPIC MODELLING

Originally, content analysis in the tech mining area has been restricted to predefined keywords or phrases that captured the meaning of a corpus. In previous research, much effort has already been devoted to identifying valuable terms and phrases from scientific publications and patent documents, using computer-aided text mining, keyword-based morphology, term clumping and so forth (Watts & Porter 2003; Yoon & Park 2005; Cascini & Russo 2007; Zhang et al. 2014). Nevertheless, the outcomes of text mining-based techniques applied to assist topic extraction are mostly keywords with a ranking. These words alone, however, are usually too general or misleading to indicate a concept, especially when there are polysemous words actually describing different topics (Tseng, Lin & Lin 2007). To conduct more accurate and understandable content analysis in the context of technology intelligence, a more managerially utilizable representation of the technology indicator's semantic property is going to be needed. In this research, 'topics', which are represented as groups of soft-clustered words that frequently show up together in a collection of documents, are applied to represent technical concepts and ideas, systematically bringing probabilistic topic modelling to the research of technology intelligence.

The basic notations for topic modelling and trend pattern presentation are listed in are shown Table 5-2. These notations will be used throughout this chapter. From the perspective of topic modelling, the corresponding patent collection associates with multiple technological topics. Before topic modelling, nothing is known 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; and applies 2000 iterations of Gibbs sampling to infer the needed distributions. Different parameter settings may improve modelling performance. Optimizing these parameters will be discussed in the next chapter.

Table 5-2. The basic notations throughout this chapter

Notation	Description
K	Number of topics
D	The overall documents in a corpus
Z	Topic assignment
Θ	The topic mixture proportion ($D \times K$ matrix)
Φ	The mixture component of topics
α	Hyperparameter on the mixing proportions
β	Hyperparameter on the mixture components
m	Total years
P	Raw publication count sequence
TP	Trend turning points matrix
TS	Trend segment sequence

In practice, Gibbs sampling produces different results each time even with exactly the same input and parameter settings. Facing this problem, this research uses USPC to help determine a more suitable topic set that better explains the actual thematic structure of the corpus. As a predefined classification hierarchy built on domain expert judgments, USPC provides a general understanding of the technical area of concern to one patent. Patents covering similar topics are usually assigned to a same main USPC. Specifically, patents are clustered, with their estimated topic distributions Θ and main USPC, using the hierarchical clustering algorithm (Steinbach, Karypis & Kumar 2000).

After each run of LDA, the results were as follows: (1) a .txt file named *DocumentName* to record the order of file reading; (2) a file containing the distribution of words in topics, Φ; (3) a file explaining the distribution of topics in document collection, Θ; (4) a file illustrating topic assignment of all the words, Z; (5) a file listing all K latent topics expressing D documents, in which each topic is represented by its top 20 words with highest possibilities.

5.3.3 Topic-based Technological Forecasting and Analysis

When estimating the future technological trend, there is a need to consider both the temporal pattern of patenting activities and semantic representation. On the one hand, as concluded in Chapter 4, by applying the proposed technological trend analysis method, from the patent counts over time $P = \{p_1, p_2, \dots, p_i, \dots, p_r\}$, where p_i represent the counts of the i^{th} time intervals (months, seasons or years) and r indicates the total month number in m years. A trend turning points matrix TP has already been generated in Chapter 4 as follows,

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

and also a corresponding trend segment sequence $TS = \{ts_1, ts_2, \dots, ts_s\}$, to quantitatively depict the temporal pattern of patenting activities. On the other hand, after topic modelling, the TTF method has 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 equals 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. A number of the most weighted topics are selected using the sum of the columns.

Since the patents are issued along a time line, while topic modelling, if processing all the documents with an ascending order of their issue ID, a topic distribution matrix in chronological order can be obtained, as shown in Figure 5-3. Then a group of elements in a column that is associated with patents published in the same year are added up, and the summation is used to present the annual weight of the corresponding topic. Specifically, matrix $W = (w_{ij})_{m \times k}$ is set 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.

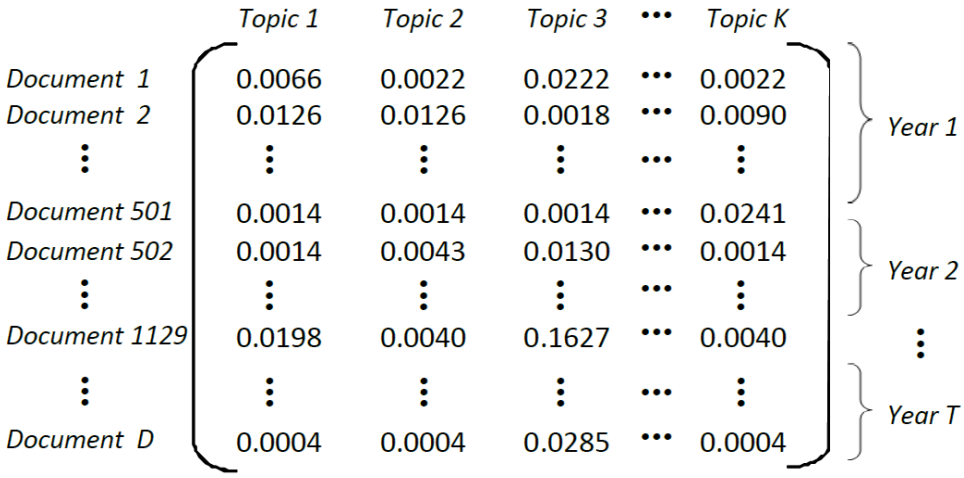


Figure 5-3. An example of topic distribution matrix in chronological order

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. The coefficients a and b are utilized 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 y-axis, the method considers the corresponding topic has a growing trend where the greater a is, the faster the growth will be. Table 5-3 lists the details of using values of a and b to forecast the developing trends of topics.

Table 5-3. Trend forecasting indicators and future trend estimation

Value of a	Symmetry	Parabola opens	Future trend
positive	$-b/2a < m$	upward	upward
positive	$-b/2a > m$	upward	downward
negative	$-b/2a < m$	downward	downward
negative	$-b/2a > m$	downward	upward

Furthermore, for more specific trend analysis, the identified trend segments and discovered topics are integrated to compute a sequence of contribution coefficients and evaluate how different topics contributed the patenting activities of the whole target area, as shown in Algorithm 5-1.

Algorithm 5-1. 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

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

$TC_n = (tc_1, tc_2, tc_3, \dots, tc_s)$

end

For the n^{th} selected topics, let $TC_n = (tc_1, tc_2, tc_3, \dots, tc_s)$ be the contribution coefficients, where tc_s indicates the topic weight on the s^{th} 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 integrate the temporal patterns of patenting activities and semantic topics together to provide topic-based technological trend explanation.

5.4 Case Study of Topic-based Technological Forecasting by Patent Data

In this sub-section, a case study using USPTO utility patent, which was published during years 2000 to 2014 in USPTO (http://www.uspto.gov/) with Australia as their assignee country, is provided to demonstrate the effectiveness of the above proposed topic-based content analysis and technological forecasting approach. The study goal is to forecast the developing trend of specific topics underlying a large volume of patent

documents, and to find to what degree each topic has contributed to the patenting activities of the whole area, in a real patent analysis context.

5.4.1 TREND PATTERN IDENTIFICATION

Using the technological trend analysis method proposed in Chapter 4, the published patent counts of each month are collected to generate a counts sequence and this is normalised 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, the value that produced the maximum absolute value of the approximate derivative, $s = 5$, is chosen as the optimal pieces number. As shown in Figure 5-4, the normalised data was decomposed into five trend segments to reveal and highlight a group of main trend shifts quantitatively. 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.

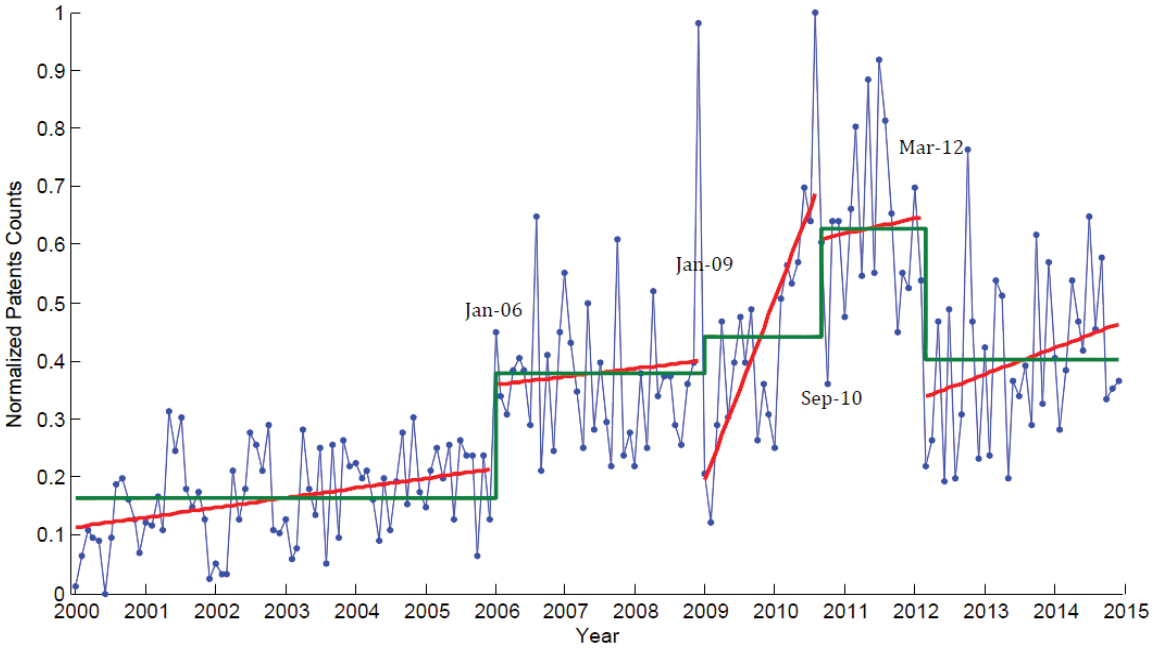


Figure 5-4. The trend turning points and trend segments generated from patenting activities

It can be observed directly from Figure 5-4 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 during the past 15 years. In the first six years from 2000 to 2005, the trend maintains a low and stable status. There is then an important trend turning point appearing in January 2006 when a sharp upward transition occurred, which implies breakthroughs in R&D activities or expansion of existing technological topics. After this trend turning point, the publication of patents almost doubled its number. In January 2009, the next trend turning point occurred, indicating another round of rapid growth of the patents. From September 2010 to February 2012, lasting one and half years, the trend has reached a peak for the time being, and started to drop down. 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 that some technological topics have diminished. Table 5-4 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 others, and had 3706 patents; while trend segment 3 contained the least documents, with a total of 1898 patents.

Table 5-4. 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

5.4.2 Topic Modelling and Prominent Topic Selection

After excluding the stop words, the high-frequency common phrases used in patent claims and general academic vocabularies from the data collection, LDA parameters $K = 50$, $\alpha = 0.5$ and $\beta = 0.1$ were applied to conduct topic modelling. Totally, this study performed 5 ($r = 5$) runs, with 2000 iterations of Gibbs sampling to decide the final topic set. After clustering all 13,910 patent using both topic distributions and their USPC, the trial with highest values of similarity is selected. 50 latent semantic topics were

estimated out, in which each is presented by the top 20 ranked words and their corresponding probabilities. For reading convenience, 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.

The top 10 most weighted topics are selected using the topic distribution matrix. Table 5-5 lists the topic weight of all estimated 50 topics, and highlights the selected 10 topics in bold. These prominent topics of utility patents owned by Australian assignees in the past 15 years cover printhead (topic 37), nozzle (topic 12), axis drive shaft (topic 17), wall body (topic 5), sensing device (topic 6), fluid valve (topic 10), amino acid sequence (topic 15), composite material (topic 31), antibody composition (topic 33) and signal & circuit (topic 13).

Table 5-5. The 50 topics generated from the patent claims collection and their weight indicator

TopicNO.	Topic	Weight	TopicNO.	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	228.530	Topic 19	Explosives	97.017

	communications				
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

5.4.3 Topic Annual Weight Matrix and Topic-based Trend Coefficients Estimation

Since the 13,910 utility patent claims documents were published following a strict time line, the smaller the patent ID, the earlier it was published. All the files were named with their patent ID. While topic modelling, the documents were processed with an ascending order of their name tag. Following the steps of TTF method, after identifying both prominent topics and trend turning points, the annual weight matrix is then generated to illustrate annual weight changes in each topic, as shown in Table 5-6.

Table 5-6. The annual weight matrix of the selected top 10 significant topics

Year	**T37**	**T12**	**T17**	**T 5**	**T 6**	**T10**	**T15**	**T31**	**T33**	**T13**
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	126.98	98.659	49.701	42.766	72.082	37.298	31.962	40.343	20.720	30.054
2007	150.04	84.661	50.609	40.772	77.192	30.371	29.372	35.444	20.144	30.130
2008	199.10	98.232	54.905	39.844	74.694	32.522	23.483	30.705	30.166	24.988
2009	130.78	91.007	44.239	39.792	60.803	26.813	38.640	28.134	29.481	25.073
2010	223.09	101.88	62.789	65.746	101.60	52.418	40.633	40.568	35.724	33.633
2011	197.48	120.11	56.470	69.065	143.48	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

To further evaluate how various topics contribute to the patenting activities of the whole target area, calculations of the topic-based trend coefficients are continued using

Algorithm 1 and the result are provided in Table 5-7. 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 what it contributes to segments 1 and 2. In summary, different latent topics contribute trend changes of patenting activities differently, and these topic-based trend coefficients can be used to measure the varying degrees of their involvement.

Table 5-7. Topic-based trend coefficients for all 10 prominent topics

	DocNO.	TS	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

5.4.4 Topic-based Trend Forecasting and Analysis

After generating the topic annual matrix and trend coefficients, the weight changes of each prominent topic are then forecast in a least-squares prospective. The latent topics generated from the document collection actually have their very own trends and different contribution levels to the patenting activities of the whole area.

Figure 5-5 presents the fitting curve for the 10 selected prominent topics. It can be observed from the figure that Printhead (topic 37) and Nozzle (topic 12), as 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 from years 2012 to 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 booming period, and they may become less important than other topics in the next few years. The significance of topics Axis drive shaft (topic 17) and

wall body (topic 5), to the contrary, are gradually growing, indicating that the two topics have development potential in future. Among the rest of the topics, sensing device (topic 6) and composite material (topic 31) have downward trends. Yet, the decline of topic 6 was more dramatic than topic 31, which basically remained steady and showed 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 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.

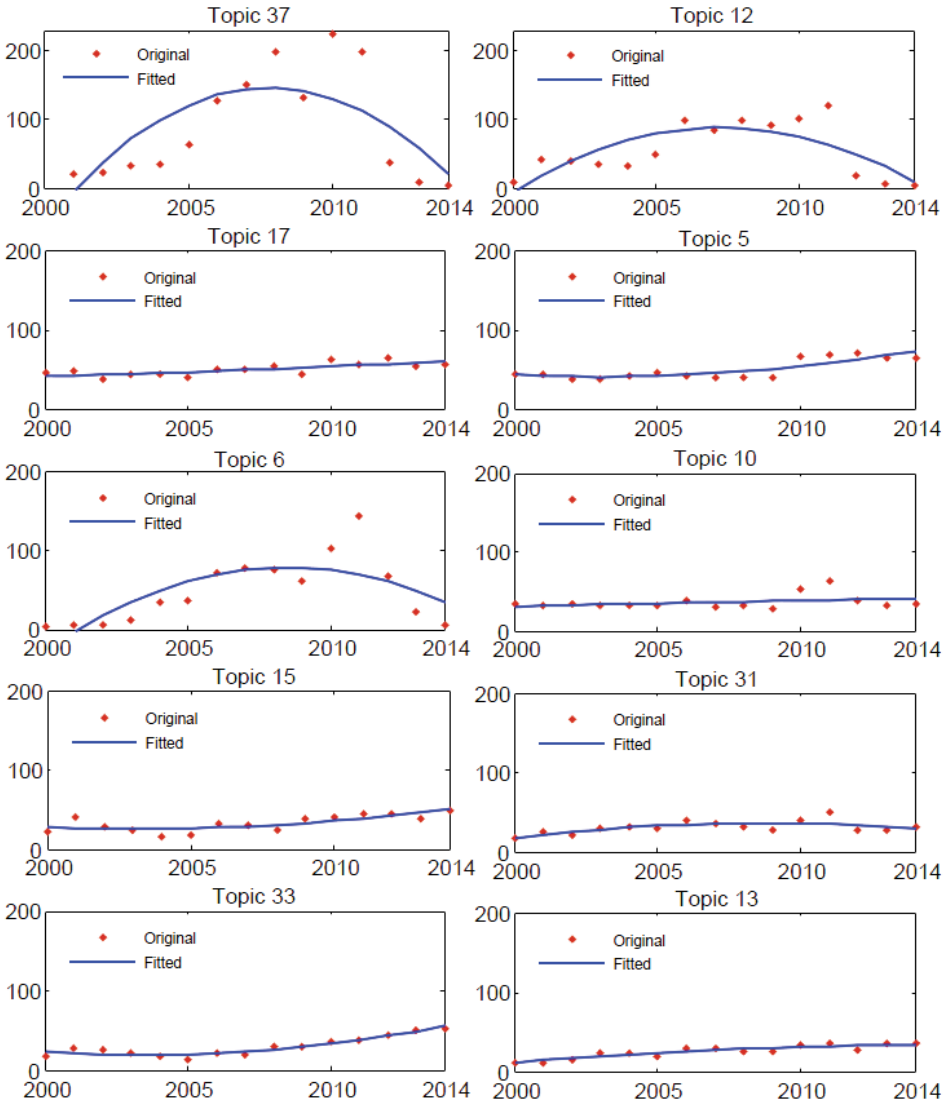


Figure 5-5. The curve fitting result for topic trend estimation of the 10 selected topics

In Table 5-8, all the quadratic polynomial fitting coefficients are examined, and a summary of the topic-based trend forecasting is provided. As mentioned, there is an important trend turning point that 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 more than any other segment, it shows that the development of printhead, nozzle and sensing device from years 2006 to 2009 increased the patent publications for the whole area. The significance of these three topics, however, all dropped quickly on the fifth trend segment, which indicates that their developing potential, compared with other topics, are limited. Topic 33, antibody composition, appeared to have a quite opposite trend. It contributed mainly to the last trend segment. Since 2005, the significance of this topic started growing continuously, from which it can be seen 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 Appendix Table 1.

Table 5-8. The details of trend estimation for the 10 selected topics

Topic NO.	Topic	a	b	Symmetry	Most Contributing	Least Contributing	Future Trend
Topic 37	Printhead	-3.25	56.99	8.78	TS 2	TS 5	Downward trend
Topic 12	Nozzle	-1.76	29.25	8.30	TS 2	TS 5	Downward trend
Topic 17	Axis drive shaft	0.03	0.87	-14.33	TS 1	TS 3	Upward trend
Topic 05	Wall body	0.26	-2.03	3.89	TS 1	TS 3	Upward trend
Topic 06	Sensing device	-1.45	27.52	9.47	TS 2	TS 5	Downward trend
Topic 10	Fluid valve	-0.02	1.07	23.52	TS 1	TS 3	Upward trend
Topic 15	Amino acid sequence	0.21	-1.64	4.01	TS 1	TS 4	Upward trend
Topic 31	Composite material	-0.24	4.73	9.83	TS 1	TS 3	Downward trend
Topic 33	Antibody composition	0.32	-2.75	4.34	TS 5	TS 3	Upward trend
Topic 13	Signal&Circuit	-0.09	3.03	17.01	TS 1	TS 3	Upward trend

5.4.5 DISCUSSION

The case study demonstrates that a patent document collection actually associates with multiple underlying technological topics, and at the same time, these latent topics have their very own trends and different contribution levels to the patenting activities of the whole area. From a methodological perspective, the main contributions of the TTF method include the following aspects:

- It proposes a stepwise methodology to quantitatively identify temporal trend patterns and semantic topics, and integrates these two features in different dimensions, to provide topic-based technological trend forecasting.
- It reveals the latent topics in massive patent claims with high accuracy, and a group of trend turning points, for detailed technological forecasting and analysis.
- It estimates the developing trends for specific latent topics, instead of a broad technological area, overcomes the limitation where simple keywords and rankings are too general or misleading to indicate a concept, and reflects their corresponding temporal pattern.

