Text Mining of Technology Roadmaps: Integrated Analysis and Visualization of Postprints

Authors: Xie Xiufang, Zhang Xiaolin

Date: 2017-11-08T00:00:00+00:00

Abstract

【Objective】To conduct knowledge discovery research on technology roadmap content and predict long-term trends in future technological development.

【Method】Based on a technology roadmap information database constructed using the “extraction-synchronization-classification” text mining method, we performed integrated analysis of global technological development demands and trends, comparative analysis of national development routes and measures, and conducted an empirical study with the renewable energy sector as a case.

【Results】The empirical research results were visualized using open-source tools such as Timeflow and Gephi, presenting from multiple perspectives the development trends of the renewable energy sector until 2050 and the strategic plans of various countries in chronological sequence.

【Limitations】The approach integrates multiple methods and tools, but the degree of automation needs improvement and personalized functions require further refinement.

【Conclusion】This research approach can rapidly extract core information from technology roadmaps and enhance intelligence acquisition efficiency.

Full Text

Integrated Analysis and Visualization of Science and Technology Roadmaps: A Text Mining Approach

Xie Xiufang[1,2], **Zhang Xiaolin**[1]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190, China

[2] School of Health Management and Education, Capital Medical University / Library of Capital Medical University, Beijing 100069, China

Abstract

[**Objective**] This study aims to enable knowledge discovery from science and technology roadmap (STR) content to predict long-term S&T development trends. [**Methods**] Based on a text mining approach employing "extraction-synchronization-classification," we constructed an STR information repository to conduct integrated analysis of global S&T development demands and trends, and to compare national development pathways and measures, using renewable energy as an empirical case study. [**Results**] We visualized the empirical findings using open-source tools including Timeflow and Gephi, presenting the development landscape of renewable energy through 2050 from multiple perspectives along a temporal sequence, as well as national strategic planning in this domain. [**Limitations**] While the study integrates multiple methods and tools, the automation level requires improvement and personalized functionalities need further development. [**Conclusions**] The proposed research framework can rapidly extract core information from STRs and enhance intelligence acquisition efficiency.

Keywords: Science and Technology Roadmap; Strategic Intelligence; Text Mining; Knowledge Discovery; Integrated Analysis; Information Visualization

1. Introduction

Science and technology roadmaps represent the most direct strategic intelligence carriers for future S&T development plans worldwide, containing rich information about national S&T development status, directions, technology evolution timelines, visions, milestone objectives, and strategic measures. They constitute intensive strategic intelligence resources. Applying text mining techniques to extract, organize, and analyze information from STRs holds significant strategic importance for grasping future S&T directions and formulating long-term development plans [1].

However, in the field of information science, STRs have primarily been treated as products of intelligence research [2-4] rather than as data resources for intelligence mining [5-6]. Common approaches involve manual interpretation of individual roadmap reports [7-8], with few studies employing text mining techniques across large collections of STR documents. This research proposes a text mining framework specifically designed for STRs. By analyzing the content organization and expression characteristics of STRs, we explore automated information extraction methods to construct an STR knowledge base. This knowledge base then enables integrated analysis, comparative analysis, and trend analysis of large volumes of STRs, thereby achieving text mining and knowledge discovery from these documents [9].

Building upon the information extraction methodology proposed in our previous work [10], this study constructs an STR information repository to integrate and analyze strategic planning information from countries worldwide across various domains. We examine the current global development status of each domain and predict future trends, thereby implementing a knowledge discovery process for STR textual content and providing strategic intelligence services to support decision-makers in developing domain-specific plans.

2. Framework for Integrated Comparative Analysis of Science and Technology Roadmaps

Our preliminary research surveyed and analyzed 166 STR documents released by 21 countries or organizations, constructing a content description framework and information classification system for STRs [9]. Based on this foundation, we explored information extraction methods for STR textual content, involving sequential processes of text cleaning, information extraction, data cleaning, synchronization matching, and information classification.

Through this analytical framework, we can comprehensively understand the global development status and trends in relevant domains, identify critical future technologies, and examine national technology choices and strategic deployments. This approach provides significant intelligence value and reference significance for seizing development opportunities and formulating development strategies.

3. Methodology for Integrated Analysis

3.1 Demand Analysis

To analyze global development demands within a specific domain and integrate survey results on current development status from countries worldwide, we employed the open-source tool Timeflow [11] for visualization. The tool presents status (Classification_1= "today") and demand (Classification_2= "need") information for a given domain along a temporal axis (Time), with classification (Classification_3) and node size weighted by relevance (W), while using different colors to represent classification indicators (Classification_4).

