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

CERTIFICATE OF AUTHORSHIP/ORIGINALITY i

ACKNOWLEDGEMENTS i

ABSTRACT iii

TABLE OF CONTENTS v

LIST OF FIGURES x

LIST OF TABLES xiii

CHAPTER 1 Introduction 1

1.1 Background 1

1.2 Research Questions and Objectives 3

1.2.1 Research Questions 3

1.2.2 Research Objectives 5

1.3 Research Significance 8

1.3.1 Theoretical Significance 8

1.3.2 Practical Significance 9

1.4 Research Methodology and Process 10

1.4.1 Research Methodology 10

1.4.2 Research Process 12

1.5 Thesis Structure 13

1.6 Publications Related to This Thesis 14

CHAPTER 2 Literature Review 17

2.1 Technology Intelligence 17

2.1.1 Concept and Framework of Technology Intelligence 17

2.1.2 Technology Intelligence in Practice 20

2.1.3 Tech Mining 21

2.2 Technology Indicators..22
2.2.1 Patent..23
2.2.2 Patent Claims..25
2.2.3 Scientific Literature..26
2.3 Empirical Technology Trend Analysis and Forecasting..27
2.3.1 Curve Fitting Based Approaches..28
2.3.2 Other Approaches..30
2.3.3 Piecewise Linear Representation..32
2.4 Topic Modelling..34
2.4.1 Semantic Space..35
2.4.2 Latent Dirichlet Allocation..36
2.4.3 Topic Modelling in Tech Mining..38
2.5 Summary..39
CHAPTER 3 Framework of Topic-based Technology Intelligence..41
3.1 Introduction..41
3.2 The Limitations of Existing Technology Intelligence Frameworks..42
3.3 Temporal Characteristics and Semantic Properties of Technology Indicators..44
3.4 Framework Description..48
3.4.1 Input and Output..48
3.4.2 Topic-based Technology Intelligence..50
3.4.3 Conceptual Model of Topic-based Technology Intelligence..51
3.4.4 Overall Framework Description..54
3.5 Description of the Framework Components..57
3.5.1 Trend Identification Component..58
3.5.2 Topic Discovery Component..60
3.5.3 Comprehensive Topic Evaluation Component..62
3.6 Summary..64
CHAPTER 4 Empirical Technology Trend Analysis Method..65
4.1 Introduction..65
4.2 Data Preparation for TTA Method..67
4.2.1 Outsider Exclusion..67

4.2.2 Parameter Setup for Piecewise Linear Representation 68

4.3 TTA Method and Afterwards Trend Forecasting 71

4.3.1 Trend Turning Points Identification 71

4.3.2 Trend Segments Identification 72

4.3.3 Trend Movement Intensity 73

4.3.4 TTA-based Trend Forecasting 74

4.4 Case Study 1: Patent Data in IP Australia 75

4.4.1 Data Sets 76

4.4.2 Outliers Exclusion 76

4.4.3 Trend States and Trend Turning Points Identification 77

4.5 Case Study 2: Patent Data in USPTO 79

4.5.1 Data Collection and Parameter Setting 79

4.5.2 Trend Forecasting for Telecommunications Technologies..............................81

4.5.3 Trend Forecasting for Solar Cell Technologies 84

4.5.4 Trend Forecasting for Radar-related Technologies in USPC 342 87

4.5.5 Comparison and Discussion 90

4.6 Summary 95

CHAPTER 5 Topic-based Technological Forecasting Method and afterwards Content Analysis 97