From the viewpoint of application, the proposed TTF method can be used to automatically uncover the thematic structure of massive patent data in a technological area of interest, and estimate the detailed developing trend of each detected topic, thereby assisting decision making for potential opportunity identification, technical strategy formation, and so forth. 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. This understanding enables entrepreneurs to prepare appropriate technical proposals with potential while at the same time providing venture capitalists with the confidence to support companies with a better understanding of the current situation in a certain industry (Holst, Nguyen & Wikander 2010).

5.5 TECHNOLOGICAL TOPIC CHANGE IDENTIFICATION APPROACH

An empirical case study and objective evidence have been presented to explain the effectiveness of using the proposed TTF method in real world patent analysis. The proposed TTF method actually can serve as a basic method of connecting the temporal patterns and semantic topics of semi-structured technology indicators. In this sub-subsection, facing the demand of revealing thematic evaluation of target corpuses, a technological topic change identification (TTCI) approach is presented, to conduct afterwards content analysis based on the proposed TTF method.

5.5.1 FRAMEWORK OF THE TTCI APPROACH

The overall framework of the proposed TTCI approach is shown in Figure 5-6. Similar to the TTF method, users need to first initiate a search statement to declare their domain analytic needs and address a group of target patents in the USPTO database. Patent ID, title, claims, issue time, assignees, 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 labelled with their corresponding issue year. For each sub-collection, patent claims and titles embodying essential technical terms, and USPC are extracted from the target patents database separately. The two plates in the figure indicate replication.

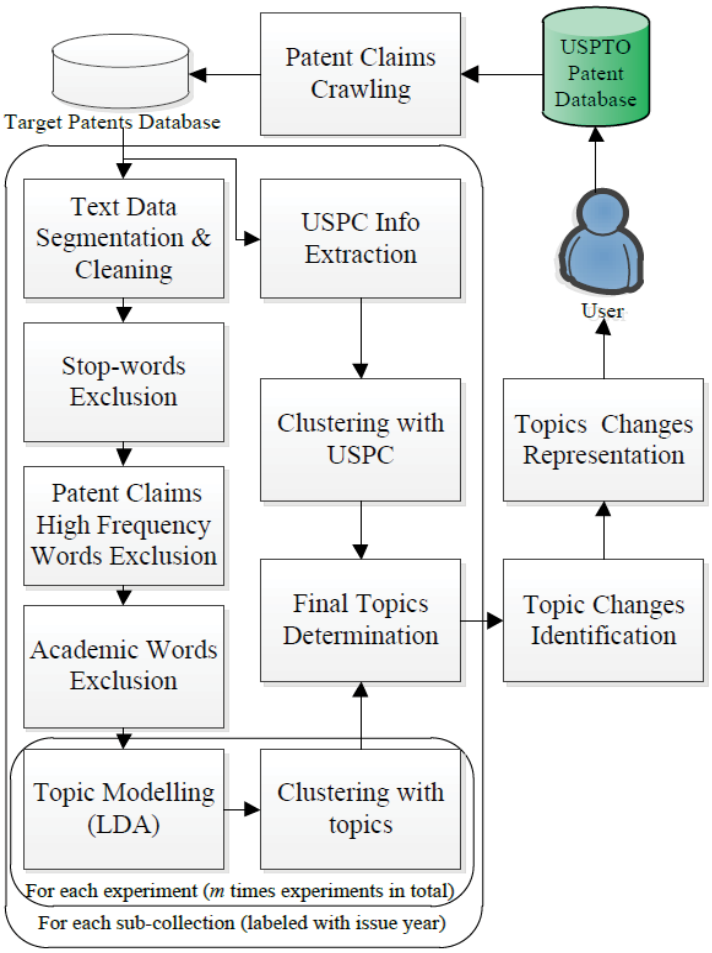


Figure 5-6. The framework of the proposed technological topic change identification approach

Textual data composed by claims and titles, after data segmentation and cleaning, is then placed into a series of words exclusion steps to filter out the most common function words, high frequency words that commonly appeared in patent claims, and academic words with vague and general meanings. Then the prepared text is passed to the topic modelling step. To acquire the most reliable topics of the corpus, based on the proposed TTF method, USPC continues to be utilized as a measurement to evaluate results from a series of experiments. Patents are clustered with both their USPC and topic proportions. The final topic modelling result is the one trial that provides the most similar clusters to the USPC clustering outcome. With all the topics estimated from patent sub-collections of different years, topic changes over time can be identified and presented to researchers and analysts by using a topic change identification model.

5.5.2 TOPIC MODELLING

In a sub-collection, the claims and title of each patent constitute one document, and the number of documents equals the number of patents. After removing all commonly used words from the corpus, LDA is used to generate a number of topics for each patent sub-collection in the corpus, which is labelled by its corresponding issue year. Here the global topics are presented as $\vec{P}_{1:t} = (\vec{P}_1, \vec{P}_2, \dots, \vec{P}_i, \dots, \vec{P}_t)$, where $\vec{P}_i$ stand for the topics of the i^{th} sub-collection of the corpus. The relationship between sub-collections and topics is illustrated in Figure 5-7.

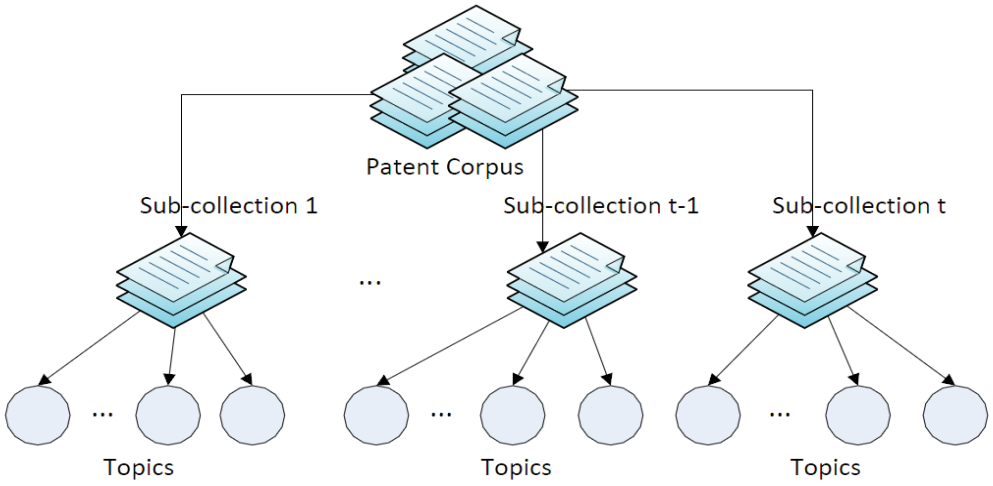


Figure 5-7. Relationships between sub-collections and topics

According to previous research, hyper-parameters α, β of the Dirichlet distribution in LDA have a smoothing effect on multinomial parameters. Based on the proposed TTF method, this case study sets $\alpha = 0.5$ and $\beta = 0.1$, which are commonly used in LDA applications (Koltcov, Koltsova & Nikolenko 2014). During the implementation, topic number K needs to be decided case by case, balancing user requirement and time consumption.

Gibbs sampling is then applied 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. 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, Figure 5-6, 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. 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 modelling result is.

5.5.3 TOPIC CHANGE IDENTIFICATION MODEL

After locating the final topics and words underlying the sub-collections of the corpus, the topic change over time can be identified. Two groups of topics deriving from different corpus sub-collections are compared, 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. Figure 5-8 provides a schematic diagram of the topic change identification model. If two topics under different sub-collections contain approximately the same group of words, then it is believed 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 it is believed these are two different topics. Finally, for documents sub-collection of year i, if there is no similar topic that can be matched in the previous year, year $i-1$, then the un-matched topic in the later year can be seen as a newly important one, which means it became more important in the year i.

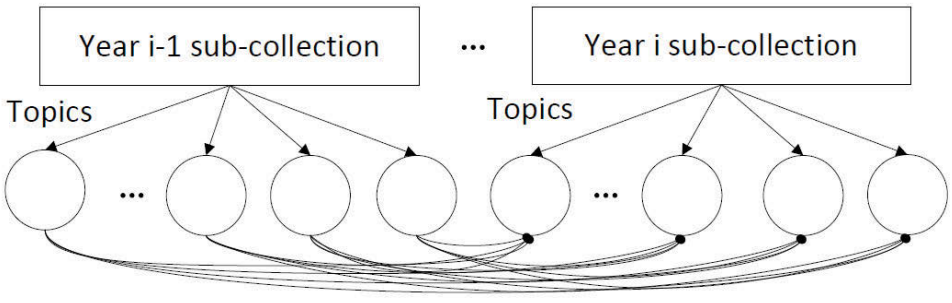


Figure 5-8. Topic change identification model

Algorithm 5-2. Topic-based trend coefficients estimation

input: Topic set $\vec{P}_i$ and topic set $\vec{P}_{i-1}$

output: a .csv file contain the wordlist, former row_number and the latter row_number .

```
1       Read topic set P_i line by line to formerlist
2       Read topic set P_{i-1} line by line to latterlist
3       Set rowsLength = formerlist.length;
4               for (int i = 0; i < rowsLength; i++)
5                       for (int j = 0; j < formerlist[i].length; j++)
6                               Set rowNumber = match(formerlist[i][j], latterlist)
7                               formerRowList.add(i)
8                               latterRowList.add(rowNumber)
9                               stringList.add(formerlist[i][j])
10                      end for
11              end for
12              for (int i = 0; i < formerRowList.size(); i++)
13                      print ( stringList.get(i) + (formerRowList.get(i)+1) +
                        latterRowList.get(i))
14              end for
15      set rowNumber = 0;
16      set rowsLength = latter.length;
17              for (int i = 0; i < rowsLength; i++)
18                      for (int j = 0; j < latter[i].length; j++)
19                              if (word.equals(latter[i][j]))
                                        rowNumber = i+1
                                        return rowNumber
20                      end for
21              end for
22      return rowNumber
```

If two topics under different sub-collections contain approximately the same group of words, which means there is only one topic evolving from year to year, besides

discovering how the detailed content of the topic evolves from year to year with the above model, the topic distribution matrix can also be used to generate historical topic-based trend and to forecast future trend using the topic distribution matrix Θ. Based on the proposed TTF method, in this TTCI approach, a topic annual weight matrix is then generated, and the weight changes is calculated in a least-squares sense to estimate the general trend of the target topics.

5.5.4 Case Study of TTCI Approach by Patent Data

To demonstrate the performance of the 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, which is a sub-corpus of the data this research used in sub-section 5.4. The claims and title for each patent constitute one document in this corpus, which totals 7071 documents on the whole. Then the whole document collection is divided into five sub-collections to present technological features and essential terms of inventions by Australia assignees in the past five years. The detailed documents number published every year from 2009 to 2010, the term number and USPC number in each corresponding sub-collection are shown in Table 5-9. 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 recent three years. It can also be observed 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.

Table 5-9. The number of documents, terms and USPC of patents published each year

Year	Doc NO.	Term NO.	USPC NO.
2009	1174	19796	199
2010	1613	24726	233
2011	1746	23757	228
2012	1256	25102	233
2013	1282	29714	227

Before topic modelling, as mentioned, a number of parameters need to be set first, including the number of topics K, and α, β of Dirichlet distribution. This case study

applies $K = 10$ with model hyper-parameters $\alpha = 0.5, \beta = 0.1$ to the target documents, to balance the topical granularity, convenience of understanding, and 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.

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, where 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 the Table 2 of the Appendix.

- 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).

After discovering main topics underlying in patent claims of each year's document collection, the topic change model is then used to identify the topic variation from years 2009 to 2013. For different groups of topics associated with two consecutive years, this research conducts traversal comparison is conducted 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 5-9 illustrates the important topics that arose each year after 2009, by presenting the top 10 words for each topic using Pajek (Batagelj & Mrvar 2002).

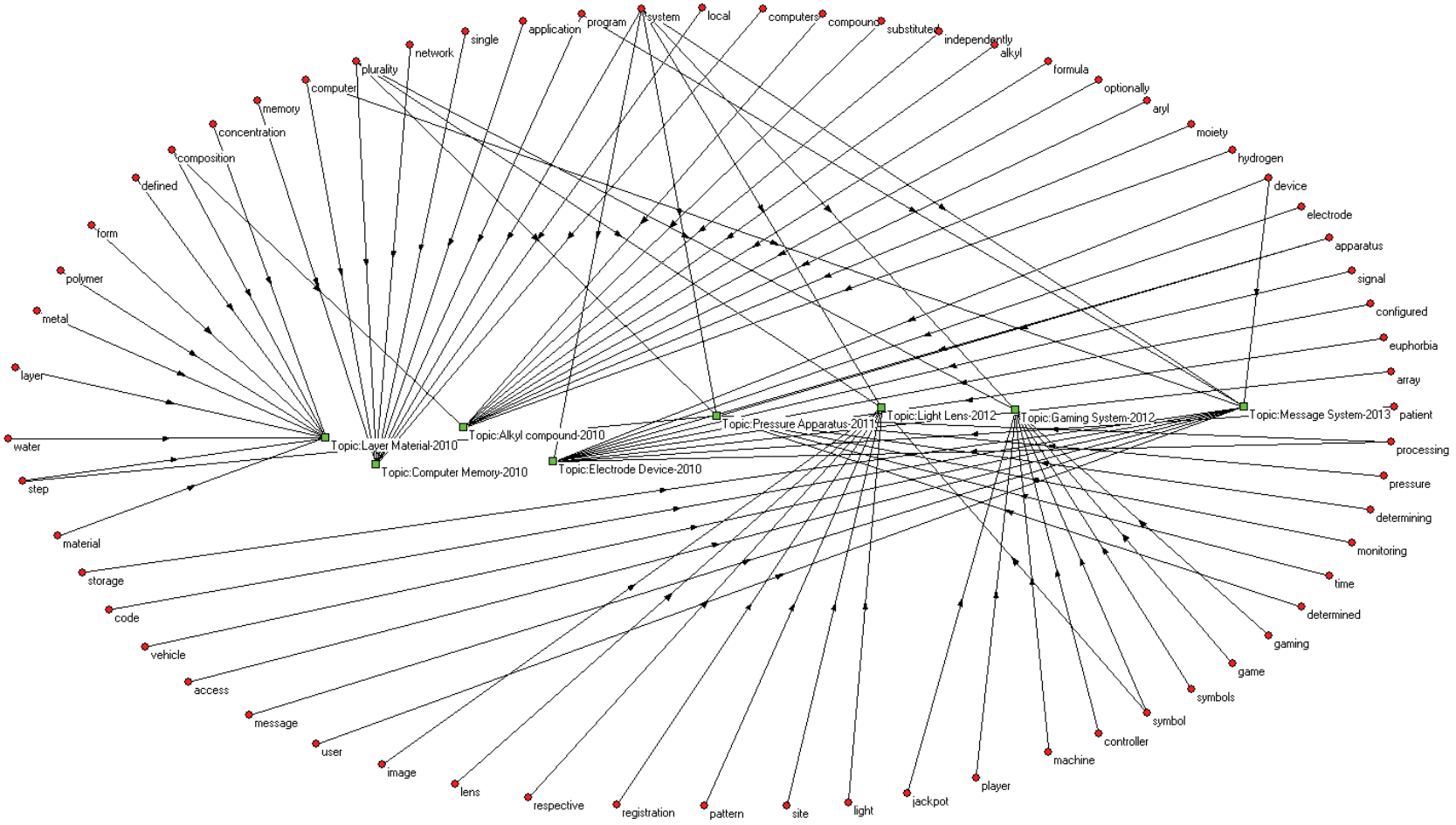


Figure 5-9. Topics became newly important in each year of 2010 to 2013 and topmost frequent words of each topic

In year 2010, 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 pre-set domain knowledge. The detailed words and their corresponding probabilities of the new topics mentioned above are highlighted in boldface in the Table 2 of the Appendix.

As mentioned, the proposed approach can be used to discover how the detailed content of a certain topic evolves from year to year and the topic-based trend can be forecast using historical status. In the case study, topic antibody fragment/sequence is chosen as an example. As shown in Figure 5-10, it can be observed 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 a 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.

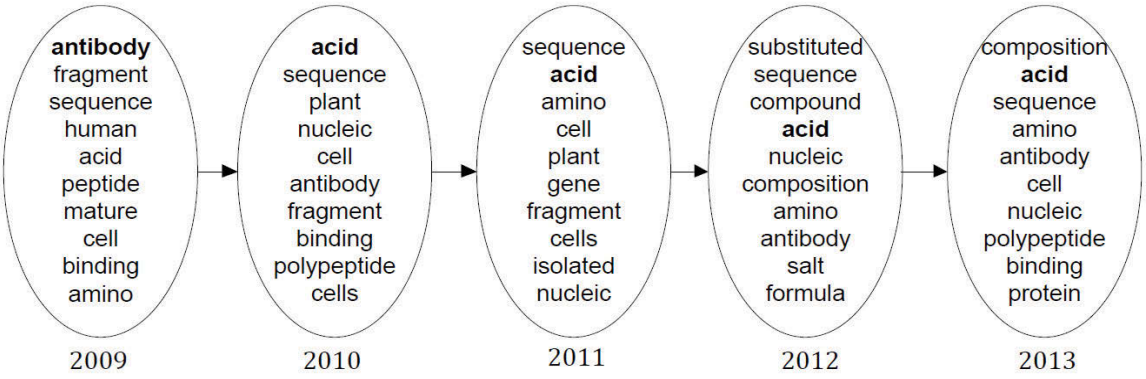


Figure 5-10. An example of the topic "antibody" evolving over time

To estimate topic-based trend of this topic, its annual weight matrix is then generated 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. The weight changes are calculated in a least-squares sense to estimate the general trend of the target topic. Figure 5-11 shows the final result of topic-based trend estimation of the theme "antibody". It can be observed directly that this topic appeared to have an upward trend. The significance of this topic kept growing continuously, from which it can be seen 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.

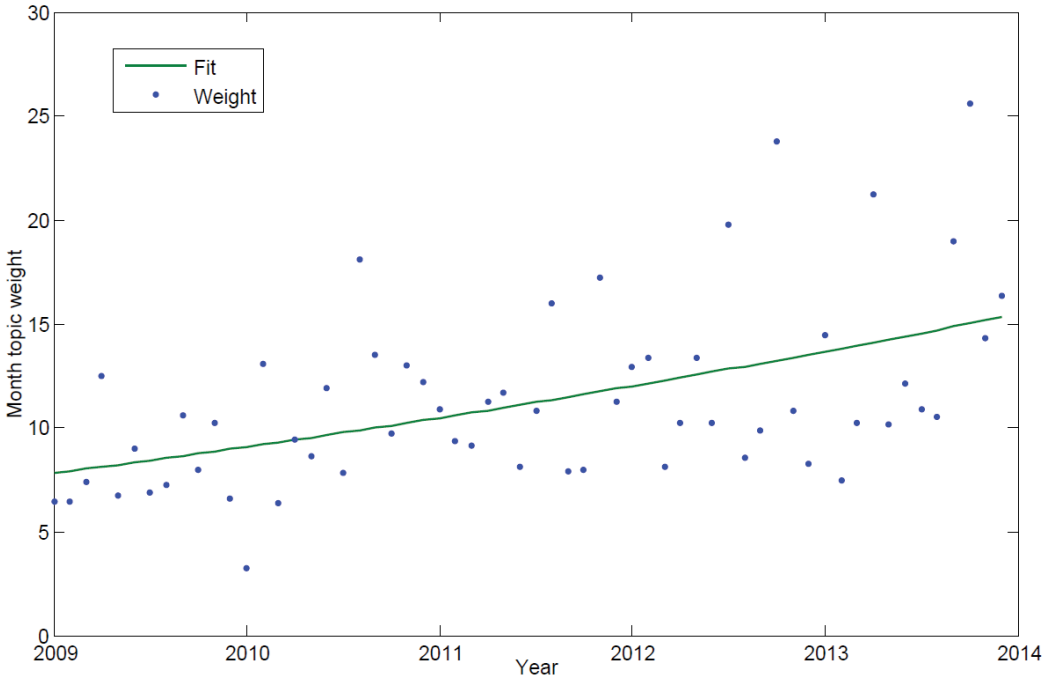


Figure 5-11. An example of the topic-based trend estimation of the theme "antibody"

In summary, this sub-section proposed an unsupervised topic change identification approach for patent mining based on the TTF method. 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. 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.