To enable this functionality, we performed the following data preprocessing steps:

(1) Keyword-Time Association Matching
Since a single sentence may contain multiple keywords and temporal expressions, we developed Python code to traverse each keyword and temporal term within the same sentence, achieving one-to-one matching based on their quantity relationships and value words. Details are available in the synchronization matching section of our previous work [10].

(2) Temporal Expression Calculation
To present various types of information on a temporal axis, we mapped temporal expressions such as "today/year/decade/century" to numerical formats. For records with empty "Time" fields, we assigned temporal information based on Classification_1 values. Letting t represent the publication year of

the report, we applied calculation rules for temporal expressions containing "year/decade/century" as shown in . For other temporal expressions not involving calculations, we applied direct assignment rules as shown in .

Although Timeflow can display time intervals, we decomposed intervals into multiple planning nodes for clearer visualization. For example, the interval (2015, 2050) was decomposed into "2015, 2020, 2030, 2050," and (2011, 2030) into "2011, 2020, 2030." Keywords from relevant sentences were then associated with each temporal node.

(3) Keyword Weight Calculation
We assigned weights to keywords based on three factors: location within the document, term frequency, and document frequency.

- **Location weight (wp):** Different weights were assigned based on positional importance [9, 12]: whead=7, wtitle=6, witem=5, wbegin=4, wlead=3, wend=2, wplain=1.
- **Term frequency weight (wt):** The natural logarithm of keyword frequency within a document: wt = ln(TF).
- **Document frequency weight (wd):** The natural logarithm of the number of documents containing the keyword in the corpus: wd = ln(DF).

The total keyword weight W = wp + wt + wd, where wp and wt reflect local importance within a document, while wd reflects global importance across the corpus. The combined weight W thus indicates the keyword's significance within the overall domain development plan.

3.2 Trend Analysis

Integrating global judgments and analyses on future development trends, potential, and opportunities in a given domain enables prediction of overall long-term development trajectories. We used Timeflow to visualize vision (Classification_1= "vision") and trend (Classification_2= "trend & potential & opportunity") information for a domain along a temporal axis (Time), with classification (Classification_3) and node size weighted by relevance, while using different colors for classification indicators (Classification_4). Temporal and weight calculation methods follow the same approach as in demand analysis.

3.3 Route Analysis

To anticipate technology development directions in a domain, we integrated future route planning information from countries worldwide. Using Gephi [13], we presented national technology choices and objectives according to temporal planning nodes, with keywords as node labels and intra-sentence co-occurrence relationships as edges. Because the core information in the content···

4. Visualization of Integrated Analysis Results

Using renewable energy as a case study, we integrated STR information on this domain from global sources to analyze current development status and trends, identify critical future technologies, and examine technology choices and strategic measures adopted by different countries.

4.1 Demand Analysis Example

Integrating current demand information for renewable energy from various countries and visualizing it with Timeflow ([Figure 3: see original paper]) reveals that development demands encompass technology, market, policy, environment, and economy, with primary emphasis on technology, market, and policy needs.

1. **Technology:** Focuses on performance (orange) and maturity (magenta), such as improving solar collector performance and commercializing advanced biofuels.
2. **Market:** Centers on production (red) and consumption (dark red), such as increasing power generation, renewable energy share, and energy demand.
3. **Policy:** Concerns planning (green) and support (blue), such as strengthening photovoltaic research, formulating national energy plans, and increasing hydropower projects.

4.2 Trend Analysis Example

Integrating development vision and trend information for renewable energy and visualizing it with Timeflow ([Figure 4: see original paper]) shows that development trends include market, technology, environment, policy, economy, and social aspects, with primary focus on technology, market, and environment trends.

1. **Market:** Production (red) and consumption (dark red), such as increasing renewable energy share and power generation while reducing prices and incremental costs.
2. **Technology:** Performance (orange) and cost (brown), such as built-in thermal storage, photovoltaic power generation, plant efficiency, and reduced investment costs.
3. **Environment:** Emissions (cyan) and resources (purple), such as reducing greenhouse gas emissions and expanding renewable resource development.