5.1 Introduction 97

5.2 Patent Crawling and Cleaning 98

5.3 Topic-based Technological Forecasting Method 101

5.3.1 Method Framework 101

5.3.2 Topic Modelling 104

5.3.3 Topic-based Technological Forecasting and Analysis 105

5.4 Case Study of Topic-based Technological Forecasting by Patent Data 108

5.4.1 Trend Pattern Identification 109

5.4.2 Topic Modelling and Prominent Topic Selection 110

5.4.3 Topic Annual Weight Matrix and Topic-based Trend Coefficients Estimation 112

5.4.4 Topic-based Trend Forecasting and Analysis 113

5.4.5 Discussion 116

5.5 Technological Topic Change Identification Approach 117

5.5.1 Framework of the TTCI Approach 117

5.5.2 Topic Modelling 119

5.5.3 Topic Change Identification Model 120

5.5.4 Case Study of TTCI Approach by Patent Data 122

5.6 Fuzzy Number-based Technological Trend Measurement Approach 128

5.6.1 Framework of the FTTM Approach 128

5.6.2 Fuzzy Set 130

5.6.3 Topic Weight Estimation 131

5.6.4 Fuzzy-based Technological Development Measurement 131

5.6.5 Case Study of FTTM Approach by Patent Data 134

5.7 Summary 138

CHAPTER 6 Topic Detection and Comprehensive Evaluation Method 139

6.1 Introduction 139

6.2 Methodology Framework 141

6.3 Topic Modelling 143

6.3.1 Parameters Setting 143

6.3.2 Final Topic Set Determination 145

6.4 Topic Evaluation Indices 146

6.4.1 Topic Weight Index 147

6.4.2 Topic Trend Index 147

6.4.3 Topic Activeness Index 149

6.4.4 Topic-based Citation 149

6.4.5 Topic-based Prominent Topics and Documents Identification Using Metadata 151

6.5 Case study: Dye-sensitized Solar Cells Scientific Literature 151

6.5.1 Data 151

6.5.2 Scientific Literature Text Cleaning 152

6.5.3 Parameters Setting and Final Topic Set Determination 153

6.5.4 Topic Evaluation Result 154

6.5.5 Topic-based Evaluation Maps 156

6.5.6 Prominent Topics and Papers Analysis 160

6.5.7 Discussion 165

6.6 Summary 166

CHAPTER 7 Conclusions and Further Study 167

7.1 Conclusions 167

7.2 Further Study 171

References 173

Abbreviations 186

Appendix 187

LIST OF FIGURES

Figure 1-1. The methodology of design research of this thesis 11
Figure 1-2. Thesis structure.. 13
Figure 2-1. Technology intelligence service define by Savioz (2004) 19
Figure 2-2. The technology intelligence process defined by Kerr at al. (2006)................ 20
Figure 2-3. An example of curve fitting approaches in technology forecasting 29
Figure 2-4. Comparison of growth curves, PLR-based approaches and time series analysis 33
Figure 2-5. Brief introduction of the 'semantic space' of topic modelling 35
Figure 2-6. The graphical model of Latent Dirichlet Allocation 37
Figure 3-1. The three-dimensional schematic structure of the semantic property and temporal characteristic 46
Figure 3-2. 'Topics' and 'Trend Segments' 48
Figure 3-3. Schematic diagram of the input and output of the topic-based technology intelligence 50
Figure 3-4. Brief introduction of topic-based technology intelligence 51
Figure 3-5. The Conceptual model of topic-based technology intelligence...................... 53
Figure 3-6. Framework of topic-based technology intelligence....................................... 56
Figure 3-7. Details of the trend identification component 59
Figure 3-8. Details of the topic discovery component 61
Figure 3-9. Details of the comprehensive topic evaluation component 63
Figure 4-1. An example of PLR threshold setup.. 70
Figure 4-2. An example of transforming original data to trend segments step by step 73
Figure 4-3. The detailed process of the TTA-based trend forecasting............................. 75
Figure 4-4. The outliers exclusion for ICT patent count sequence 77