5.6 FUZZY NUMBER-BASED TECHNOLOGICAL TREND MEASUREMENT APPROACH

To achieve the aim of understanding valuable thematic knowledge from massive textual data, there are two main phases that need to be considered. The first one is automatically learning topics, and the second one is assisting further thematic evaluation using the estimated topics. This chapter has proposed the TTF method and the TTCI approach, to discover latent knowledge from the textual data of patent documents using LDA. However, for the second phase of assisting further thematic evaluation, simple term frequency may not be sufficient to support the measurement of development states that various topics have. Especially, the evaluation result in a real case is often expected to be a group of linguistic terms, such as 'growing', 'stable', 'have potential' and so forth, other than numerical values. Under such circumstances, approaches that are capable of discovering multiple words topics automatically and dealing with the vagueness of linguistic terms are needed.

In a real situation, as mentioned, the judgement on certain states, relations or tendency are often expressed by linguistic terms. To deal with the vagueness nature of these terms and manipulate imprecise values in real life, fuzzy sets were introduced by Zadeh (Zadeh 1965) as a classical notation of 'set' extension. Since fuzzy sets can effectively handle linguistic terms in measurement and deal with the uncertainty, in this research, this thesis propose a fuzzy number-based technological trend measurement (FTTM) approach to estimate and evaluate the topics hidden in a large volume of patent claims based on the proposed TTF method, which provides a fuzzy linguistic description to better explain the estimated trend status.

5.6.1 FRAMEWORK OF THE FTTM APPROACH

The framework of the FTTM approach is shown in Figure 5-12. After a target technological area is determined, the titles and claims of patent documents, and their corresponding patent ID and Issue dates, are extracted separately. On the one hand, the

textual data are passed to several segmentation and cleaning steps to remove all the punctuations, meaningless symbols, stop-words, general words used in claims and high frequency academic words; afterwards, LDA is utilized to generate latent topics from the prepared corpus. On the other hand, the result of topic modelling and the patent issue date information are gathered to serve the topic weight estimation. An annual weight matrix is then created to illustrate how the weights of all topics change over time. Then a group of linguistic terms are decided for depicting different states of topic development. According to the outcome of the topic weight estimation step, a fuzzy membership function can be created case by case. For each topic, a temporal-weight coefficient is calculated, which is associated with a set of linguistic terms to describe its development state over time. After choosing a suitable linguistic term set, fuzzy membership functions are created for each term. Finally, the temporal-weight coefficients can be transformed to membership vectors related to the linguistic terms in the step of fuzzy number-based technological trend measurement.

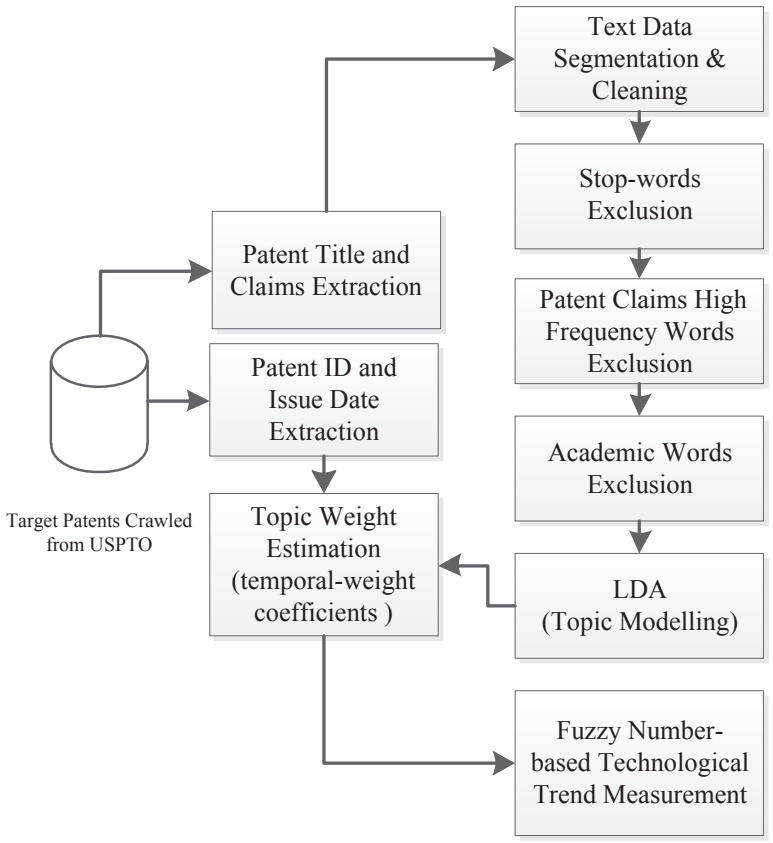


Figure 5-12. The framework of FTTM approach

5.6.2 FUZZY SET

Fuzzy sets are characterized by membership functions that assign each object to a grade of membership ranging from 0 to1 (Zhang 1998). The mathematical definition and notations of a fuzzy set is reviewed from Zhang (2009) and Lu (2009) as follows:

Definition 5-1. Let X be a universe discourse. A fuzzy set $\tilde{A}$ in X is characterized by its membership function $\mu_{\tilde{A}}(x)$

$$x \vdash \mu_{\tilde{A}}(x) \in [0,1]. \tag{5-1}$$

where the membership function $\mu_{\tilde{A}}(x)$ associates with each element x in X a real number in the interval of [0,1]. This real number is interpreted as the grade of x belongs to $\tilde{A}$. That is, the closer the value of $\mu_{\tilde{A}}(x)$ is to 1, the more it belongs to the fuzzy set $\tilde{A}$. A fuzzy set can be presented as a set of ordered pairs of elements x and its corresponding grade $\mu_{\tilde{A}}(x)$, which is noted by,

$$\tilde{A} = \{(x,\ \mu_{\tilde{A}}(x))|\ x \in X\}. \tag{5-2}$$

Definition 5-2. A fuzzy set $\tilde{A}$ in a universe of discourse X is convex if and only if for any $x_1, x_2 \in X$,

$$\mu_{\tilde{A}}(\lambda x_1 + (1-\lambda)x_2) \geqq min(\mu_{\tilde{A}}(x_1),\ \mu_{\tilde{A}}(x_2)), \tag{5-3}$$

where $\lambda \in [0,1]$.

Definition 5-3. A fuzzy set $\tilde{A}$ in a universe of discourse X is called a normal fuzzy set implying that there exists $x_0 \in$ X such that $\mu_{\tilde{A}}(x_0) = 1$.

Definition 5-4. A fuzzy number $\tilde{a}$ is a fuzzy subset on the space of real number R that is both convex and normal. A triangular fuzzy number $\tilde{a}$ can be defined by a triplet (a_0^L, a, a_0^R) and the membership function $\mu_{\tilde{a}}(x)$ is defined as:

$$\mu_{\tilde{A}}(x) = \begin{cases} \frac{(x-a_0^L)}{(a-a_0^L)}, a_0^L \le x \le a, \\ \frac{(a_0^R-x)}{(a_0^R-a)}, a \le x \le a_0^R, \\ 0, \qquad\qquad others. \end{cases} \tag{4}$$

Definition 5-5. A linguistic variable is a variable whose values are linguistic terms, such as 'good', 'stable', 'young' and 'old'.

In this research, three sets of linguistic terms are utilized to describe the developing states of estimated topics case by case. For a technological area showing strong growing potential, this research uses terms I={Steady (IS), Gradual Increasing (GI), Rapid Growing (RG)}; for technologies that have been comparatively mature, this research chooses a term set D={Rapid Declining (RD), Gradual Declining (GD), Steady (DS)}; for technologies show a wave-type of development, this research uses term set W={Declining (WD), Steady (WS), Growing (WG)}(Chen, Zhang, Lu, et al. 2015).

5.6.3 Topic Weight Estimation

After removing all commonly used words from the corpus, LDA is utilized to generate K topics in D documents in the prepared corpus. This research sets $\alpha = 0.5$ and $\beta = 0.1$, which are commonly used in LDA applications. Gibbs sampling is then applied to infer the needed distributions in LDA. As mentioned, this has provided K topics to express D documents totally, which gives a topic distribution matrix Θ with D rows and K columns. Each row of the topic distribution indicates how different topics distribute over a document in the corpus. Based on the proposed TTF method, a topic-based annual weight matrix is then generated for further content analysis.

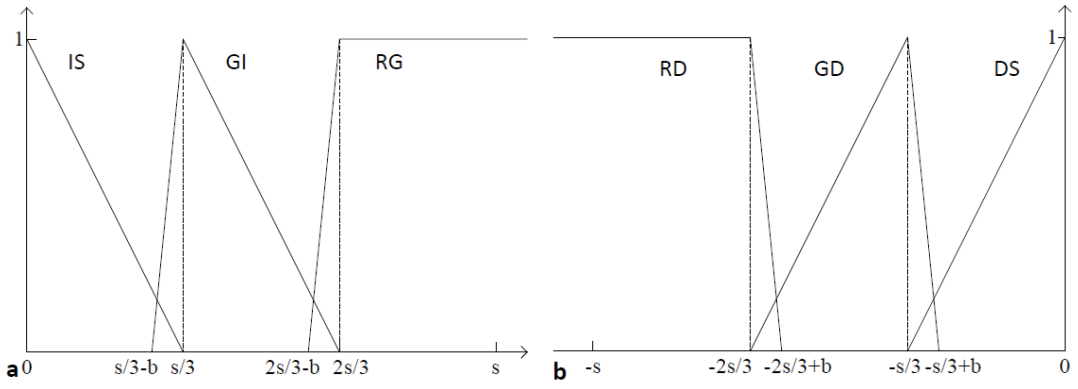
5.6.4 Fuzzy-based Technological Development Measurement

After the annual weight matrix is estimated, the weight changing rate of each topic is calculated using the method of least-squares, which fit each column in the matrix to a straight line. The first coefficient in one degree polynomial is then defined as the

temporal-weight coefficient of the corresponding topic. For all the generated topics, this research gets a temporal-weight coefficient vector TW, where $TW = (tw_1, tw_2, tw_3, \dots, tw_K)$ and tw_K stands for the coefficient of the K^{th} topic.

The vector TW is actually an attribute that associates with a set of linguistic terms to describe the development states of all estimated topics. It is known that, in existing research, the type of membership function that is suitable depends on the application context (Pedrycz & Gomide 1998; Wu, Zhang & Lu 2013). In this research, fuzzy membership functions can be inferred from the analysis of TW, or they may be determined by domain experts. Specifically, a domain value s is determined based on the domain of TW, where $s \geq |TW_{max} - TW_{min}|$. In addition, a value b is set in each function as a buffer. Because the TW will change for topics in a different technological area, the membership functions for the linguistic terms associated with the attribute will be changed accordingly. In this research, fuzzy numbers are used to present different terms.

The shapes of the membership functions of three sets of linguistic terms are illustrated in Figure 5-13, in which the sub-figure a shows the membership function of term set I; the sub-figure b illustrates the membership function of term set D; the sub-figure c provides the membership function of term set W.



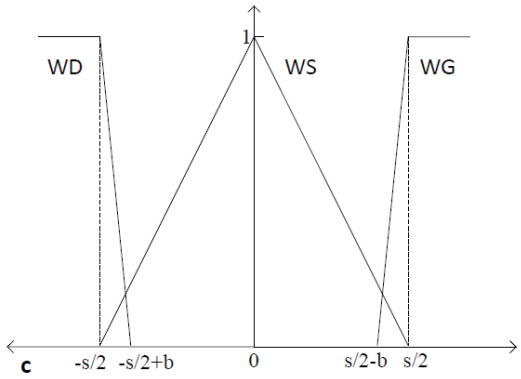


Figure 5-13. Linguistic terms and their membership functions

A mapping between *TW* and linguistic terms that describe topics development tendency can be built then. Each mapping denotes what development state a topic is on. The fuzzy numbers related to the linguistic terms IS, GI, DS, GD and WS are shown in Table 5-10.

Table 5-10. Linguistic terms and fuzzy numbers

Linguistic Terms	Fuzzy Numbers
IS	(0, 0, s/3)
GI	(s/3-b, s/3, 2s/3)
DS	(-s/3, 0, 0)
GD	(-2s/3, -s/3, -s/3+b)
WS	(-s/2, 0, s/2)

For linguistic terms RG, RD, WD and WG, however, it is hard to use fuzzy numbers to build the mapping. In this research, fuzzy membership functions $\mu_{RG}(x), \mu_{RD}(x), \mu_{WD}(x)$ and $\mu_{WG}(x)$ are defined respectively for these linguistic terms as follows:

$$\mu_{RG}(x) = \begin{cases} 0, & x < 2s/3 - b \\ \frac{x}{b} + 1 - \frac{2s}{3b}, & 2s/3 - b \le x \le 2s/3 \\ 1, & x > 2s/3 \end{cases} \quad (5\text{-}5)$$

$$\mu_{RD}(x) = \begin{cases} 1, & x < -2s/3 \\ -\frac{x}{b} + 1 - \frac{2s}{3b}, & -2s/3 \le x \le -2s/3 + b \\ 0, & x > -2s/3 + b \end{cases} \quad (5\text{-}6)$$

$$\mu_{WD}(x) = \begin{cases} 1, & x < -\mathrm{s}/2 \\ -\frac{x}{b} + 1 - \frac{s}{2b}, & -\mathrm{s}/2 \le x \le -\mathrm{s}/2 + \mathrm{b} \\ 0, & x > -\mathrm{s}/2 + \mathrm{b} \end{cases} \quad (5\text{-}7)$$

$$\mu_{WG}(x) = \begin{cases} 0, & x < \mathrm{s}/2 - \mathrm{b} \\ \frac{x}{b} + 1 - \frac{s}{2b}, & \mathrm{s}/2 - \mathrm{b} \le x \le \mathrm{s}/2 \\ 1, & x > \mathrm{s}/2 \end{cases} \quad (5\text{-}8)$$

After observing vector TW, it can be determined what type of technological development the target area has, and the most suitable term set selected from I, D and W, for further topic development states analysis.

5.6.5 Case Study of FTTM Approach by Patent Data

In order to demonstrate the effectiveness of the proposed FTTM approach, this case study chooses the solar cell area as a case study, to discover the development states of various detailed topics in it. This case study collects all the patents related to solar cell (ABST/"solar cell") and published during years 1985 to 2014 in USPTO (http://www.uspto.gov/). Totally, there are 3277 target patents covering 3271 utility patents, 5 re-issue patents and 1 statutory invention registration in the solar cell area. Their patent ID, titles, issue date and claims are crawled from USPTO and placed in a patent database for further processing. Patents ID and their issue time are put into one single document, while the claims and title for each patent constitute one document in the corpus, which totals 3277 documents in all. While volume cleaning, besides all the general words, the word ‘solar’ and ‘cell’ are also excluded, which are the highest frequency words in this area.

Before topic modelling, as mentioned, a number of parameters need to be set first, including the number of topics, α, β of Dirichlet distribution and the number of iterations for Gibbs sampling. This case study applies $K = 30$ with model hyper-parameters $\alpha = 0.5$, $\beta = 0.1$and 2000 iterations of Gibbs sampling to the target document collection, to balance the topical granularity, convenience of understanding, and the speed of processing. Totally, 34607 unique terms were obtained for the final corpus for topic modelling.

After topic modelling, this research obtained 30 latent semantic topics, in which each of them is presented by the top 20 ranked words and their corresponding probabilities. This research then generated the annual weight matrix and temporal-weight coefficients,TW, for all topics based on the topic distribution matrix. Table 5-11 shows the temporal-weight coefficients of all the 30 semantic topics sorted from the largest to the smallest. The larger the TW coefficient is for a topic, the more rapid it is developing.

Table 5-11. The Temporal-weight coefficients of topic 1 to topic 30

Topic	TW coefficient	Topic	TW coefficient
Topic 14	1.3451	Topic 4	0.2741
Topic 23	0.8643	Topic 20	0.2121
Topic 7	0.6482	Topic 13	0.1911
Topic 18	0.6354	Topic 1	0.1816
Topic 26	0.5562	Topic 17	0.1761
Topic 6	0.4509	Topic 29	0.1417
Topic 24	0.4416	Topic 30	0.1339
Topic 22	0.4009	Topic 19	0.1214
Topic 28	0.3684	Topic 15	0.1133
Topic 8	0.3430	Topic 9	0.1094
Topic 21	0.3343	Topic 5	0.1032
Topic 12	0.2990	Topic 3	0.0833
Topic 25	0.2965	Topic 11	0.0803
Topic 2	0.2863	Topic 27	0.0650
Topic 10	0.2776	Topic 16	0.0646

After observing the scope of the domain of TW, linguistic term set I is selected as the suitable one, since all the topics showed growing potential. For fuzzy membership functions creation, $s = 1.5$ and $b = 0.2$ were set in this case study. Figure 5-14 illustrates the final membership functions for term set I.

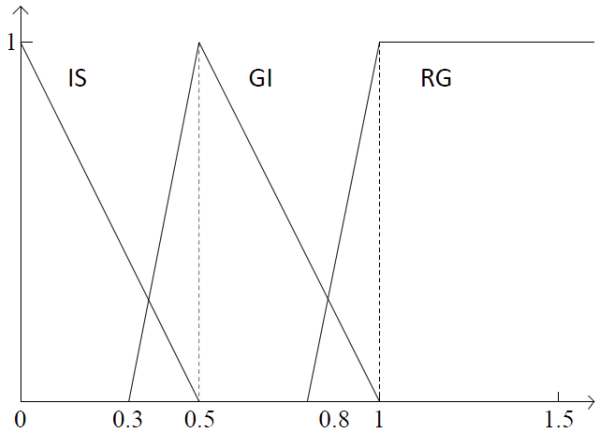


Figure 5-14. Membership functions of linguistic terms in term set I

To measure the development state of a semantic topic, this case study then then maps each topic coefficient to a linguistic term, by calculating its fuzzy membership degree vector (FMDV). Table 5-12 listed 15 topics of the final result of using FTTM approach on solar cell related patents, which have the highest temporal-weight values. For each topic, its fuzzy development measurement and 5 top ranked words are illustrated. There is a probability value indicating how probable is that a word belongs to its current topic. It can be seen from the form that topic No. 14 that relates to 'silicon substrate', and topic No.23 that concerns 'oxide polymer precursor' are the most rapid growing topics on the whole. Its development state can be measured as 'Rapid Growing'. Topics No. 7, No. 18, No. 26, No. 6, No. 24, No. 22 and No. 28 are measured as 'Gradual Increasing'. The rest of topics, including the topics does not show in the form, are all 'Steady', which means their growing potential are not as strong as the topics mentioned above.

Table 5-12. Development states measure result

Topic	Words	RROB	FMDV	Linguistic Terms
Topic 14	Silicon Substrate Surface Dopant metal	0.0564 0.0430 0.0320 0.0227 0.0154	(0,0,1)	RG (Rapid Growing)
Topic 23	metal oxide solution precursor polymer	0.0207 0.0182 0.0162 0.0141 0.0128	(0,0.27, 0.32)	RG (Rapid Growing)
Topic 7	conversion photoelectric glass organic paste	0.0511 0.0358 0.0316 0.0236 0.0186	(0, 0.7, 0)	GI (Gradual Increasing)

Topic 18	electrode plurality surface rear substrate	0.0846 0.0360 0.0274 0.0245 0.0139	(0, 0.73, 0)	GI (Gradual Increasing)
Topic 26	substituted formula unsubstituted compound alkyl	0.0230 0.0191 0.0180 0.0156 0.0127	(0, 0.89, 0)	GI (Gradual Increasing)
Topic 6	film transparent thin substrate conductive	0.1102 0.0517 0.0498 0.0432 0.0377	(0.1,0.75,0)	GI (Gradual Increasing)
Topic 24	power voltage circuit control signal	0.0705 0.0497 0.0220 0.0205 0.0174	(0.12,0.71,0)	GI (Gradual Increasing)
Topic 22	acid encapsulant copolymer composition weight	0.0220 0.0199 0.0137 0.0119 0.0118	(0.2,0.5,0)	GI (Gradual Increasing)
Topic 28	housing upper mounting body plate	0.0173 0.0136 0.0122 0.0116 0.0108	(0.26,0.34,0)	GI (Gradual Increasing)
Topic 8	conductive contact structure electrical surface	0.1054 0.0722 0.0431 0.0278 0.0267	(0.31,0.22,0)	IS (Steady)
Topic 21	surface radiation reflective optical concentrator	0.0586 0.0308 0.0308 0.0209 0.0176	(0.33,0.17,0)	IS (Steady)
Topic 12	system wireless source communication sensor	0.0246 0.0190 0.0180 0.0169 0.0165	(0.4, 0, 0)	IS (Steady)
Topic 25	surface upper semiconductor barrier silicon	0.0240 0.0193 0.0155 0.0154 0.0152	(0.41, 0, 0)	IS (Steady)
Topic 2	plurality elongated absorber internal transparent	0.0320 0.0252 0.0214 0.0159 0.0149	(0.42, 0, 0)	IS (Steady)
Topic 10	panel array plurality support located	0.0357 0.0253 0.0231 0.0153 0.0130	(0.44, 0, 0)	IS (Steady)

It is known that, as a green energy source, the solar cell has experienced very vigorous growth during the past decade. The consumption of solar cell technologies

continues to rise from the market's perspective (Tseng et al. 2011). The topic development measurement result on the whole shows the same tendency. More specifically, from the result it can be seen that the importance of topic "silicon substrate" and "oxide polymer precursor" are growing most rapidly, followed by topics on "photoelectric conversion", "plurality electrode substrate", "alkyl compound", "conductive thin substrate", "voltage and circuit", "acid encapsulant" and "housing (shell) mounting". These topics are more active than other topics in the solar cell area. Their developing potentials are stronger. The result of FTTM provides domain experts a direct view and a foresight in topics themselves and their corresponding development in a target area. In summary, FTTM can be used to obtain the main topics and their development states automatically from a large volume of documents, which makes it possible to set domain experts and analysts free from the heavy work of understanding and evaluating massive technological content. As patents and other technical indicators are still generating and accumulating in an increasing rate, approaches for automatically identifying and analysing latent topics will continue to be emphasized. The FTTM approach can also be used in understanding and evaluating scientific literatures.