4.3 Route Analysis Example

(1) Technology Development Landscape

Integrating global information on renewable energy technology trends, potential, and opportunities, we visualized the technology development landscape from 2015 to 2050 using Gephi ([Figure 5: see original paper]). The analysis reveals that development primarily centers on concentrated solar power (CSP), biomass, photovoltaic (PV), and wind power technologies, with performance improvements in cost, CO_2 emissions, and power generation.

(2) Technology Development Direction
Integrating global technology target information for renewable energy and visualizing it with Gephi ([Figure 6: see original paper]) shows the technology development direction from 2015 to 2050. The analysis indicates that development focuses on applications of CSP, PV, hydropower, wind power, biomass, and biofuel technologies in transportation, building, and heating sectors.

(3) Technology Development Pathways
Selecting specific technology keywords, we compared different national target plans for technology development pathways and visualized them using Timeflow. [Figure 7: see original paper] presents development pathway planning for four key renewable energy technologies—CSP, carbon capture and storage (CCS), wind power, and PV—across different countries. For CSP, hovering over each node reveals the technology' s development status at that time point, enabling comparison of national development pathways.

4.4 Strategic Analysis Example

Comparing the development visions, routes, and measures of the United States and China in renewable energy, we conducted comprehensive visualization using Timeflow ([Figure 8: see original paper]).

(1) Development Vision Comparison
- **United States:** Vision encompasses market, policy, technology, environment, and economy, focusing primarily on market, policy, and technology. Market aspects emphasize production (red), such as increasing renewable energy generation and share. Policy aspects focus on planning (green), such as sustainable energy development plans and renewable energy plant construction. Technology aspects target performance (orange), such as improving solar and wind efficiency. - **China:** Vision includes market, policy, technology, economy, society, and environment, also focusing on market, policy, and technology. Market aspects emphasize production and consumption, such as increasing renewable energy generation and capacity, making wind power cheaper than coal power, and reducing wind generation costs. Policy aspects focus on planning and support, such as formulating wind development strategies and incentivizing grid companies to purchase renewable power. Technology aspects target maturity (magenta), such as advancing wind power technology, large wind farms, and offshore wind farms.

(2) Development Route Comparison
- **United States:** Routes encompass technology, environment, policy, and market. Technology focuses on performance (orange), such as improving wind and solar efficiency. Environment emphasizes emissions (cyan), such as reducing CO_2 emissions by 17% by 2020, 26-28% by 2025, 33% by 2030, and 80% by 2050 compared to 2005 levels. Policy centers on planning (green), such as devel-

oping routes to accelerate renewable energy development. Market emphasizes production, such as increasing onshore wind capacity from 63 GW in 2014 to 314 GW by 2030. - **China:** Routes encompass technology, market, and policy. Technology focuses on performance (orange) and maturity (magenta), such as wind capacity targets of 15 GW (2010-2015), 200 GW by 2020, 400 GW by 2030, and 1 TW by 2050, with a progression from onshore wind (2010-2020) to nearshore wind (2020-2030) to offshore wind and micro-siting technology (2030-2050). Market emphasizes production and consumption, such as wind power substituting 130 million tonnes of coal equivalent (mtce) by 2020, 260 mtce by 2030, and 660 mtce by 2050. Policy focuses on planning, such as formulating renewable energy development strategies.

(3) Development Measures Comparison
- **United States:** Measures encompass policy, economy, technology, environment, and market, focusing on policy, economy, and technology. Policy emphasizes support (blue), such as supporting renewable energy R&D. Economy focuses on investment (brown), such as investing $86 billion annually in renewable energy technologies. Technology targets performance (orange), such as improving energy carriers and solar PV performance. - **China:** Measures encompass policy, market, technology, society, environment, economy, and cooperation, focusing on policy and market measures. Policy includes planning (green), management (blue-green), evaluation (dark blue), and support (blue), such as formulating renewable energy industry development plans, electricity pricing regulations, improving wind power technical standards, testing and certification systems, and providing technical consulting and strategic research services. Market measures focus on market mechanisms, such as accelerating power market reform, establishing wind power management and market rules, and electricity pricing marketization.

5. Conclusion

This study employed our previously proposed text mining framework and information extraction methodology to construct an STR information repository. Using renewable energy as a case study, we integrated and analyzed global development demands and long-term trends, compared national technology development routes and strategic measures, and visualized the results using open-source tools such as Timeflow and Gephi. Experimental results demonstrate that our methodology is feasible and effective, enabling rapid grasp of global development status and trends while comprehensively understanding long-term development routes and measures worldwide, thereby providing decision-makers with comprehensive and efficient strategic intelligence services for developing nationally appropriate plans.