Figure 4-5. Original data, PLR segmentation and trend segments of technologies in ICT industry 78
Figure 4-6. Normalized original and cumulative patent data in the case study 80
Figure 4-7. PLR threshold setup for the three technologies in the case study 81
Figure 4-8. Original data, PLR segmentation result and trend segments of telecommunication technologies 82
Figure 4-9. The forecasting result of the trend segments of Telecommunication technologies 84
Figure 4-10. Original data, PLR segmentation result and trend segments of solar cell technologies solar cell technologies 85
Figure 4-11. The forecasting result of the trend segments of solar cell technologies 87
Figure 4-12. Original data, PLR segmentation result and trend segments of radar-related technologies 88
Figure 4-13. The forecasting result of the trend segments of radar-related technologies . 90
Figure 4-14. The growth curves fitting result for the three technologies 92
Figure 4-15. Experimental comparison between the proposed approach and the growth curves model 93
Figure 5-1. An example of webpage crawling result 100
Figure 5-2. Framework for the topic-based technological forecasting method 103
Figure 5-3. An example of topic distribution matrix in chronological order 107
Figure 5-4. The trend turning points and trend segments generated from patenting activities 109
Figure 5-5. The curve fitting result for topic trend estimation of the 10 selected topics 114
Figure 5-6. The framework of the proposed technological topic change identification approach 118
Figure 5-7. Relationships between sub-collections and topics 119
Figure 5-8. Topic change identification model 120
Figure 5-9. Topics became newly important in each year of 2010 to 2013 and topmost frequent words of each topic 125
Figure 5-10. An example of the topic “antibody” evolving over time 126
Figure 5-11. An example of the topic-based trend estimation of the theme “antibody”. 127

Figure 5-12. The framework of FTTM approach..129
Figure 5-13. Linguistic terms and their membership functions 133
Figure 5-14. Membership functions of linguistic terms in term set I............................136
Figure 6-1. The framework of the topic detection and comprehensive evaluation method 142
Figure 6-2. An example of annual weight matrix 148
Figure 6-3. The log likelihood of the probability of the observation under models with different setting of the number of topics 154
Figure 6-4. The annual weight growth of the top 10 high weighted topic.....................155
Figure 6-5. The topic-based evaluation map based on weight, trend and activeness characteristics of DSSCs corpus 157
Figure 6-6. The topic-based citations map based on the total citation...........................159
Figure 6-7. The topic evaluation map based on weight, trend, activeness and topic-based citations of DSSCs corpus...160
Figure 6-8. Graphical illustration of the main content of the 6 prominent topics...........161
Figure 6-9. The cumulative citation distribution of DSSCs corpus from year 1991 to 2014 162
Figure 6-10. Publications country distribution of DSSCs corpus from year 1991 to 2014 164

LIST OF TABLES

Table 4-1. The detailed outliers' values of ICT patent count sequence 77

Table 4-2. Trend signal value and trend tags of ICT technologies in Australia............... 79

Table 4-3. Description of the data sets for case study 2 80

Table 4-4. Curve fitting information for PLR threshold determination 81

Table 4-5. The trend segments information of Telecommunication technologies 83

Table 4-6. The trend segments information of solar cell technologies 86

Table 4-7. The trend segments information of radar-related technologies 89

Table 4-8. Growth curves selection for comparison experiments................................... 91

Table 4-9. Forecasting result comparison 94

Table 5-1. The start and end of webpage source code while patent crawling.................. 99

Table 5-2. The basic notations throughout this chapter 105

Table 5-3. Trend forecasting indicators and future trend estimation 107

Table 5-4. The trend turning points, document numbers, and term numbers for each trend segment.. 110

Table 5-5. The 50 topics generated from the patent claims collection and their weight indicator 111

Table 5-6. The annual weight matrix of the selected top 10 significant topics.............. 112

Table 5-7. Topic-based trend coefficients for all 10 prominent topics 113

Table 5-8. The details of trend estimation for the 10 selected topics............................ 115

Table 5-9. The number of documents, terms and USPC of patents published each year122

Table 5-10. Linguistic terms and fuzzy numbers 133

Table 5-11. The Temporal-weight coefficients of topic 1 to topic 30 135

Table 5-12. Development states measure result 136

Table 6-1. The number of documents, terms and subjects of documents each year 152

Table 6-2. Similarity evaluation result for the final topic set selection 154

Table 6-3. The top 10 topics with the highest evaluation indices values 156

Table 6-4. Top 10 topics with the highest values (normalized) of the three indices....... 158

Table 6-5. Detected topics and the publications they covered 163

Table 6-6. The most contributory papers of the prominent topic set 165