5.7 SUMMARY

In summary, this chapter mainly deals with technological content analysis and trend forecasting tasks in real technology intelligence applications. This research proposed an empirical TTF method to generate topics from massive patent claims documentation, and then forecast their very own trends and different contribution levels for the patenting activities of the whole target area. This study then expanded afterwards technological content analysis based on TTF, presented a topic change identification approach to analyse the thematic evolution in a target technical area and a fuzzy number-based technological trend measurement approach to provide a fuzzy linguistic description to better explain the estimated trend status. All the above method and approaches can be used on scientific literature as well, to provide valuable topic-based knowledge and corresponding temporal patterns to facilitate further technological decision making.

CHAPTER 6
TOPIC DETECTION AND COMPREHENSIVE EVALUATION METHOD

6.1 INTRODUCTION

This research has applied topic modelling to provide promising applications for technological topics discovery and trend estimation, by successfully linked the semantic property and temporal characteristics of target technology indicators. However, topic models themselves cannot provide tight connections between the descriptive metadata of a corpus and the discovered topics. This chapter corresponds with the topic-based comprehensive analysis component in the proposed technology intelligence framework, proposes a topic detection and comprehensive evaluation (TDCE) method, which further improve the application of LDA in the scientometrics context, and access effective post-topic-modelling evaluation to gain a comprehensive overview of technological landscape and in-depth insight of technical details.

Specifically, to achieve the aim of improving the approaches and methods of understanding valuable thematic knowledge from massive technical documents, there are two main phases need to be considered. The first is to automatically detect latent topics from textual data more accurately. The second is to conduct further thematic evaluation using discovered topics or emblematic keywords. Although effort has been devoted to the existing research of implementing LDA for the first phase, there are still several limitations need to be considered a real world application. First of all, a fixed topic

number needs to be set before processing any documents, which raises a new requirement for researchers, analysts or decision makers to understand the possible latent thematic structure of the corpus before topic modelling (Suominen & Toivanen 2015). In addition, Gibbs sampling, one of the most commonly used approximate inference algorithm for parameter estimation, produces different results each time the LDA is executed, making determination of the final topic set even more complicated. Although LDA identifies topics efficiently, it can only evaluate a limited number and cannot link metadata to the target corpus (De Battisti, Ferrara & Salini 2015). Metadata, such as scientific papers' authors, affiliation, publication year, subject category and so forth, or patents' inventor, country, assignee, issue year, USPC and so on, gives information about more aspects of the target technology indicators. It summarizes the basic information about the target corpus, provides valuable records on how people classify and cite these documents, which can make tracking and understanding with specific data much easier. Thus, methods that can quantitatively evaluate and characterize further themes in the second phase are still in great demand.

To address the above three issues, the proposed TDCE method promises to discover and characterize the estimated topic after applying topic modelling to a target corpus. Aim to set user of topic-based technology intelligence free from making assumption on the number of latent topics, data likelihood and the subject category of scientific literature, or USPC for patents, are used to decide the final topic set, which better explains the actual semantic structure of a target corpus. Furthermore, by comprehensively considering the metadata of the corpus, this research proposes three topic evaluation indices to quantitatively characterizing the weight, developing trend and activeness for all the discovered topics, after executing LDA. This research also presents the topic-based citation indicator and corresponding citation distribution to feature the influence and potential usefulness of each topic. Topic evaluation maps are then applied to identify a number of prominent topics in the target area and the significant document supporting these topics. A case study with DSSCs data derive from searches in WoS is presented to demonstrate the proposed method.

The remainder of the chapter is organised as follows. Section 6.2 describes the full process of the proposed empirical method to topic evaluation. The parameters setting for the final topic set determination is presented in Section 6.3, followed by Section 6.4, which describes a series of topic evaluation indices. A case study using WoS papers from DSSC's data is then given in Section 6.5 to examine the method and then explains how to use it in the context of real scientific literature analysis. Finally, a summary of this chapter is given in Section 6.6.

6.2 METHODOLOGY FRAMEWORK

Motivated by the increasing number of research using LDA in tech mining and technology intelligence, facing the limitations of pre-setting the topic number and evaluating the discovered topics afterwards, the proposed TDCE method aims to conduct empirical topic extraction and evaluation on a large volume of target documents. In this section, a process framework of the proposed TDCE method is illustrated, which determines a final topic set that explains a target corpus more reasonably, then evaluates and characterize the estimated topic after executing LDA.

In Figure 6-1, the overall input, output and the whole process of the proposed TDCE method are examined. After a target technological area has been determined, the corresponding corpus is then built: for scientific literature, it consists of the titles and abstracts of all the WoS papers within the target technological scope; for patents, it covers the titles and claims of USPTO patents matching the search statement. For textual data, the title and abstract of a paper or the title and the claims of a patent, constitute one single txt document in the corpus, named with the paper's ISI article identifier or the patent issued number, as shown in step 1. Meanwhile, all metadata comprises one single file, as shown in step 2.

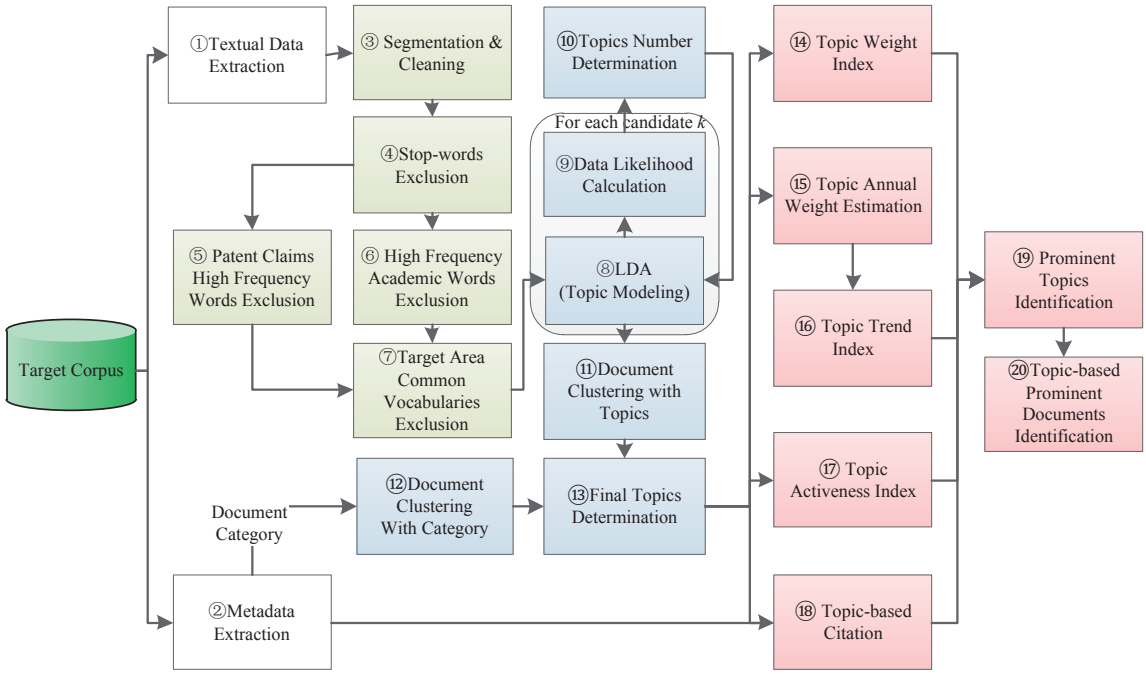


Figure 6-1. The framework of the topic detection and comprehensive evaluation method

In preparation for topic modelling, the textual data is first passed to a segmentation & cleaning step which removes all punctuation and meaningless symbols, as shown in steps 3 to 7. Stop words, high frequency academic words, high frequency patent claims words and common vocabulary used in target technical area are removed using a series of subsequent steps, marked in green in Figure 6-1. The LDA then generates latent topics and a distribution of topics of the corpus, which is the observation of the model. To overcome the limitation of assuming topic number and determining the final topic set, this research then applies data likelihood and pre-set document categories, to decide which model fits the given data best. All the detailed steps related to textual data topic modelling, from step 8 to step 13, are marked in blue. The round-cornered outline indicates replication. Once the final topic set has been confirmed, this research then evaluates the weights, developing trends and activeness of all the topics using the metadata and estimated topic distribution. Meanwhile, topic-based citations and their corresponding distribution are also calculated to quantitatively characterize the influence and contribution these estimated topics have made to the target area. The steps relate to comprehensive analysis on detected topics, from step 14 to step 20, are marked in red. Evaluation of the three indices and the topic-based citations provides the final number of

topics and the most significant documents supporting them, to assist researchers and analysts in understanding the landscape and analysing the interesting themes of technological investigation in the area.

6.3 TOPIC MODELLING

Following the research presented in Chapter 5, the TDCE method is built on the perspective of probabilistic topic modelling, which means the target corpus for a target technological area actually relates to multiple latent topics, each containing a probability distribution of words. This section focuses on how to set parameters and select the most representative topic set to better explain the observation, thus providing comparatively accurate semantic presentation of discovered topics for further evaluation and analysis.

6.3.1 PARAMETERS SETTING

Before topic modelling, the topic distribution in a target's corpus and a topic's word distribution are unknown, so assumptions must be drawn to determine the parameters K, α, β of LDA. As mentioned, α, β are the hyper-parameters of the Dirichlet distribution and have a smoothing effect on multinomial parameters (Heinrich 2005). This research lets $\alpha = 0.5$ and $\beta = 0.01$, since a relatively small value for β is expected to provide a fine-grained decomposition of the document collection (Griffiths & Steyvers 2004). Different values for the topic number K will impact final performance, that is, a higher value for K, will reduce the topical granularity but increase processing time. During implementation, the value of K needs to be decided case by case, to balance the number of multinomial distributions with the words that more reasonably represent the corpus and the time consumption.

This research uses a criterion named 'data likelihood' to reasonably select a suitable value of K. It evaluates how well an estimated model with a specific choice of K, fits the given data. As shown in Expression 6-1, the data log likelihood is proportional to the probability of the observation w under a model of z and φ, where D presents the overall documents, W_d denotes all the n unique words in document d, z is the topic assignment,

and φ stands for the distribution over vocabularies, which actually means the illustration of the topics themselves. The larger the data log likelihood value of $P(w|z,\varphi)$, the bigger the probability that the given data can be observed under a trail; then the result it produced fits the observation better, which also means the corresponding K is more reasonable. The detailed steps of calculating likelihood values for K selection are presented as Algorithm 6-1.

$$log\prod_{d}^{D}\prod_{n}^{W_d} P(w|z,\varphi) \tag{6-1}$$

Algorithm 6-1. Likelihood estimation for model with a specific choice of K topics

Input: the distribution over vocabularies φ and topic assignment z

Output: the log data likelihood value of the model, as sumOfAll

```
1       import the distribution over vocabularies φ
2       spamreaderPhi = φ
3       set rowNum = 0;
4       colNum = 0;
5       for row in spamreaderPhi:
6           colNum = len(row)
7           rowNum += 1
8       ldamap = numpy.zeros((colNum, rowNum));
9       rowIdx  = 0
10      for row in spamreaderPhi:
11          colIdx = 0
12              for item in row:
13                  try:
14                      number = float(item)
15                      except ValueError,e:
16                      break
17                  ldamap[colIdx][rowIdx] = number
18          colIdx +=1
19          rowIdx += 1
```

```
20      import the topic assignment z
21      spamreaderTassign = z;
22      sumOfAll = 0
23      for row in spamreaderTassign:
24          rowSum = 0
25                for item in row:
26                 r = item.split(':')
27                      if len(r) != 2:
28                      break
29          rowIdx = int(r[0])
30          colIdx = int(r[1])
31      rowSum += ldamap[rowIdx][colIdx]
32      sumOfAll += ldamap[rowIdx][colIdx]
33      print 'sum: ', sumOfAll
34      end
```

6.3.2 Final Topic Set Determination

Despite a suitable topic number, the final topic set still cannot be finalized, since Gibbs sampling in LDA produces different results each time even with the exactly same input and parameter settings. To overcome this limitation, this research uses the subject categories in WoS and USPC in patent classification system as document category labels to test candidate topic set, and help determine the final topic set for further analysis and evaluation.

The subject category in WoS is a predefined classification hierarchy assigned to a record by Thomson Reuters, built upon related journal scopes and the judgment of editors. It provides a general understanding of the technical domain that a scientific publication relates to and publications with similar topics should be assigned to the same category. In USPTO system, this research selects USPC as the criteria. Patents covering similar topics are usually assigned to a same main USPC. As a predefined classification hierarchy built on domain expert judgments, USPC provides a general understanding of

the technical domain of concern to one patent. For two main technology indicators, this research thus chooses subject categories and USPC to judge in multiple LDA experiments, and decide which trial is closest to the observation.

After parameters setting step, this research denotes the suitable topic number as k, the total number of papers in the final corpus is d. The unique document categories are then numbered and presented as $S = (s_1, s_2, s_3, \cdots, s_j, \cdots, s_d)$, where s_j is the subject category of the j^{th} document. This research first runs p times LDA with k topics. After performing each run, documents are clustered, with their estimated topic distributions Θ and category S, using the hierarchical clustering approach (Steinbach, Karypis & Kumar 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 p times experiments are used to measure the similarity between clustering results based on two different attributes (Halkidi, Batistakis & Vazirgiannis 2001). The three indices are listed in Equation 6-2, Equation 6-3, Equation 6-4, as follows:

$$J = a/(a + b + c)\,, \tag{6-2}$$

$$FM = a/\sqrt{r_1 \cdot r_2}\,, \tag{6-3}$$

$$F_\beta = \frac{(\beta^2+1)\cdot r_1 \cdot r_2}{\beta^2 \cdot r_1 + r_2}\,, \tag{6-4}$$

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 papers that belong to the same cluster of topics and to the same document categories, b is the number of papers that are assigned to the same cluster of topics but to different document categories, and c is the number of papers that are associated with different clusters of topics but to the same document categories. The topic modelling result that provides the highest index values is more reasonable than others.

6.4 Topic Evaluation Indices

In this section, by comprehensively considering metadata of a target corpus, three topic evaluation indices, which characterize the weight, developing trend and activeness of the discovered topics quantitatively, are proposed. This research also presents a topic-based citation indicator and corresponding citation distribution to feature the influence and potential usefulness of each topic. Topic evaluation maps are then applied to identify a number of prominent topics in the target area and the significant papers supporting them.

6.4.1 Topic Weight Index

After topic modelling, as mentioned, d documents were explained by k latent topics, in which each is presented as the number of top ranked words and those words' corresponding probabilities. The proportion of topics comprising the whole corpus, is shown as a topic distribution matrix $\Theta = (\vartheta_{ij})_{d\times k}$. Each row of the matrix indicates how different topics are distributed over one single document in the corpus, which makes the summation of every row equal to 1. The sum values of each column, however, are different. The larger the sum of a column, the bigger the proportion the corresponding topic has in the whole corpus, which also means it has a larger weight than other topics. Thus this research defines the weight of topics as $WI = (wi_1, wi_2, wi_3, \cdots, wi_k)$,

$$wi_k = \sum_{i=1}^{D} \vartheta_{ik}, \tag{6-5}$$

where wi_k indicates the sum-up of column k of Θ, D is the total number of documents in the target corpus.

6.4.2 Topic Trend Index

As discussed in Chapter 5, within the scope of the target corpus, no matter scientific papers or patents, are all published along a time line. The publication years have been considered as temporal labels for documents; while topic modelling, if documents are processed in ascending order, a topic distribution matrix in chronological order can be obtained. This sub-section continues to use the integration of the four elements: unique

document identifier, publication year, txt file and the topic distribution matrix, to link the topic's temporal attributes with its semantic properties, more importantly, that is, to link the metadata with the textual data.

This research then sums the group of elements in each column that was associated with target documents published in the same year; the total is the annual weight of the corresponding topic. Specifically, matrix $W_{m\times k}$ represents the annual weight of all estimated k topics that appeared during m years. Figure 6-2 depicts an example of the annual weight matrix with m rows and k columns to reveal the importance of the selected topics in different years.

	Topic 1	*Topic 2*	*Topic 3*	•••	*Topic k*
Year 1 weight	0.0109	1.6190	0.4411	•••	0.5875
Year 2 weight	0.0171	2.4317	1.6732	•••	0.8146
Year 3 weight	0.0282	2.7350	0.4394	•••	0.8069
⋮	⋮	⋮	⋮	•••	⋮
Year m weight	6.0745	16.5351	4.6622	•••	7.5101

Figure 6-2. An example of annual weight matrix

To define the trend index of topics, the publication years are first denoted with $t = (t_1, t_2, t_3, \cdots, t_m)$, where t_m indicates the m^{th} year. Once the annual weight matrix is complete, this research calculates the annual weight changes to estimate the developing trend of each underlying semantic topic. In a least-squares sense, the annual weight values of the k^{th} topic are fitted to a univariate quadratic polynomial,

$$\vartheta_{ik} = a_k t^2 + b_k t + c_k, \tag{6-6}$$

where ϑ_{ik} stands for the topic weight, and t represents the year $(1 < t \leq m)$. This research uses the coefficients $a_k \; and \; b_k$ to measure the developing trends of the corresponding topic, since a_k controls the speed of increase (or decrease) of the quadratic function, $-b_k/2a_k$ control the axis of symmetry. For instance, if coefficient a_k is positive and the symmetry is on the left of y-axis, it is considered the corresponding topic

has a growing trend. The greater a_k is, the faster the growth will be. In summary, the topic trend index is defined as $TI = (ti_1, ti_2, ti_3, \cdots, ti_k)$,

$$ti_k = (\vartheta_{ik} - b_k t - c_k)/t^2, \tag{6-7}$$

where ti_k equals to the coefficients a_k of the quadratic function that the annual weight is fitted to.

6.4.3 TOPIC ACTIVENESS INDEX

Besides the weight and trend indices, this research also interests in the activeness of the discovered topics as an essential criterion in evaluating how a topic is keeping up with the times. Before defining an indicator for activeness, the relationship between documents and topics needs to be narrowed from multi-to-multi to multi-to-one, so each document can be assigned to the topic with the highest probability. Mathematically, for document d, if T_d represents the main theme it serves, then the probability of d belonging to T_d is P_{T_d}, where $P_{T_d} = max(\vartheta_{dk})$. Likewise, for each topic, there is an array $\vec{D}$ to illustrate its corresponding documents, where $\vec{D}$ stores the unique identifier from the metadata of the document.

Therefore, for each topic, a group of documents can be identified, which mainly explain the topic and distribute along a timeline. This research then presents the number of papers published in each year that belong to each topic, as a matrix $PN = (pn_{ij})_{m \times k}$, where pn_{ij}denotes the number of papers explaining topic j and published in year i. An assumption is made that, the more recent-published papers a topic associated to, the more active it is. Thus the activeness weights of papers are proportional to the publication year $t = (t_1, t_2, t_3, \cdots, t_m)$, and increase year by year. In summary, the activeness index is defined as $AI = (ai_1, ai_2, ai_3, \cdots, ai_k)$, where ai_k is presented as follows:

$$ai_k = Sum(pn_{1k}, 2 \times pn_{2k}, 3 \times pn_{3k}, \cdots, i \times pn_{ik}, \cdots, m \times pn_{mk}). \tag{6-8}$$

6.4.4 TOPIC-BASED CITATION

Citation is a well-accepted metric to measure the quality of papers, journals and patents. According to previous research, the more citations a paper or patent carries, the more relevant and the more contribution it can be considered to have given to a scientific area. Moreover, the more likely it is to be cited itself (Baird & Oppenheim 1994; Oppenheim 1996; Radicchi & Castellano 2012). It follows that the more citations a journal, or in the case of this research, a topic receives, the greater its value or importance. This subsection aims to quantitatively characterize how influential and potentially useful a topic is.