The research remains in the methodological exploration stage with several limitations requiring improvement. The approach integrates multiple programming languages and software tools, but limited by time and resources, we have not yet achieved full integration of all tools, leaving room for enhanced automation.

Additionally, visualizations currently utilize existing functions of open-source tools; future research could develop customized functionalities based on specific needs. For instance, current capabilities only allow hovering over keyword nodes to display relevant information, but we could add click-through functionality to access associated paragraphs, chapters, and detailed pages—numerous enhancements remain to be developed.

References

[1] Liu Xiwen, Ke Chunxiao. The Applications of Technology Roadmap and Its Enlightenment to Strategic Intelligence Research[J]. Library and Information Service, 2007, 51(6): 37-40, 112.

[2] Zhang Y, Zhang G, Chen H, et al. Topic Analysis and Forecasting for Science, Technology and Innovation: Methodology with a Case Study Focusing on Big Data Research[J]. Technological Forecasting and Social Change, 2016, 105: 179-191.

[3] Amer M, Daim T U, Jetter A. Technology Roadmap Through Fuzzy Cognitive Map-Based Scenarios: The Case of Wind Energy Sector of a Developing Country[J]. Technology Analysis & Strategic Management, 2016, 28(2): 131-155.

[4] Jin G, Jeong Y, Yoon B. Technology-driven Roadmaps for Identifying New Product/Market Opportunities: Use of Text Mining and Quality Function Deployment[J]. Advanced Engineering Informatics, 2015, 29(1): 126-138.

[5] Ye Chunlei, Leng Fuhai. Building the Future-Oriented Technology Thesaurus of Technology Roadmap[J]. New Technology of Library and Information Service, 2013(5): 59-63.

[6] Ye Chunlei, Leng Fuhai. Study on the Keyword Extraction from Roadmap Based on the Lexical Chains[J]. New Technology of Library and Information Service, 2013(1): 50-56.

[7] Amer M, Daim T U. Application of Technology Roadmaps for Renewable Energy Sector[J]. Technological Forecasting and Social Change, 2010, 77(8): 1355-1370.

[8] Bader B, Richardson C, Tsuriya M. Technology Roadmap Overviews and Future Direction through Technology Gaps[C]//Proceedings of the 2015 International Conference on Electronics Packaging. 2015.

[9] Xie Xiufang, Zhang Xiaolin. Text-mining Framework and Feature Analysis on Science and Technology Roadmap[J]. Information Science. (In Press).

[10] Xie Xiufang, Zhang Xiaolin. The Research on Text-mining of Science and Technology Roadmap: Method of Information Extraction[J]. Information Studies: Theory & Application. (In Press).

[11] Timeflow[EB/OL]. [2016-07-22]. https://github.com/FlowingMedia/TimeFlow/wiki.

[12] Shi Lei, Wang Yongcheng. Research and Development of an Automatic Abstracting System for English Documents[J]. Chinese High Technology Letters, 1999, 9(11): 22-26.

[13] Gephi[EB/OL]. [2016-07-22]. https://gephi.org/.

Author Contributions

Xie Xiufang: Designed the research framework, collected and analyzed data, wrote the manuscript; Zhang Xiaolin: Defined research direction, proposed research ideas, revised the manuscript.

Conflict of Interest

All authors declare no conflict of interest.

Supporting Data

The supporting data is self-archived by the authors and available upon request at: xiexiufang@mail.las.ac.cn.

- [1] Xie Xiufang, Zhang Xiaolin. Keywords-network.py. Code for keyword network construction.
- [2] Xie Xiufang, Zhang Xiaolin. data-20160608-timesplit.csv. Supporting dataset for demand analysis (Figure 3), trend analysis (Figure 4), pathway analysis (Figure 7), and strategic analysis (Figure 8).
- [3] Xie Xiufang, Zhang Xiaolin. Keywords_{trend}.rar, node_{trend}.rar, edge_{trend}.rar. Supporting data for technology development landscape analysis (Figure 5).
- [4] Xie Xiufang, Zhang Xiaolin. Keywords_{VsTg}.rar, node_{VsTg}.rar, edge_{VsTg}.rar. Supporting data for technology development direction analysis (Figure 6).

Received: September 30, 2016; Revised: November 2, 2016

Note: Figure translations are in progress. See original paper for figures.