As this research has already identified a group of target documents $\vec{D}$ mainly explained a topic (for topic j, $\overrightarrow{D_j}$ is used), the sum-up of documents' citations of $\vec{D}$ can be used to define the existing influence and potential usefulness of the topic. Specifically, if the citations of all the documents under topic j are set as $\overrightarrow{C_j}$, the topic-based citation of topic j can be denoted as $ct_j = \sum \overrightarrow{C_j}$, where ct_j indicates the sum of citations received by $\overrightarrow{D_j}$. This research goes on to estimate the citation distribution of those topics have accumulated much more citations than others over time, to identify the citation variation patterns of these most influential topics. The average number of citations per year is calculated to generate the citation distribution, since when each citation occurred cannot be accurately tracked using just metadata. The topic citation distribution can be presented as $CP = (cp_{ij})_{y \times k}$, where cp_{ij} indicates the estimated citations of topic j in year i. As the citations of all the documents under topic j is $\overrightarrow{C_j}$, particularly, this research set c_{ir} as the citied time of paper r in $\overrightarrow{D_j}$ (totally contain n paper) with publication year i, and the most recent year as mentioned is denote as m. Then cp_{ij} can be delineated as follows:

$$cp_{ij} = \sum_{n}^{1} \frac{c_{ir}}{(m+1-i)} + cp_{(i-1)j}. \tag{6-9}$$

An example of identifying high-cited topics and their corresponding citation distribution will be presented later in the case study of this chapter.

6.4.5 Topic-based Prominent Topics and Documents Identification Using Metadata

After characterizing the weight, trend, activeness and citation level of all topics with quantitative indices, topic-based evaluation maps to assist informative post-topic modelling analysis are drawn. A number of very significant topics that have high research values, and therefore greater impact, are now recognizable. Prominent topics are specifically explained with their top 20 ranked topic words constituting and corresponding probabilities. This research then labels each topic with its most relevant terms and link it to other metadata, to identify the total number of documents supporting it, the distribution across countries, and the most active affiliation or assignee, and so forth. Furthermore, the most highly-cited paper or patent, which usually represents the most reliable work in an area, is highlighted with its title, authors/inventor, affiliations/assignee, and number of citations to give a focal understanding of the whole theme.

6.5 Case study: Dye-sensitized Solar Cells Scientific Literature

This section provides a case study with one of the main target technology indicator, scientific literature, which plays an important role in profiling R&D and estimates scientific trends for potential innovation assistance. Specifically, to demonstrate the proposed method above, a case study with dye-sensitized solar cells (DSSCs) data derive from searches in WoS, to show how to apply the method in real context of scientific literature topic identification and analysis.

6.5.1 Data

This case study built and expanded a DSSCs target corpus with titles and abstracts of 12,435 scientific papers, which published between 1991, the year that DSSCs was first announced in Nature, and 2014. The title and the abstract of each paper constitute one

single text document in the corpus, named and identified by a unique ISI article identifier. The metadata of all publications each compose a single file. Metadata is parsed into publication year, country, authors, affiliations, subject category number, subject category and number of citations. Corpus details including the number of unique terms and subjects in each corresponding sub-collection are shown in Table 6-1. Note that the number of papers, terms, and subjects increased from the year 2009.

Table 6-1. The number of documents, terms and subjects of documents each year

Year	Doc NO.	Term NO.	Subject NO.	Year	Doc NO.	Term NO.	Subject NO.
1991	5	93	5	2003	175	1714	12
1992	6	119	5	2004	237	2078	13
1993	13	252	8	2005	268	2167	16
1994	10	163	5	2006	338	2617	13
1995	9	177	5	2007	427	3436	14
1996	15	243	8	2008	538	3735	17
1997	35	507	11	2009	740	5006	19
1998	35	589	10	2010	1104	6571	18
1999	49	688	10	2011	1506	8466	21
2000	87	989	11	2012	1889	9968	21
2001	101	1256	16	2013	2200	11781	17
2002	146	1567	13	2014	2502	13657	24

6.5.2 Scientific Literature Text Cleaning

Scientific literature contains plenty of technical terms and academic-related words that describe research outcomes. In order to maintain vocabularies with technical semantic meaning only, the terms in the target corpus are cleaned and consolidated, as the four steps after Textual Data Extraction that shown in Figure 6-1. First, titles and abstracts of scientific papers are segmented into a unique vocabulary list. Then all punctuation and non-alphabetic characters are removed, leaving only plain English terms. Additionally, this research uses a stop words list, publication-related thesauri and high frequency academic words list to remove meaningless terms that provide little or no contribution to the technological topics. Examples follow:

- Stop words (David et al. 2004) , such as *the, that, these*;

- Publication-related terms, names of organizations, governments, and companies (Zhang et al. 2014), such as *United States Abstract, Springer Science, Copyright*;
- Academic words that frequently used in scientific publications (Haywood 2003), such as *methodology, error, approach*;

Given this research wants the final model to return latent, but potentially useful concepts and topics, not general ideas, it then consolidates and excludes all the common words used in that scientific area. Before topic modelling, this research assembled a group of commonly used DSSCs terms and chemical compounds, e.g., dye-sensitized solar cell, DSSCs and DSSC, TiO2 and titanium-dioxide and applied a common DSSCs term thesauri to exclude words like solar, cell and DSSC.

6.5.3 Parameters Setting and Final Topic Set Determination

Before topic modelling, the hyper-parameters are set to $\alpha = 0.5$ and $\beta = 0.01$, since a smaller β will decompose the corpus into fine-grained topics (Griffiths & Steyvers 2004). This case study then calculated a data log likelihood value of $P(w|z, \varphi)$ for K values 10, 20, 30, 40, 50, 100, 200, 300, continuing in increments of 100 up to 1000, and with 2000 iterations of Gibbs sampling. The model with the highest log likelihood is selected to determine a suitable topic number. Figure 6-3 illustrates the estimates of $P(w|z, \varphi)$ log likelihood values against K values, i.e., the number of topics, and suggests that the model explains the observation better when $K = 400$.

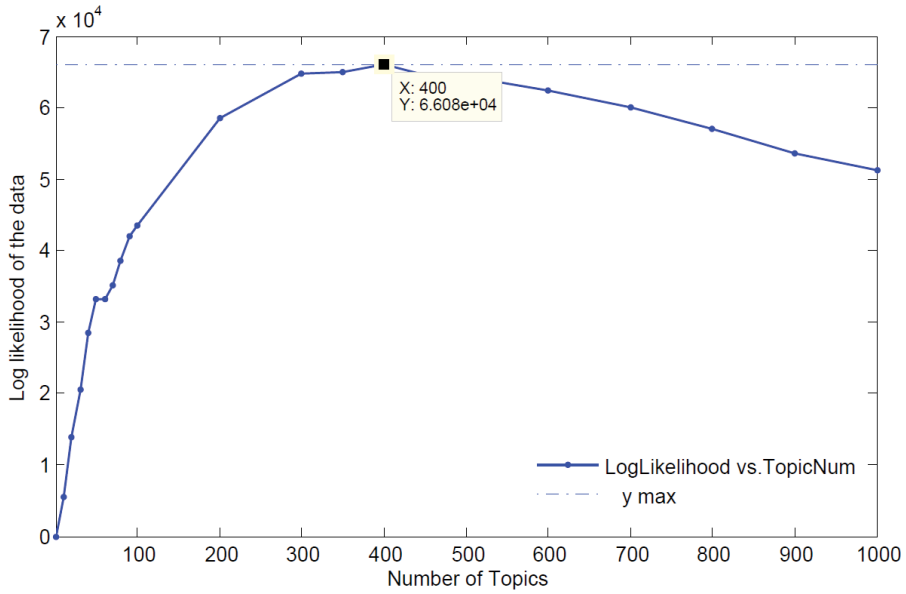


Figure 6-3. The log likelihood of the probability of the observation under models with different setting of the number of topics

This case study then performed 5 ($p = 5$) runs, with 2000 iterations of Gibbs sampling, to select the final topic set selection. After clustering all the abstract using both the generated topic distributions and their WoS subject categories, the similarity evaluation indices Folkes & Mallows, Jaccard, and F1 are calculated. Table 6-2 lists the index values from five runs with $K = 400$. The 5th run had the largest values of all three indices among the experimental trials and is highlighted in bold. This trial fits the observations better, thus those topics and parameters were selected as the final topic modelling result.

Table 6-2. Similarity evaluation result for the final topic set selection

Index	400_Topics_R1	400_Topics_R2	400_Topics_R3	400_Topics_R4	400_Topics_R5
F	0.5384	0.5397	0.5385	0.5375	0.5397
DJC	0.3684	0.3696	0.3684	0.3675	0.3695
FM	0.6039	0.6057	0.6040	0.6039	0.6059

6.5.4 Topic Evaluation Result

After determining the final topic set, the results showed 41,435 unique vocabularies explaining 400 latent semantic topics in the field of DSSCs. Each of them were represented by the top 20 ranked words in the topic their corresponding probability. The

topic distribution matrix is presented as a separate file named 'theta', where $\Theta = (\vartheta_{ij})_{12435 \times 400}$.This research then: evaluated all 400 topics using the proposed topic weight, trend and activeness indices; identified a group of prominent topics; and explored the citation distribution of these themes.

First, an annual weight matrix was calculated to illustrate annual weight changes in each topic. The sum of each column of the matrix shows the proportion of the corresponding topic in the topic collection. Figure 6-4 illustrates the annual weight growth of the top 10 most weighted topics. It shows that topic 28 has the greatest weight and is developing the fastest, followed by topic 67, topic 271 and so on. Actually, all 400 estimated topics, including the 10 topics shown in Figure 6-4 have positive trend index values. This also suggests that solar cells are a relatively young technological field that emerged in the 1990s.

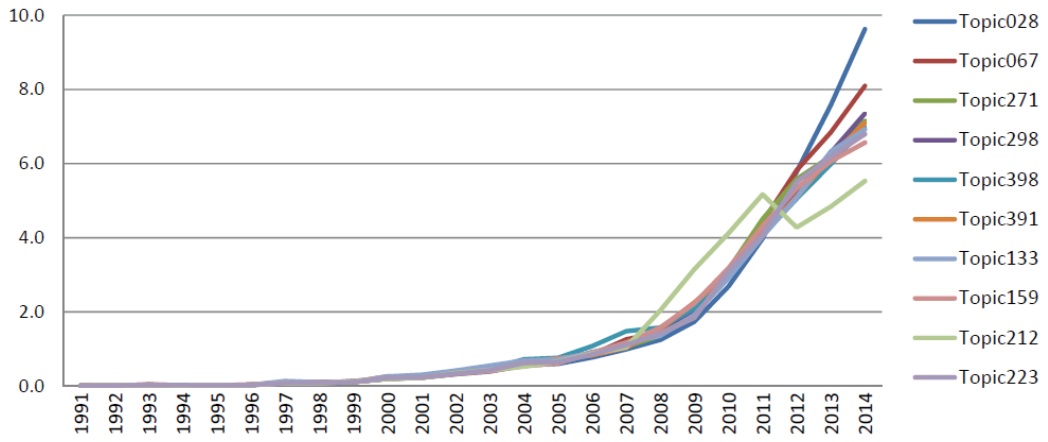


Figure 6-4. The annual weight growth of the top 10 high weighted topic

To continue evaluating all discovered topics from different perspectives, each paper is assigned the single topic with the maximum probability that the paper belongs to it. For the three mentioned indices, the topics with top 10 largest values are listed in Table 6-3. Topics are associated with two selected words from the word distributions, as topic labels to highlight their main themes.

Table 6-3. The top 10 topics with the highest evaluation indices values

Rank	Weight			Trend			Activeness		
	Topic	Label	Index	Topic	Label	Index	Topic	Label	Index
1	028	counter-electrodes platinum	37.1112	028	counter-electrodes platinum	0.0331	028	counter-electrodes platinum	4894
2	067	photocatalytic degradation	36.6833	067	photocatalytic degradation	0.0290	067	photocatalytic degradation	3742
3	271	td-dft calculations	34.9770	389	grapheme electrocatalytic	0.0263	047	nanowires single-crystalline	3066
4	298	xrd microscopy	34.6255	271	td-dft calculations	0.0262	391	porphyrin soret	3003
5	398	grapheme electrocatalytic	34.5680	298	xrd microscopy	0.0262	271	td-dft calculations	2657
6	391	porphyrin soret	34.4260	223	platinum H2PtCl6	0.0252	101	nio p-type	2555
7	133	ligands MLCT	34.0478	238	scattering submicrometer	0.0251	133	ligands MLCT	2441
8	159	nanotube anodization	34.0380	391	porphyrin soret	0.0251	159	nanotube anodization	2419
9	212	TiO2 Al2O3	33.8619	250	nanostructures one-dimension	0.0251	014	triphenylamine TPA	2233
10	223	platinum H2PtCl6	33.7870	210	thiophene cyanoacrylic	0.0249	119	sno2 f-doped	2233

Topic 28 characterized by words '*counter electrodes* (*CEs*)' and '*platinum* (*Pt*)', as well as topic 67 described by terms '*photocatalytic*' and '*degradation*' rank the highest in the three indices, which means from the perspective of topic modelling and evaluation, these two topics are more weighted, fast-developing and being recently focused, than other content in this area. Also, topic 271 '*td-dft* (*Time-dependent Density Functional Theory*) *calculations*' and topic 391 '*porphyrin soret*' are considered as two important topics as well, since they both showed in top 10 records of all the three indices. As expected, it can be observed that not all the topics with top-ranked activeness have large weight proportion, because fresh themes appear in relatively fewer papers than old ones. These comparatively young themes mainly covered technical content such as '*single-crystalline nanowires*', '*p-type NiO*', '*MLCT* (*Metal To Ligand Charge Transfer*)', '*nanotube anodization*', '*triphenylamine*' and '*f-doped SnO2*', as shown in the above table.

6.5.5 Topic-based Evaluation Maps

To further identify topics have attracted considerable research efforts and has a larger possibility that will continue being emphasized in future, the mentioned three indices are

considered comprehensively. As shown in Figure 6-5, after normalization, this research mapped the values for topic weight, trend and activeness on a scatter plot chart using Matlab, where the x-axis represents the trend index, the y-axis denotes the weight index, and the size of the dots indicates how up-to-date the corresponding topic is. The topics shown in the upper right corner of the figure are the topics considered to be more valuable of research prior. Additionally, the warmer colour and larger its dot, the more active it is.

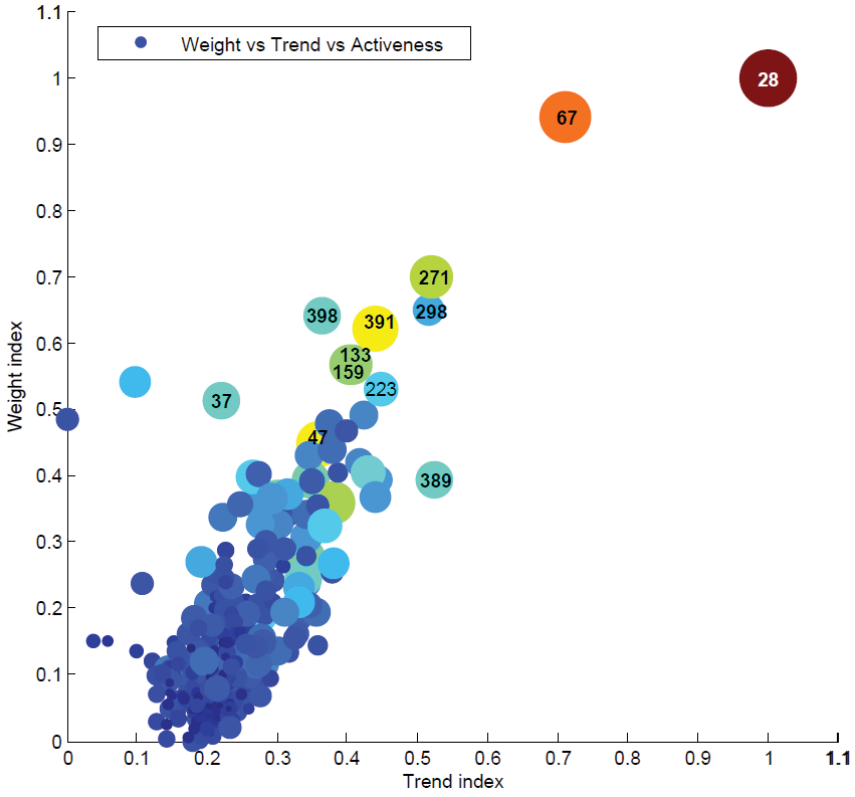


Figure 6-5. The topic-based evaluation map based on weight, trend and activeness characteristics of DSSCs corpus

Based on comprehensive evaluation of the three indices, topics 28, 67, 271, 298, 391, 398, 133, 159, 223, 47, 389 and 37 are the outstanding topics, with at least one index larger than 0.5 (half range). Table 6-4 provides the normalized indices and averaged values of the above topics in a quantitative statement, and highlights the five most related terms within each topic. This research scaled all the indices values to a range between $[0, 1]$ for convenience of comparison. Among all 400 topics that were generated to

uncover the latent thematic structure of Dye-sensitized Solar Cells, the 12 themes below were considered as candidates of the final topic set. For reading convenience, the top 10 ranked words of these themes and their corresponding probabilities are listed in Table 3 of the Appendix.

Table 6-4. Top 10 topics with the highest values (normalized) of the three indices

Topic	Weight	Trend	Activeness	Average	More relevant terms
028	1.0000	1.0000	1.0000	1.0000	counter-electrodes, platinum, electrocatalytic, platinum-free, Tafel
067	0.9397	0.7121	0.7625	0.8048	photocatalytic, degradation, photocatalyst, methylene, visible-light
271	0.6991	0.5198	0.5388	0.5859	time-dependent, td-dft, calculations, geometries, B3LYP
391	0.6214	0.4410	0.6101	0.5575	porphyrin, soret, red-shifted, YD2-o-C8, meso
133	0.5681	0.4067	0.4942	0.4897	ligands, MLCT, metal-to-ligand, bpy, heteroleptic
159	0.5667	0.4033	0.4897	0.4866	nanotube, anodization, NH4F, potentiostatic
298	0.6495	0.5169	0.2905	0.4856	xrd, microscopy, diffraction, x-ray, SEM
047	0.4479	0.3601	0.6231	0.4770	nanowires, single-crystalline, nanowire-based, htitanium-dioxide, nanoparticle
398	0.6414	0.3645	0.4031	0.4697	gel, quassolid-state, electrolytes, QS-DSSC, gelator
223	0.5313	0.4475	0.3381	0.4390	pt, H2PtCl6, homogeneously, electrocatalyst, spin-coating
389	0.3934	0.5251	0.3926	0.4370	grapheme, electrocatalytic, nanoplatelets, gnp, electro-catalytic
037	0.5130	0.2202	0.3926	0.3753	IPCE, photon-to-current, monochromatic, photon-to-electron, APCE

This case study then took topic-based citations into consideration to quantitatively define the existing influence and potential usefulness of the topic set. Since not all the papers are equally cited, a topic with a small number of papers, but with many more citations than another, indicates the topic includes several very highly cited papers and may suggest fundamental works in that theme.

For the 400 latent topics that have been estimated, Figure 6-6 visualizes the number of publications belonging to each of them and the number of corresponding topic-based citations. The x-axis denotes all the topics (topic 1 to topic 400), the y-axis illustrates the number of papers assigned to each topic, and most importantly, the size of dots shows the total citations a topic has received. In this figure, the top 12 topics with the highest citations are labelled (cited more than 4000 times). Topic 163 received the largest number - 14,258 - yet the number of papers that include that topic is quite small. As previously mentioned, this phenomenon implies the topic includes one or several fundamental studies. Within topic 163 the most highly-cited document related to photovoltaic and photocurrent-voltage is identified as A1991GL69600062, which was published in 1991 and has been cited 14,112 times. Likewise, topic 195 (*high-efficiency*

photoelectrochemical cell) and topic 177 (*redox iodide/triiodide*) show a similar phenomenon, and the documents 172150700053 (cited 6,024 times) and A1993LT17300063 (cited 4,376 times) can be correspondingly identified as the most representative papers in each theme.

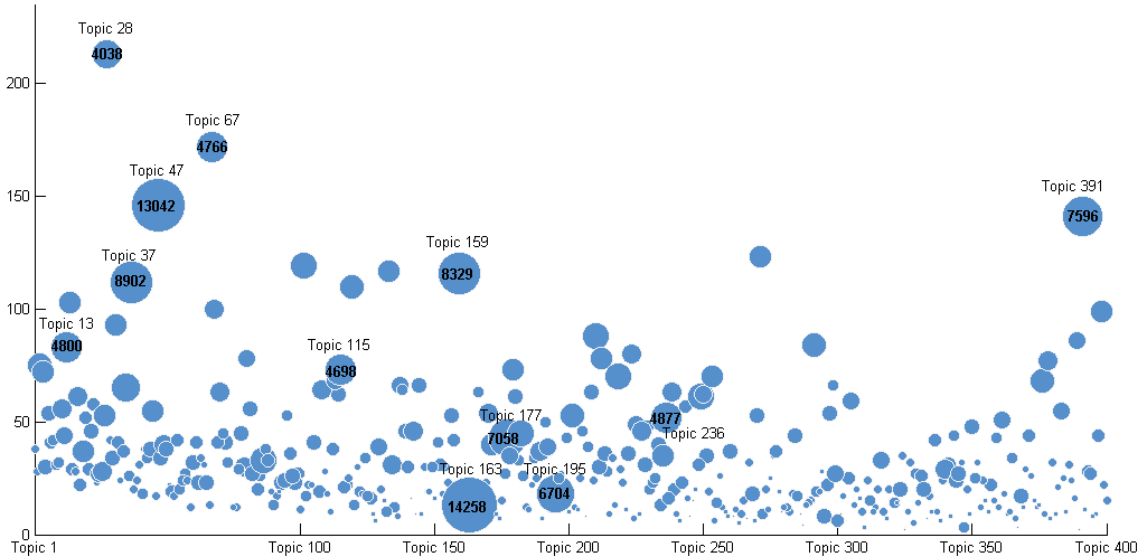


Figure 6-6. The topic-based citations map based on the total citation

After establishing the influence of the identified topics, this case study continued to evaluate them by simultaneously considering their topic-based citations and the three indices. Figure 6-7 illustrates the final topic-based evaluation map, based on the quantitative representation of the weight, trend, activeness and citations. In this 3D map, the x-axis explains the trend index, the y-axis denotes the weight index and the z-axis expresses the activeness of a topic. The colour temperature of dots represents the average values of the three indices, the warmer the colour, the larger the average value. Finally, the size of dots represents the total number of citations a topic has received.

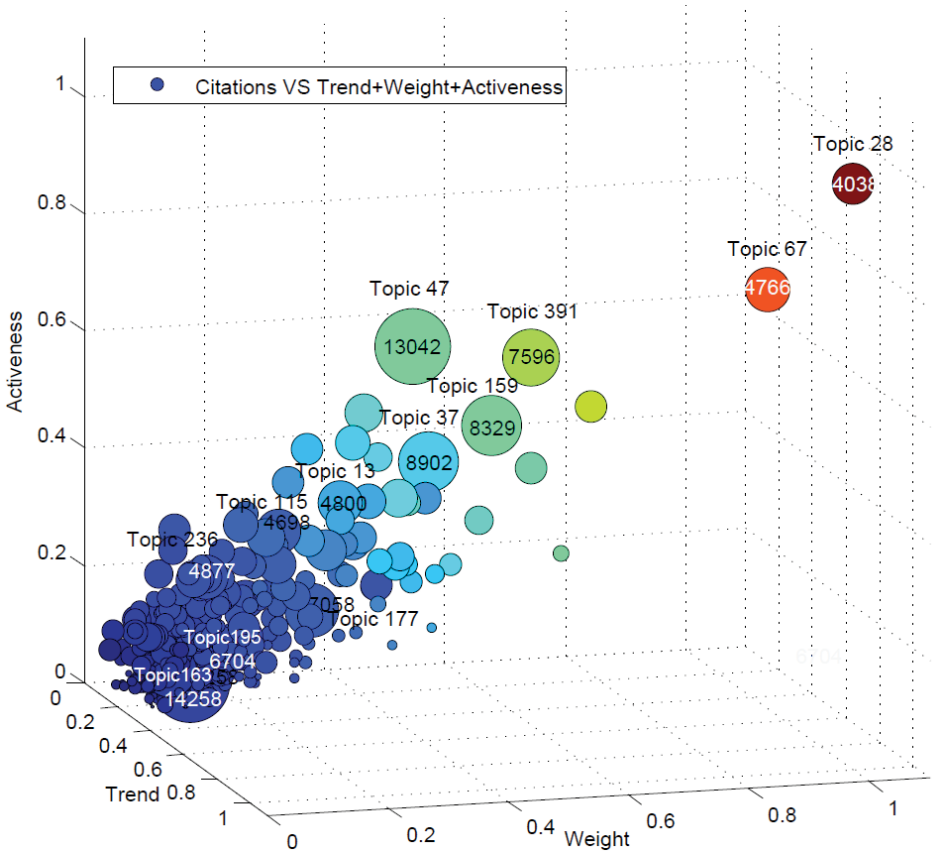


Figure 6-7. The topic evaluation map based on weight, trend, activeness and topic-based citations of DSSCs corpus

In the above figure, this research is interested in the dots close to the upper rightmost corner, with larger size and warmer colour. Six prominent topics stand out: topics 28, 67, 391, 47, 159, and 37. These topics are highly weighted, fast-developing and have been focused on recently, meanwhile, compared to other topics, they have performed stronger existing influence and potential usefulness.

6.5.6 Prominent Topics and Papers Analysis

Instead of using single terms or phrases, prominent topics are represented by probability distributions over words, explained with the top 20 ranked words constituting the topic and their corresponding probabilities specifically. A graphical illustration of the six prominent topics discussed in Figure 6-8. The thicker the connection between a term and the topic it belongs to, the higher possibility that this term is associated with the topic. Topic 028 is mainly characterized by the terms *counter electrodes, platinum,*

electrocatalytic, platinum-free, Tafel-polarization. Topic 067's key terms are *photocatalytic, degradation, photocatalyst, methylene, visible-light*. Topic 391: *porphyrin, soret, red-shifted, YD2-o-C8, meso*; Topic 159 can be characterized with *nanotube, anodization, NH4F, potentiostatic, free-standing*; Topic 047 is represented by terms nanowires, single-crystalline, nanowire-based, htitanium-dioxide, nanoparticle; Finally, Topic 037 can be explained by words *ipce, photon-to-current, monochromatic, photon-to-electron, apce*.

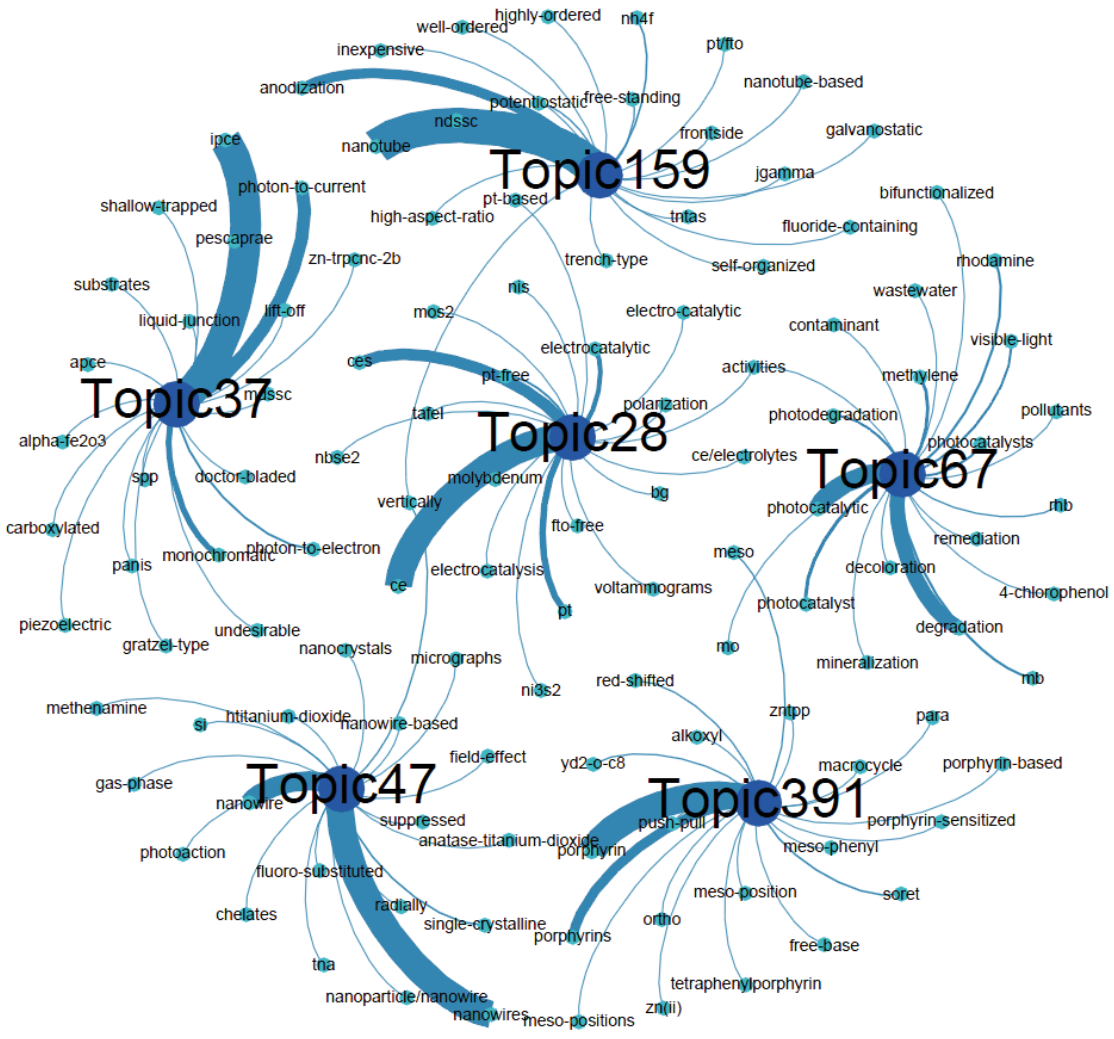


Figure 6-8. Graphical illustration of the main content of the 6 prominent topics

Continuing the examination, this case study investigated the cumulative citation distribution, of these six prominent topics, over time, to identify variations in their citation pattern. As shown in Figure 6-9, the earliest publication year of each topic from

the documents' assignment and the metadata of the DSSCs corpus are illustrated as follows: Topic 67 (*photocatalytic degradation*) and topic 37 (*photon-to-electron*) – 1997; topic 391(*porphyrin soret*) – 2000; topic 28 (*counter electrodes*) and topic 47 (*single-crystalline nanowires*) – 2004; and topic 159 (*nanotube anodization*) – 2005, and visualized the citation distribution with approximate curves in Figure 6-9. It can be observed that topic 47 (*single-crystalline nanowires*) and topic 391(*porphyrin soret*) are the two with the most rapid citation growth, while topic 37's (*photon-to-electron*) growth is relatively slow.

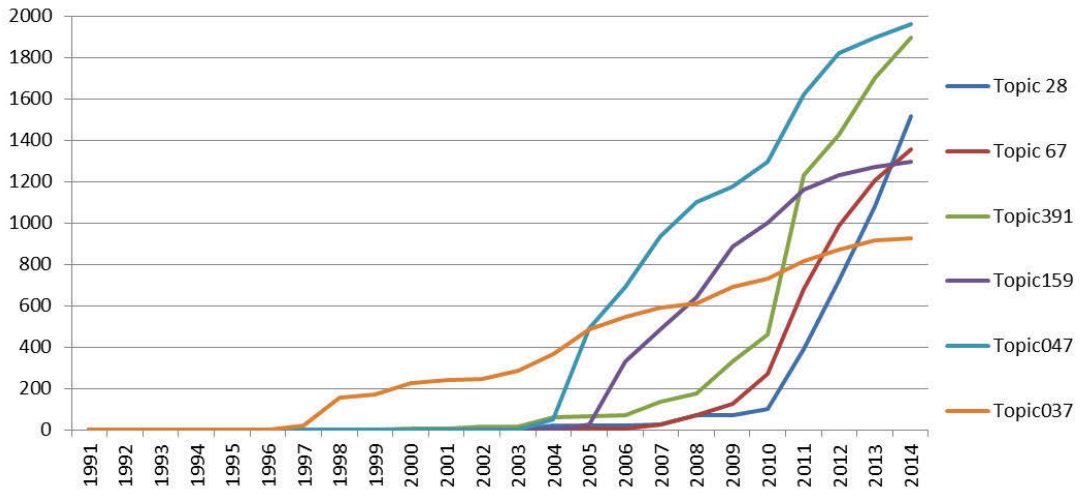


Figure 6-9. The cumulative citation distribution of DSSCs corpus from year 1991 to 2014

Closer examination of each of these topics reveal there are a number of very relevant papers contributing to their main thematic structure. To analyse detected topics across-the-board and gain more comprehensive insight, this case study continued to identify the number of the scientific papers, the most active affiliation, the distribution across countries, the most highly cited paper, as well as the authors and affiliates of the paper, to give a focal understanding of the whole theme. Table 6-5 illustrates the total paper number and the most active affiliation that the 6 prominent topics covered, where MAA means "most active affiliation".

Table 6-5. Detected topics and the publications they covered

Topic	Content label	Total paper No.	Most active affiliation	Paper NO. MAA published
Topic 028	counter electrodes	213	Dalian Univ Technol	18
Topic 067	photocatalytic degradation	172	Chinese Acad Sci	14
Topic 391	porphyrin soret	141	KYOTO Univ	10
Topic 159	nanotube anodization	116	Penn State Univ	9
Topic 047	single-crystalline nanowires	146	Sun Yat Sen Univ	5
Topic 037	photon-to-electron	112	Chinese Acad Sci	10

Six most active affiliations correspondingly belong to China, Japan and USA, which are also three top contributory countries in publication country distribution, as shown in Figure 6-10. China published the most paper for all 6 topics, especially for topic 28, counter electrodes, and topic 67, photocatalytic degradation. It contributed 59% and 54% of the publications in the scope of these two topics, in which Dalian University of Technology and Chinese Academy of Sciences play a very important role. USA, Japan, South Korea and Taiwan are also top contributory countries. USA focused more on topics porphyrin soret, nanotube anodization and single-crystalline nanowires; Japan concentrated more on topics porphyrin soret and photon-to-electron; South Korea interested more in research of counter electrodes, photocatalytic degradation and photon-to-electron; Taiwan mainly focused on topic of porphyrin soret.

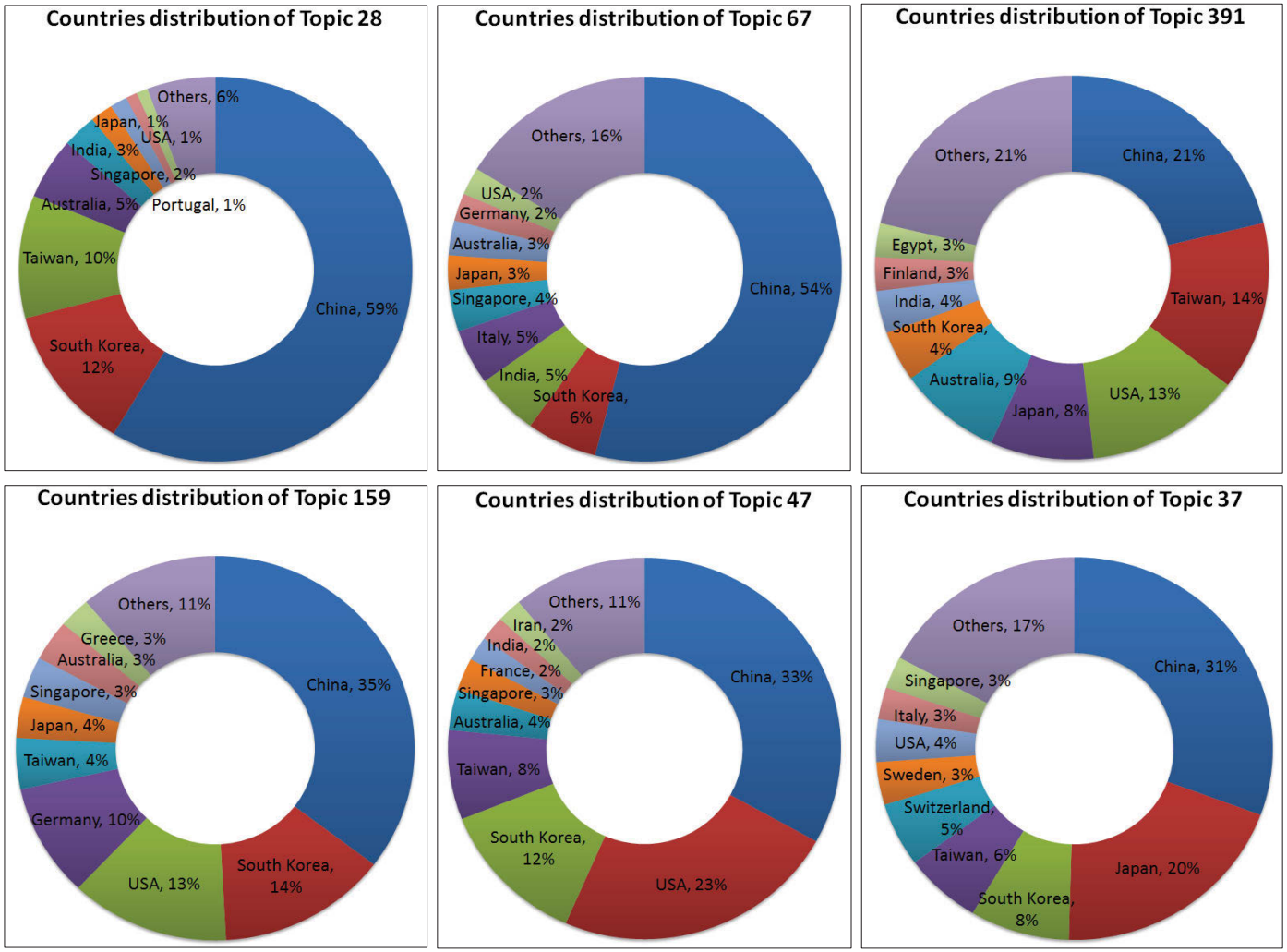


Figure 6-10. Publications country distribution of DSSCs corpus from year 1991 to 2014

After analysing the publication quantity to present research effort that has been devoted to a certain topic, eventually, the most highly-cited paper serving each theme can be detected, to highlight the most reliable and representative work in the area. Table 6-6 lists these most contributory papers for the 6 mentioned topics, presents their ISI article identifier (SCI accession number), title, authors, affiliations, and number of citations, to give a focal understanding of the whole theme.

Table 6-6. The most contributory papers of the prominent topic set

Topic	ISI article identifier	Title	First author	Affiliation	Citations
028	000254426600002	Counter electrodes for DSC: Application of functional materials as catalysts	Murakami, T N	Ecole Polytech Fed Lausanne	308
067	000283141300015	Semiconductor-mediated photodegradation of pollutants under visible-light irradiation	Chen, C C	Chinese Acad Sci	453
391	000296494700044	Porphyrin-Sensitized Solar Cells with Cobalt (II/III)-Based Redox Electrolyte Exceed 12 Percent Efficiency	Yella, A	Natl Chiao Tung Univ	2636
159	000235532700014	Use of highly-ordered TiO2 nanotube arrays in dye-sensitized solar cells	Mor, G K	Penn State Univ	1416
047	000229502700011	Nanowire dye-sensitized solar cells	Law, M	Univ Calif Berkeley	3453
037	000076362900042	Solid-state dye-sensitized mesoporous TiO2 solar cells with high photon-to-electron conversion efficiencies	Bach, U	Swiss Fed Inst Technol	2184

6.5.7 DISCUSSION

Since its introduction to the field of tech mining, LDA has been used as an efficient tool for key content extraction and thematic analysis. It will continue to be emphasized in this field due to the ability of estimating the ‘co-occurrence’ of words and semantically meaningful outcomes. In such circumstances, a full process including beforehand topic modelling preparation and afterwards evaluation is very helpful in providing a comprehensive overview of technological landscape, and assisting researchers and analysts to identify focal themes and publications supporting these themes. The topic extraction and evaluation method presented in this paper, accordingly, provides a practical method for determining the final topic set to explain a scientific literature corpus, when using LDA; and a set of indices, based on empirical research, to

quantitatively measure the discovered topics. The most significant and potentially useful themes in the area of interest are identified automatically and in an empirical way.

6.6 SUMMARY

This chapter proposed a topic detection and comprehensive evaluation method, based on LDA, aims to select the most representative topic set to explain a target corpus of technology indicator, as well as link metadata with the estimated topics to provide valuable evaluation on various themes. The three indices presented in this chapter comprehensively bring the metadata of publications into consideration, and quantitatively characterize the weight, developing trend and activeness of all discovered topics after executing the LDA. Topic-based citation and the corresponding citation distribution are also designed to feature the influence and potential usefulness of each topic. This research can provide a better and more comprehensive overview of the scientific research landscape, which can help the users of comprehensive-analysis technology intelligence to identify the most prominent topics and representative documents in their area of interest.

CHAPTER 7
CONCLUSIONS AND FURTHER STUDY

This chapter concludes the whole thesis and its contributions to the field of technology intelligence and technological decision-making support. It also provides some further research directions in this area.

7.1 CONCLUSIONS

This study is motivated by an awareness of practical issues in technology intelligence. The continuous technological advances are producing a wealth of information regarding technology development in both the public and private domain. Facing the increasingly severe information overload problem, the concept and tools of technology intelligence have been widely developed for various application domains.

Even though existing research on technology intelligence has gained considerable attention and undergone developments, one of the big challenges is that, the frameworks and applications of previous studies lack a comprehensive perspective on trend analysis of the detailed content within an area. In addition, single keywords and their ranking alone, are too general or ambiguous to represent complex concepts and their corresponding temporal patterns. In summary, systematic post-processing, forecasting and evaluation on both content analysis and trend identification outputs are still in great demand, for diverse and flexible technological decision support and opportunity discovery. This research attempts to address these issues.

The main contributions of this study are as follows:

(1) It proposes a three-dimensional schematic structure of the semantic properties and temporal characteristics of semi-structured technology indicators, and measures the two features quantitatively (to achieve Objective 1) as presented in Chapter 3.

Facing the difficulties in integrating the semantic property and temporal characteristic in existing research, this thesis uses 'topics' to represent the semantic property of all target technology indicators, and uses 'trend segments' and 'trend turning points' to define the temporal characteristic. It then provides a three-dimensional structure to clarity the relationship of the two important features. The entire knowledge presentation is considered as a semantic space with a temporal dimension. The different development levels of concepts in the semantic space can be quantitatively measured in this way. This knowledge structure provides decision makers with more comprehensive awareness of technological content and trend, at the same time setting a strong foundation for this research.

(2) It proposes a topic-based technology intelligence framework that systematically processes and analyses technological publication count sequence, textual data and metadata of typical semi-structured technology indicators (to achieve Objective 2) as presented in Chapter 3.

Technology indicators, represented by scientific literature and patent documents, hold explicit technical information and hidden knowledge that indicates technological concepts, themes, trend and related R&D activities, which are significant resources for decision making support, early warning signals or trend pointers. To take the full advantage of technology indicators, this research develops a topic-based technology intelligence framework to systematically process technological publication count sequence, textual data and metadata, to improve the framework and applications of existing technology intelligence. The framework includes three main functionalities: (1) trend identification functionality; (2) topic discovery functionality; (3) comprehensive topic evaluation functionality.

(3) It proposes a data-driven technological trend analysis method to capture the underlying trend patterns of technological publication activity and further technology forecasting (to achieve Objective 3) as presented in Chapter 4.

This research proposes a data-driven technological trend analysis method with technological trend identification, analysis and forecasting functionalities. From a data-driven perspective, it develops an innovative solution for empirical technology trend patterns identification and future trend forecasting by quantitatively identifying and depicting the concept "trend" with trend segments. In addition, it overcomes the limitations of model choosing and upper limits estimating which is experienced by existing empirical technology forecasting approaches. Thirdly, it can be used to learn valuable trend patterns from historical patent counts record, then can use the learned trend turning points and trend segments to predict future technology trend.

(4) It proposes a topic-based technological forecasting method, which can be used to uncover the latent topics and temporal trends underlying massive technical documents (to achieve Objective 4) as presented in Chapter 5.

To integrate the temporal trend patterns and semantic topics quantitatively, based on the outcomes of Chapter 3 and Chapter 4, this research proposes a topic-based technological forecasting method to discover and estimate the trends for specific topics underlying large volumes of patent claims using LDA. This method can effectively generate latent topics from massive technical documents, as well as estimate their very own trend and different contribution levels to the publication activities, thus providing valuable topic-based knowledge and corresponding temporal patterns to facilitate further technological decision making. This research then expands afterwards technological content analysis based on this method, presents a topic change identification approach and a fuzzy number-based technological trend measurement approach. The former one builds a topic change identification model to analyse the thematic evolution in a target technical area; the later one provides a fuzzy linguistic description to better explain the estimated trend status.

(5) It proposes an empirical topic detection and evaluation method to discover and characterize the estimated topic after applying topic modelling to a target technical corpus (to achieve Objective 5) as presented in Chapter 6.

This research proposes an empirical topic detection and evaluation method to set researchers, analysts and other decision makers free from making assumption on the number of latent topics, by using data likelihood and the aid of subject category for scientific literature, USPC for patents, to decide the final topic set, which better explains the actual semantic structure of the target corpus. In addition, this research links metadata with the estimated topics to provide valuable evaluation on various themes in depth and presents three indices, which comprehensively bring the metadata of publications into consideration, and quantitatively characterize the weight, developing trend and activeness of all discovered topics. Finally, it conducts topic-based citation and the corresponding citation distribution analysis to feature the influence and potential usefulness of each topic.

To demonstrate the proposed topic-based technology intelligence framework and all the related methods, this thesis presented case studies with Australian patents, United States patents and scientific papers from Web of Science database. Experimental results showed that the proposed framework and methods are well-suited in dealing with semi-structured technology indicators analysis, and can provide valuable topic-based knowledge to facilitate further technological decision making with good performance.

In practice, this research can be used to assist enterprises, venture capitalists and universities in understanding the underlying technological topics distribution and trends in target areas with much less human intervention. From the perspective of industrial use, it can provide entrepreneurs with topic-based summarization of a certain industry for proposal preparation or strategy making support. From the perspective of academic use, it can provide a full understanding of topic-based research strength for seeking potential funding and collaboration opportunities. In conclusion, the framework and methods proposed in this research can serve as a tool of automatically uncovering the thematic structure of massive technical data, estimate the detailed developing trend of each detected topic, and link the metadata with discovered topics, thereby assisting decision

making for potential opportunity identification, technical strategy formation, and so forth, for both industrial and academic purposes.

7.2 FURTHER STUDY

There are still some limitations of the current study:

(1) Even though this research proposes a framework and a set of methods for topic-based technology intelligence, it concentrates on the level of algorithms and methodologies. A physical technology intelligent system is still needed.

(2) While topic modelling parameters are optimized, the approach used in this research is time consuming, since a series of candidate topics numbers need to be tested.

(3) This research is built on the assumption of at least one target corpus in the technological area of interests is prepared. The corpuses preparation effects all the afterwards calculation and analysis.

The theoretical and applied research on technology intelligence will continue to be emphasized, for its ability to assist decision makers with learning knowledge in massive data efficiently. This research can be fully advanced in the following aspects:

(1) In the future, a physical topic-based technology intelligent system will be developed, to accomplish the full process of data preparation, cleaning, feature extraction, forecasting, topic modelling, and evaluation.

(2) From this research, it can be noticed that the latent topics underlying in massive technological indicators are correlated with each other. Future study will continue to examine the dynamic evaluation of the latent topics, using approaches of concept drift and other machine learning methods.

(3) The proposed framework and methods can be further modified to work on other backgrounds, more than just technological analysis. Since this research works for semi-structured inputs, all suitable documents can be considered as the target of applying the

framework or methods of this research. Future study will keep exploring its possible implementation in other fields.

REFERENCES

Artstein, R. & Poesio, M. 2008, 'Inter-coder agreement for computational linguistics', *Computational Linguistics*, vol. 34, no. 4, pp. 555-96.

AusPat, *IP Australia*, <http://pericles.ipaustralia.gov.au/ols/auspat/advancedSearchPage.do>.

Baird, L.M. & Oppenheim, C. 1994, 'Do citations matter?', *Journal of Information Science*, vol. 20, no. 1, pp. 2-15.

Barker, D. & Smith, D.J. 1995, 'Technology foresight using roadmaps', *Long Range Planning*, vol. 28, no. 2, pp. 21-8.

Basberg, B.L. 1983, 'Foreign patenting in the US as a technology indicator: the case of Norway', *Research Policy*, vol. 12, no. 4, pp. 227-37.

Baskurt, O.K. 2010, 'Time series analysis of publication counts of a university: what are the implications?', *Scientometrics*, vol. 86, no. 3, pp. 645-56.

Batagelj, V. & Mrvar, A. 2002, 'Pajek—analysis and visualization of large networks', *Graph Drawing*, Springer, pp. 477-8.

Bengisu, M. 2003, 'Critical and emerging technologies in Materials, Manufacturing, and Industrial Engineering: A study for priority setting', *Scientometrics*, vol. 58, no. 3, pp. 473-87.

Bengisu, M. & Nekhili, R. 2006, 'Forecasting emerging technologies with the aid of science and technology databases', *Technological Forecasting and Social Change*, vol. 73, no. 7, pp. 835-44.

Bishop, C.M. 2006, *Pattern recognition and machine learning*, springer.

Blei, D. & Lafferty, J. 2006, 'Correlated topic models', *Advances in neural information processing systems*, vol. 18, p. 147.

Blei, D.M. 2012, 'Probabilistic topic models', *Communications of the ACM*, vol. 55, no. 4, pp. 77-84.

Blei, D.M., Griffiths, T.L. & Jordan, M.I. 2010, 'The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies', *Journal of the ACM (JACM)*, vol. 57, no. 2, p. 7.

Blei, D.M., Ng, A.Y. & Jordan, M.I. 2003, 'Latent dirichlet allocation', *the Journal of machine Learning research*, vol. 3, pp. 993-1022.

Campbell, R.S. 1983, 'Patent trends as a technological forecasting tool', *World Patent Information*, vol. 5, no. 3, pp. 137-43.

Camus, C. & Brancaleon, R. 2003, 'Intellectual assets management: from patents to knowledge', *World Patent Information*, vol. 25, no. 2, pp. 155-9.

Carrillo, M. & González, J.M. 2002, 'A new approach to modelling sigmoidal curves', *Technological Forecasting and Social Change*, vol. 69, no. 3, pp. 233-41.

Cascini, G. & Russo, D. 2007, 'Computer-aided analysis of patents and search for TRIZ contradictions', *International Journal of Product Development*, vol. 4, no. 1, pp. 52-67.

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 Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 39, no. 1, pp. 80-92.

Chang, P.-C., Liao, T.W., Lin, J.-J. & Fan, C.-Y. 2011, 'A dynamic threshold decision system for stock trading signal detection', *Applied Soft Computing*, vol. 11, no. 5, pp. 3998-4010.

Chen, H., Zhang, G. & Lu, J. 2013, 'A Time-Series-Based Technology Intelligence Framework by Trend Prediction Functionality', *2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC),* , pp. 3477-82.

Chen, H., Zhang, G., Lu, J. & Zhu, D. 2014, 'A Two-Step Agglomerative Hierarchical Clustering Method for Patent Time-Dependent Data', in F. Sun, T. Li & H. Li (eds), *Foundations and Applications of Intelligent Systems*, vol. 213, Springer Berlin Heidelberg, pp. 111-21.

Chen, H., Zhang, G., Lu, J. & Zhu, D. 2015, 'A fuzzy approach for measuring development of topics in patents using Latent Dirichlet Allocation', *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE),* , pp. 1-7.

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*, vol. 26, no. 2, pp. 345-53.

Chen, H., Zhang, Y., Zhang, G., Zhu, D. & Lu, J. 2015, 'Modeling technological topic changes in patent claims', *2015 Portland International Conference on Management of Engineering and Technology (PICMET)*, IEEE, pp. 2049-59.

Chen, Y.-H., Chen, C.-Y. & Lee, S.-C. 2011, 'Technology forecasting and patent strategy of hydrogen energy and fuel cell technologies', *International Journal of Hydrogen Energy*, vol. 36, no. 12, pp. 6957-69.

Ciarli, T., Coad, A. & Rafols, I. 2015, 'Quantitative analysis of technology futures: A review of techniques, uses and characteristics', *Science and Public Policy*.

Coates, V., Farooque, M., Klavans, R., Lapid, K., Linstone, H.A., Pistorius, C. & Porter, A.L. 2001, 'On the future of technological forecasting', *Technological Forecasting and Social Change*, vol. 67, no. 1, pp. 1-17.

Cozzens, S., Gatchair, S., Kang, J., Kim, K.S., Lee, H.J., Ordóñez, G. & Porter, A. 2010, 'Emerging technologies: quantitative identification and measurement', *Technology Analysis & Strategic Management*, vol. 22, no. 3, pp. 361-76.

Cunningham, S.W., Porter, A.L. & Newman, N.C. 2006, 'Special issue on tech mining', *Technological Forecasting and Social Change*, vol. 73, no. 8, pp. 915-22.

Daim, T. & Suntharasaj, P. 2009, 'Technology diffusion: forecasting with bibliometric analysis and Bass model', *Foresight*, vol. 11, no. 3, pp. 45-55.

Daim, T.U. 2015, 'Technology analytics: Enhancing technology assessment with technology intelligence', *Technological Forecasting and Social Change*.

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*, vol. 78, no. 2, pp. 197-8.

Daim, T.U., Rueda, G., Martin, H. & Gerdsri, P. 2006, 'Forecasting emerging technologies: Use of bibliometrics and patent analysis', *Technological Forecasting and Social Change*, vol. 73, no. 8, pp. 981-1012.

David, L., Yiming, Y., Tony, G.R. & Fan, L. 2004, *SMART stopword list*, MIT Press, <http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop>.

De Battisti, F., Ferrara, A. & Salini, S. 2015, 'A decade of research in statistics: a topic model approach', *Scientometrics*, vol. 103, no. 2, pp. 413-33.

Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W. & Harshman, R.A. 1990, 'Indexing by latent semantic analysis', *JAsIs*, vol. 41, no. 6, pp. 391-407.

Dereli, T., Baykasoğlu, A., Durmuşoğlu, A. & Durmuşoğlu, Z.D.U. 2011, 'Enhancing technology clustering through heuristics by using patent counts', *Expert Systems with Applications*, vol. 38, no. 12, pp. 15383-91.

Ding, Y. 2011, 'Topic-based PageRank on author cocitation networks', *Journal of the American Society for Information Science and Technology*, vol. 62, no. 3, pp. 449-66.

Ernst, H. 1997, 'The use of patent data for technological forecasting: the diffusion of CNC-technology in the machine tool industry', *Small Business Economics*, vol. 9, no. 4, pp. 361-81.

Faust, K. & Schedl, H. 1983, 'International patent data: Their utilization for the analysis of technological developments', *World Patent Information*, vol. 5, no. 3, pp. 144-57.

Florita, A.R. & Henze, G.P. 2009, 'Comparison of short-term weather forecasting models for model predictive control', *HVAC&R Research*, vol. 15, no. 5, pp. 835-53.

Gallupe, R.B. 2007, 'The tyranny of methodologies in information systems research', *ACM SIGMIS Database*, vol. 38, no. 3, pp. 20-8.

Garfield, E., Malin, M.V. & Small, H. 1983, 'Citation data as science indicators'.

Gerybadze, A. 1994, 'Technology forecasting as a process of organisational intelligence', *R&D Management*, vol. 24, no. 2, pp. 131-40.

Griffiths, T.L. & Steyvers, M. 2004, 'Finding scientific topics', *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228-35.

Griffiths, T.L., Steyvers, M., Blei, D.M. & Tenenbaum, J.B. 2004, 'Integrating topics and syntax', *Advances in neural information processing systems*, pp. 537-44.

Griliches, Z. 1990, 'Patent Statistics as Economic Indicators: A Survey', *Journal of Economic Literature*, vol. 28, no. 4, pp. 1661-707.

Gupta, B., Kumar, S., Sangam, S. & Karisiddappa, C. 2002, 'Modeling the growth of world social science literature', *Scientometrics*, vol. 53, no. 1, pp. 161-4.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M. 2001, 'On clustering validation techniques', *Journal of Intelligent Information Systems*, vol. 17, no. 2-3, pp. 107-45.

Haywood, S. 2003, *Academic Vocabulary*, Nottingham University, Nottingham University2014, <http://www.nottingham.ac.uk/alzsh3/acvocab/wordlists.htm>.

Heinrich, G. 2005, *Parameter estimation for text analysis*, Fraunhofer IGD, Darmstadt, Germany.

Herce, J.L. 2001, 'WIPO patent information services for developing countries', *World Patent Information*, vol. 23, no. 3, pp. 295-308.

Hofmann, T. 1999, 'Probabilistic latent semantic indexing', *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 50-7.

Holst, H.v., Nguyen, H. & Wikander, J. 2010, *Innovation driven research education: volume I : an introduction*, PIEp, Product Innovation Engineering Program, Stockholm, Sweden.

Hongshu Chen, G.Z., Jie Lu 2013, 'A Time-series-based Technology Intelligence Framework by Trend Prediction Functionality', *2013 IEEE International Conference on System, Man and Cybernetics, 2013IEEE SMC.*, vol. In press, IEEE, Manchester, UK.

Hossain, M.D., Moon, J., Kang, H.G., Lee, S.C. & Choe, Y.C. 2011, 'Mapping the dynamics of knowledge base of innovations of R&D in Bangladesh: triple helix perspective', *Scientometrics*, vol. 90, no. 1, pp. 57-83.

Jabłońska-Sabuka, M., Sitarz, R. & Kraslawski, A. 2014, 'Forecasting research trends using population dynamics model with Burgers' type interaction', *Journal of Informetrics*, vol. 8, no. 1, pp. 111-22.

Jeong, D.-H. & Song, M. 2014, 'Time gap analysis by the topic model-based temporal technique', *Journal of Informetrics*, vol. 8, no. 3, pp. 776-90.

Jun, S. & Uhm, D. 2010, 'Technology forecasting using frequency time series model: Bio-technology patent analysis', *Journal of Modern Mathematics and Statistics*, vol. 4, no. 3, pp. 101-4.

Keogh, E., Chu, S., Hart, D. & Pazzani, M. 2001, 'An online algorithm for segmenting time series', *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, IEEE, pp. 289-96.

Keogh, E., Chu, S., Hart, D. & Pazzani, M. 2004, 'Segmenting time series: A survey and novel approach', *Data mining in time series databases*, vol. 57, pp. 1-22.

Keogh, E. & Kasetty, S. 2003, 'On the need for time series data mining benchmarks: a survey and empirical demonstration', *Data Mining and Knowledge Discovery*, vol. 7, no. 4, pp. 349-71.

Keogh, E. & Pazzani, M. 1998, 'An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback', *KDD-98 Proceedings.*, pp. 239-41.

Kerr, C.I., Mortara, L., Phaal, R. & Probert, D. 2006, 'A conceptual model for technology intelligence', *International Journal of Technology Intelligence and Planning*, vol. 2, no. 1, pp. 73-93.

Ketata, I., Sofka, W. & Grimpe, C. 2015, 'The role of internal capabilities and firms' environment for sustainable innovation: evidence for Germany', *R&D Management*, vol. 45, no. 1, pp. 60-75.

Kim, D. & Oh, A. 2011, 'Topic Chains for Understanding a News Corpus', in A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, vol. 6609, Springer Berlin Heidelberg, pp. 163-76.

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 Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 396-407.

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

Lai, K.-K. & Wu, S.-J. 2005, 'Using the patent co-citation approach to establish a new patent classification system', *Information Processing & Management*, vol. 41, no. 2, pp. 313-30.

Lee, C., Cho, Y., Seol, H. & Park, Y. 2012, 'A stochastic patent citation analysis approach to assessing future technological impacts', *Technological Forecasting and Social Change*, vol. 79, no. 1, pp. 16-29.

Lee, C., Jeon, J. & Park, Y. 2011, 'Monitoring trends of technological changes based on the dynamic patent lattice: A modified formal concept analysis approach', *Technological Forecasting and Social Change*, vol. 78, no. 4, pp. 690-702.

Lee, H.-j., Lee, S. & Yoon, B. 2011, 'Technology clustering based on evolutionary patterns: The case of information and communications technologies', *Technological Forecasting and Social Change*, vol. 78, no. 6, pp. 953-67.

Lee, S., Kim, M.-S. & Park, Y. 2009, 'ICT Co-evolution and Korean ICT strategy—An analysis based on patent data', *Telecommunications Policy*, vol. 33, no. 5, pp. 253-71.

Lee, S., Lee, H.-j. & Yoon, B. 2012, 'Modeling and analyzing technology innovation in the energy sector: Patent-based HMM approach', *Computers & Industrial Engineering*, no. 0, p. http://www.sciencedirect.com.ezproxy.lib.uts.edu.au/science/article/pii/S0360835211003755.

Lichtenthaler, E. 2007, 'Managing technology intelligence processes in situations of radical technological change', *Technological Forecasting and Social Change*, vol. 74, no. 8, pp. 1109-36.

Liu, S. & Chen, C. 2013, 'The differences between latent topics in abstracts and citation contexts of citing papers', *Journal of the American Society for Information Science and Technology*, vol. 64, no. 3, pp. 627-39.

Lu, J., Zhu, Y., Zeng, X., Koehl, L., Ma, J. & Zhang, G. 2009, 'A linguistic multi-criteria group decision support system for fabric hand evaluation', *Fuzzy Optimization and Decision Making*, vol. 8, no. 4, pp. 395-413.

Lukins, S.K., Kraft, N.A. & Etzkorn, L.H. 2010, 'Bug localization using latent Dirichlet allocation', *Information and Software Technology*, vol. 52, no. 9, pp. 972-90.

Luo, L. & Chen, X. 2013, 'Integrating piecewise linear representation and weighted support vector machine for stock trading signal prediction', *Applied Soft Computing*, vol. 13, no. 2, pp. 806-16.

Lv, P.H., Wang, G.-F., Wan, Y., Liu, J., Liu, Q. & Ma, F.-c. 2011, 'Bibliometric trend analysis on global graphene research', *Scientometrics*, vol. 88, no. 2, pp. 399-419.

Mailänder, L. 2013, *Topic 3: Claims*, Riyadh, Saudi Arabia.

Mandal, S. & Prabaharan, N. 2006, 'Ocean wave forecasting using recurrent neural networks', *Ocean engineering*, vol. 33, no. 10, pp. 1401-10.

Manning, C.D. & Schütze, H. 1999, *Foundations of statistical natural language processing*, MIT press.

Martino, J.P. 1993, *Technological forecasting for decision making*, McGraw-Hill, Inc.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D. & Barton, D. 2012, 'Big data', *The management revolution. Harvard Bus Rev*, vol. 90, no. 10, pp. 61-7.

Mcauliffe, J.D. & Blei, D.M. 2008, 'Supervised topic models', *Advances in neural information processing systems*, pp. 121-8.

Mohammadi, E., Thelwall, M., Haustein, S. & Larivière, V. 2015, 'Who reads research articles? An altmetrics analysis of Mendeley user categories', *Journal of the Association for Information Science and Technology*, vol. 66, no. 9, pp. 1832-46.

Morris, S., DeYong, C., Wu, Z., Salman, S. & Yemenu, D. 2002, 'DIVA: a visualization system for exploring document databases for technology forecasting', *Computers & Industrial Engineering*, vol. 43, no. 4, pp. 841-62.

Mortara, L., Kerr, C.I., Phaal, R. & Probert, D.R. 2009a, 'Technology intelligence practice in UK technology-based companies', *International Journal of Technology Management*, vol. 48, no. 1, pp. 115-35.

Mortara, L., Kerr, C.I., Phaal, R. & Probert, D.R. 2009b, 'A toolbox of elements to build technology intelligence systems', *International Journal of Technology Management*, vol. 47, no. 4, pp. 322-45.

Ngai, E.W., Cheng, T.E., Au, S. & Lai, K.-h. 2007, 'Mobile commerce integrated with RFID technology in a container depot', *Decision Support Systems*, vol. 43, no. 1, pp. 62-76.

Nishijima, Y., Anzai, T. & Sengoku, S. 2013, 'Application of bibliometric analysis to market analysis', *Technology Management in the IT-Driven Services (PICMET), 2013 Proceedings of PICMET '13:*, pp. 2365-77.

Niu, L. 2009, 'The cognition-driven decision process for business intelligence: a model and techniques', University of Technology, Sydney.

Niu, L., Lu, J. & Zhang, G. 2008, 'Cognitive orientation in business intelligence systems', *Intelligent Decision and Policy Making Support Systems*, Springer Berlin Heidelberg, pp. 55-72.

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*, no. In press.

OECD 2005, *OECD handbook on economic globalisation indicators*, OECD, Paris.

OECD 2008, *Compendium of Patent Statistics*.

Oppenheim, C. 1996, 'Do citations count? Citation indexing and the Research Assessment Exercise (RAE)', *Serials*, vol. 9, no. 2, pp. 155-61.

Palit, A.K. & Popovic, D. 2005, *Computational intelligence in time series forecasting: theory and engineering applications*, Springer-Verlag New York Inc.

Pedrycz, W. & Gomide, F. 1998, *An introduction to fuzzy sets: analysis and design*, Mit Press.

Philips, F. 1999, 'A method for detecting a shift in a trend', *Management of Engineering and Technology, 1999. Technology and Innovation Management. PICMET'99. Portland International Conference on*, vol. 1, IEEE, p. 238 vol. 1.

Piateski, G. & Frawley, W. 1991, *Knowledge discovery in databases*, MIT press.

Porter, A.L. 2009, 'Tech mining for future-oriented technology analyses', in J.C. Glenn & T.J. Gordon (eds), *Futures Research Methodology - Version 3.0*.

Porter, A.L. & Cunningham, S.W. 2004, *Tech mining: exploiting new technologies for competitive advantage*, vol. 29, John Wiley & Sons.

Porter, A.L., Cunningham, S.W., Banks, J., Roper, A.T., Mason, T.W. & Rossini, F.A. 1991, *Forecasting and management of technology*, Wiley, New York.

Porter, A.L. & Detampel, M.J. 1995, 'Technology opportunities analysis', *Technological Forecasting and Social Change*, vol. 49, no. 3, pp. 237-55.

Publishing, O. 2010, *Measuring Innovation: A New Perspective*, OECD Pub.

Radicchi, F. & Castellano, C. 2012, 'Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts', *Journal of Informetrics*, vol. 6, no. 1, pp. 121-30.

Rao, I.R. & Srivastava, D. 2010, 'Growth of journals, articles and authors in malaria research', *Journal of Informetrics*, vol. 4, no. 3, pp. 249-56.

Robinson, D.K.R., Huang, L., Guo, Y. & Porter, A.L. 2011, 'Forecasting Innovation Pathways (FIP) for new and emerging science and technologies', *Technological Forecasting and Social Change*, no. 0.

Safavieh, E., Andalib, S. & Andalib, A. 2007, 'Forecasting the Unknown Dynamics in NN3 Database Using a Nonlinear Autoregressive Recurrent Neural Network', *2007 International Joint Conference on Neural Networks, IJCNN 2007.* , pp. 2105-9.

Safdari Ranjbar, M. & Tavakoli, G.R. 2015, 'Toward an inclusive understanding of technology intelligence: a literature review', *Foresight*, vol. 17, no. 3, pp. 240-56.

Salton, G. & Buckley, C. 1988, 'Term-weighting approaches in automatic text retrieval', *Information Processing & Management*, vol. 24, no. 5, pp. 513-23.

Salton, G., Wong, A. & Yang, C.-S. 1975, 'A vector space model for automatic indexing', *Communications of the ACM*, vol. 18, no. 11, pp. 613-20.

Savioz, P. 2004, *Technology Intelligence: Concept Design and Implementation in Technology Based SME's*, Palgrave macmillan.

Schuh, G., Brakling, A. & Apfel, K. 2014, 'Identification of requirements for focused crawlers in technology intelligence', *Management of Engineering & Technology (PICMET), 2014 Portland International Conference on*, IEEE, pp. 2918-23.

Shapiro, R.D. 1985, *Toward effective supplier management: international comparisons*, Division of Research, Harvard Business School.

Sheikh, N., Gomez, F.A., Yonghee, C. & Siddappa, J. 2011, 'Forecasting of advanced electronic packaging technologies using bibliometric analysis and Fisher-Pry diffusion model', *Technology Management in the Energy Smart World (PICMET), 2011 Proceedings of PICMET '11:*, pp. 1-20.

Sheldon, J.G. 2001, *How to Write a Patent Application*.

Srivastava, A.N. & Sahami, M. 2009, *Text mining: Classification, clustering, and applications*, CRC Press.

Steinbach, M., Karypis, G. & Kumar, V. 2000, 'A comparison of document clustering techniques', *KDD workshop on text mining*, vol. 400, Boston, pp. 525-6.

Steyvers, M. & Griffiths, T. 2007, 'Probabilistic topic models', in D.M. T. Landauer, S. Dennis, and W. Kintsch (ed.), *Latent Semantic Analysis: A road to meaning*, Laurence Erlbaum.

Suominen, A. & Toivanen, H. 2015, 'Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification', *Journal of the Association for Information Science and Technology*, pp. n/a-n/a.

Tang, J., Wang, B., Yang, Y., Hu, P., Zhao, Y., Yan, X., Gao, B., Huang, M., Xu, P. & Li, W. 2012, 'PatentMiner: topic-driven patent analysis and mining', *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 1366-74.

Teh, Y.W., Jordan, M.I., Beal, M.J. & Blei, D.M. 2006, 'Hierarchical dirichlet processes', *Journal of the American statistical association*, vol. 101, no. 476.

TFAMWG, T.F.A.M.W.G. 2004, 'Technology futures analysis: Toward integration of the field and new methods', *Technological Forecasting and Social Change*, vol. 71, no. 3, pp. 287-303.

Tong, X. & Frame, J.D. 1994, 'Measuring national technological performance with patent claims data', *Research Policy*, vol. 23, no. 2, pp. 133-41.

Trajtenberg, M. 1990, 'A penny for your quotes: patent citations and the value of innovations', *The Rand Journal of Economics*, pp. 172-87.

Trappey, C.V., Wu, H.-Y., Taghaboni-Dutta, F. & Trappey, A.J. 2011, 'Using patent data for technology forecasting: China RFID patent analysis', *Advanced Engineering Informatics*, vol. 25, no. 1, pp. 53-64.

Trippe, A.J. 2003, 'Patinformatics: Tasks to tools', *World Patent Information*, vol. 25, no. 3, pp. 211-21.

Tseng, F.-M., Hsieh, C.-H., Peng, Y.-N. & Chu, Y.-W. 2011, 'Using patent data to analyze trends and the technological strategies of the amorphous silicon thin-film solar cell industry', *Technological Forecasting and Social Change*, vol. 78, no. 2, pp. 332-45.

Tseng, Y.-H., Lin, C.-J. & Lin, Y.-I. 2007, 'Text mining techniques for patent analysis', *Information Processing & Management*, vol. 43, no. 5, pp. 1216-47.

Tseng, Y.-H., Wang, Y.-M., Lin, Y.-I., Lin, C.-J. & Juang, D.-W. 2007, 'Patent surrogate extraction and evaluation in the context of patent mapping', *Journal of Information Science*.

USPTO, <http://www.uspto.gov/patents/index.jsp>.

USPTO 2012a, *CLASSES WITHIN THE U.S. CLASSIFICATION SYSTEM*, p. http://www.uspto.gov/patents/resources/classification/classescombined.pdf.

USPTO 2012b, *Classes within the U.S. Classification System*, <http://www.uspto.gov/patents/resources/classification/classescombined.pdf>.

USPTO 2012c, *Manual of Patent Examining Procedure:Claim Interpretation*, USPTO, <http://www.uspto.gov/web/offices/pac/mpep/s2111.html>.

Vaishnavi, V.K. & Kuechler, W. 2015, *Design science research methods and patterns: innovating information and communication technology*, CRC Press.

Veugelers, M., Bury, J. & Viaene, S. 2010, 'Linking technology intelligence to open innovation', *Technological Forecasting and Social Change*, vol. 77, no. 2, pp. 335-43.

Wang, W.M. & Cheung, C.F. 2011, 'A Semantic-based Intellectual Property Management System (SIPMS) for supporting patent analysis', *Engineering Applications of Artificial Intelligence*, vol. 24, no. 8, pp. 1510-20.

Watts, R.J. & Porter, A.L. 2003, 'R&D cluster quality measures and technology maturity', *Technological Forecasting and Social Change*, vol. 70, no. 8, pp. 735-58.

Wikipedia 2014, *Transitional phrase*, Wikipedia, 2014, <http://en.wikipedia.org/wiki/Transitional_phrase>.

WIPO 2002, *Patent Cooperation Treaty (PCT) Article 6*, WIPO, Washington, <http://www.wipo.int/pct/en/texts/articles/a6.htm>.

WIPO 2004, *WIPO Intellectual Property Handbook: Policy, Law and Use*, 2 edn, vol. 489.

Woon, W.L., Zeineldin, H. & Madnick, S. 2011, 'Bibliometric analysis of distributed generation', *Technological Forecasting and Social Change*, vol. 78, no. 3, pp. 408-20.

Wu, D. 2014, 'Tree similarity measure-based recommender systems'.

Wu, D., Zhang, G. & Lu, J. 2013, 'A Fuzzy Tree Similarity Measure and Its Application in Telecom Product Recommendation', *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, IEEE, pp. 3483-8.

Xie, Z. & Miyazaki, K. 2013, 'Evaluating the effectiveness of keyword search strategy for patent identification', *World Patent Information*, vol. 35, no. 1, pp. 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', *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, ACM, pp. 99-108.

Yang, S.-Y. & Soo, V.-W. 2012, 'Extract conceptual graphs from plain texts in patent claims', *Engineering Applications of Artificial Intelligence*, vol. 25, no. 4, pp. 874-87.

Yoon, B. 2008, 'On the development of a technology intelligence tool for identifying technology opportunity', *Expert Systems with Applications*, vol. 35, no. 1, pp. 124-35.

Yoon, B. & Park, Y. 2005, 'A systematic approach for identifying technology opportunities: Keyword-based morphology analysis', *Technological Forecasting and Social Change*, vol. 72, no. 2, pp. 145-60.

Yoon, B. & Park, Y. 2007, 'Development of new technology forecasting algorithm: hybrid approach for morphology analysis and conjoint analysis of patent information', *Engineering Management, IEEE Transactions on*, vol. 54, no. 3, pp. 588-99.

Yoon, J. & Kim, K. 2012, 'TrendPerceptor: A property–function based technology intelligence system for identifying technology trends from patents', *Expert Systems with Applications*, vol. 39, no. 3, pp. 2927-38.

Zadeh, L.A. 1965, 'Fuzzy sets', *Information and control*, vol. 8, no. 3, pp. 338-53.

Zhang, G. 1998, 'Fuzzy number-valued measure theory', Tsinghua University Press, Beijing.

Zhang, G. & Lu, J. 2009, 'A linguistic intelligent user guide for method selection in multi-objective decision support systems', *Information Sciences*, vol. 179, no. 14, pp. 2299-308.

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*, vol. 85, pp. 26-39.

Zhang, Y., Robinson, D.K.R., Porter, A.L., Zhu, D., Zhang, G. & Lu, J. 2015, 'Technology roadmapping for competitive technical intelligence', *Technological Forecasting and Social Change, in press*.

Zhu, D. & Porter, A.L. 2002, 'Automated extraction and visualization of information for technological intelligence and forecasting', *Technological Forecasting and Social Change*, vol. 69, no. 5, pp. 495-506.

Abbreviations

DSSCs	Dye-sensitized Solar Cells
FMDV	Fuzzy Membership Degree Vector
FTTM	Fuzzy Number-based Technological Trend Measurement
ICT	Information and Communications Technologies
IP Australia	Australian Government Intellectual Property Department
IPC	International Patent Classification
LDA	Latent Dirichlet Allocation
NARNNs	Nonlinear Autoregressive Neural Networks
OECD	Organization for Economic Cooperation and Development
PLR	Piecewise Linear Representation
R&D	Research and Development
RSS	Residual Sum of Squares
TDCE	Topic Detection and Comprehensive Evaluation
TF-IDF	Term Frequency-Inverse Document Frequency
TTA	Technological Trend Analysis
TTCI	Technological Topic Change Identification
TTF	Topic-based Technological Forecasting
USPC	United States Patent Classification
USPTO	United States Patent and Trademark Office
WoS	Web of Science

Appendix

Table 1. The top 10 ranked words of 50 topics in USPTO patents and their corresponding probabilities in Section 5.4

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		Topic12		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		Topic23		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
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

Table 2. The top 10 ranked words of topics for years from 2009 to 2013 and their corresponding probabilities in Section 5.5

Year 2009									
Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
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	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
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
Year 2010									
Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
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	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
device	0.0269	acid	0.0199	apparatus	0.0370	**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**

Year 2011

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
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	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
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**

Year 2012									
Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
signal	0.0325	fluid	0.0209	portion	0.0240	**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	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
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

Year 2013									
Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
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	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
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	0.0070	**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

Table 3. The top 10 ranked words of the 12 candidate topics of the final topic set in Section 6.5

Topic 028		**Topic 067**		**Topic 271**		**Topic 391**		**Topic 133**	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
counter electrode	0.3605	photocatalytic	0.2571	dft	0.1277	porphyrin	0.3616	ligands	0.1533
counter electrodes	0.1526	degradation	0.2019	electronic	0.1156	porphyrins	0.1300	ligand	0.1040
pt	0.0871	photocatalyst	0.0488	time-dependent	0.0756	soret	0.0251	mlct	0.0801
electrocatalytic	0.0637	methylene	0.0439	td-dft	0.0657	red-shifted	0.0208	metal-to-ligand	0.0483
pt-free	0.0275	visible-light	0.0334	calculations	0.0603	yd2-o-c8	0.0199	bpy	0.0400
tafel	0.0135	photodegradation	0.0323	geometries	0.0603	meso	0.0194	heteroleptic	0.0361
mos2	0.0124	photocatalysts	0.0319	b3lyp	0.0342	macrocycle	0.0175	22-bipyridine	0.0322
nis	0.0110	rhodamine	0.0300	oscillator	0.0229	porphyrin-sensitized	0.0109	ru(ii)	0.0220
pt-based	0.0103	mb	0.0300	functionals	0.0229	porphyrin-based	0.0109	polypyridyl	0.0166
activities	0.0092	rhb	0.0206	excitations	0.0189	zn(ii)	0.0090	phenanthroline	0.0161

Topic 159		Topic 298		Topic 047		Topic 398		Topic 223	
Word	Probability	Word	Probability	Word	Probability	Word	Probability	Word	Probability
nanotube	0.4654	film	0.5682	nanowires	0.3626	gel	0.4066	pt	0.6622
anodization	0.1356	bcdsc	0.0070	nanowire	0.2590	quassolid-state	0.1324	h2ptcl6	0.0192
nh4f	0.0237	tn	0.0070	single-crystalline	0.0235	electrolytes	0.1239	homogeneously	0.0078
potentiostatic	0.0202	cytochrome	0.0042	vertically	0.0201	qs-dssc	0.0137	electrocatalyst	0.0078
free-standing	0.0197	znhc	0.0042	nanowire-based	0.0151	gelator	0.0114	spin-coating	0.0062
vertically	0.0121	chronoamperometry	0.0035	htitanium-dioxide	0.0034	polyacrylic	0.0080	electroless	0.0062
self-organized	0.0101	bce	0.0035	nanoparticle/nanowire	0.0022	paa-peg	0.0071	pvp-capped	0.0047
tntas	0.0081	nano-sheets	0.0035	anatase-titanium-dioxide	0.0022	qs-dsscs	0.0057	counterelectrodes	0.0047
galvanostatic	0.0061	hsmm-titanium-dioxide	0.0035	suppressed	0.0017	polyhydroxyethyl	0.0052	loadings	0.0042
nanotube-based	0.0050	absorption	0.0028	nanocrystals	0.0017	gelling	0.0047	pdat	0.0042

Topic 389		Topic 037	
Word	Probability	Word	Probability
graphene	0.5049	ipce	0.4118
electrocatalytic	0.0595	photon-to-current	0.1748
nanoplatelets	0.0189	monochromatic	0.0801
graphene-based	0.0130	photon-to-electron	0.0272
gnp	0.0124	apce	0.0115
electro-catalytic	0.0118	spp	0.0026
synergistic	0.0112	alpha-fe2o3	0.0026
graphene-titanium-dioxide	0.0088	substrates	0.0021
electrocatalyst	0.0077	piezoelectric	0.0016
few-layer	0.0077	mdssc	0.0016