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 development trends in future science and technology. Method: Utilizing a technology roadmap information database constructed through the "extract-synchronize-classify" text mining approach, this study performs integrated analysis of global technology development demands and trends, comparative analysis of national development routes and measures, and presents an empirical study using the renewable energy sector as a case. Results: The empirical study results are visualized using open-source tools such as Timeflow and Gephi, chronologically presenting from multiple perspectives the development prospects of the renewable energy sector through 2050 and the strategic plans of various countries. Limitations: While integrating multiple methods and tools, the level of automation requires improvement and personalized functionalities need further refinement. Conclusion: This research framework enables rapid extraction of core information from technology roadmaps and enhances the efficiency of intelligence acquisition.

Full Text

Text Mining Research on Science and Technology Roadmaps: Integrated Analysis and Visualization

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 achieve knowledge discovery from science and technology roadmap (STR) content to predict long-term future S&T development trends. [**Methods**] Based on an STR information database constructed using a "extraction–synchronization–classification" text mining approach, we conducted integrated analysis of global S&T development demands and trends, compared national development routes and measures, and performed an empirical case study in the renewable energy domain. [**Results**] We visualized the empirical findings using open-source tools including Timeflow and Gephi, presenting multi-perspective development scenarios in renewable energy through 2050 in chronological sequence along with national strategic planning. [**Limitations**] The study employs multiple methods and tools, requiring improved automation and customized functionality. [**Conclusions**] The proposed research framework enables rapid extraction of core information from STRs, enhancing intelligence acquisition efficiency.

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

Science and technology roadmaps represent the most direct strategic intelligence carriers for future S&T development planning worldwide, containing intensive strategic information about national S&T development status, directions, technology evolution processes, visions, phased targets, and strategic measures. Applying text mining techniques to extract, organize, and analyze STR information 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 specific roadmap reports [7-8], with few studies conducting text mining across large collections of STR documents. This study proposes a text mining framework for STRs that analyzes content organization and expression characteristics, explores automated STR information extraction methods, and establishes an STR knowledge base. This foundation enables integrated analysis, comparative analysis, and trend analysis of large-scale STR collections, achieving text mining and knowledge discovery from roadmaps [9].

Building upon the information extraction methodology proposed in [10], this paper constructs an STR information database to integrate and analyze strategic planning information from countries worldwide across various domains. By examining each domain' s global development status and forecasting future trends, we implement a knowledge discovery process for STR textual content, providing strategic intelligence services for decision-makers formulating domain-specific development plans.

Framework for Integrated Comparative Analysis of Science and Technology Roadmaps

Our preliminary research analyzed 166 STR documents published by 21 countries or organizations, constructing an STR content description framework and information classification system [9]. Based on this foundation, we explored information extraction methods for STR textual content through text cleaning, information extraction, data cleaning, synchronization matching, and information classification. This process yielded an STR information database comprising four categories—basic text information, semantic classification information, core content information, and original sentence information—totaling 19 fields (see [Figure 1: see original paper]). Semantic classification values derive from classification items in the information classification system [9], as illustrated in [Figure 2: see original paper].

[Figure 1: see original paper] Composition Fields of the Science and Technology Roadmap Information Database
[Figure 2: see original paper] Value Sources for Semantic Classification Information in Science and Technology Roadmaps

Leveraging this STR information database, we can perform analyses at both global and national levels based on domain (Area), time (Time), focus object (Object), and semantic classification:

1. **Demand Analysis:** Integrate global reviews of a domain' s development status to analyze worldwide development demands and identify strategic positioning.
2. **Trend Analysis:** Integrate global analyses of a domain' s development trends, potential, and opportunities to forecast long-term worldwide development trajectories and identify emerging opportunities.
3. **Route Analysis:** Integrate and compare national technology development routes within a domain to inform domestically appropriate planning:
    - **Technology Development Situation Analysis:** Integrate technology selection information across different time periods to analyze technological evolution patterns.
    - **Technology Development Direction Analysis:** Integrate technology target information across different periods to predict future technology directions.
    - **Technology Development Path Analysis:** Integrate development targets for a specific technology across countries to compare different national path planning.
4. **Strategic Analysis:** Integrate and compare strategic measures adopted by different countries in a domain to inform timely and appropriate strategic planning:
    - **National Vision Comparison:** Compare development vision targets of different countries within the same timeframe.
    - **National Route Comparison:** Compare development route plan-

ning of different countries within the same timeframe.
- **National Measures Comparison:** Compare different strategic measures adopted by countries within the same timeframe.

As STRs are typically authoritative and forward-looking documents published by credible institutions for specific domains, these analyses provide comprehensive understanding of a domain' s global development status and trends, future key technologies, and national technology choices and strategic deployments—offering significant intelligence value for seizing development opportunities and formulating strategies.

Demand Analysis

To analyze a domain's worldwide development demands, we integrate global survey results on the domain' s current status. This study employs the open-source tool Timeflow [11] for visualization, presenting status (Classification_1="today") and demand (Classification_2= "need") information along a timeline (Time) with classification (Classification_3). Node sizes represent keyword weights (W), while colors indicate classification metrics (Classification_4). Data preprocessing involves:

1. **Keyword-Time Association Matching**
   Since sentences may contain multiple keywords and temporal expressions, we use Python to iterate through each keyword and time word within a sentence, achieving one-to-one matching based on value words and quantity relationships. Details are available in the core information synchronization matching section of [10].

2. **Temporal Expression Value Calculation**
   To present information on a timeline, temporal expressions like "today/year/decade/century" are uniformly mapped to numerical values. Empty Time fields are assigned values based on Classification_1 results. Letting t represent the report publication year, we calculate temporal information as "t $\pm$ n" using the rules in for expressions involving "year/decade/century," and direct assignment rules in for other cases.

   Temporal Expression Calculation Rules
   Temporal Expression Assignment Rules

   Although Timeflow can display time intervals, we decompose intervals into multiple planning nodes for clarity (e.g., (2015, 2050) becomes "2015, 2020, 2030, 2050"). Keywords are then associated with each time node.

3. **Keyword Weight Calculation**
   Weights are assigned based on location, term frequency (TF), and document frequency (DF):

   - **Location Weight (wp):** Values assigned by positional importance [9, 12]: whead=7, wtitle=6, witem=5, wbegin=4, wlead=3, wend=2,

wplain=1.
- **Term Frequency Weight (wt):** Natural logarithm of TF: wt = ln(TF).
- **Document Frequency Weight (wd):** Natural logarithm of DF: wd = ln(DF).

Total keyword weight W = wp + wt + wd, where wp and wt reflect local importance within a document, wd reflects global importance across the corpus, and W represents overall importance in domain development planning.

Trend Analysis

Integrating global judgments on future domain trends, potential, and opportunities enables prediction of long-term worldwide development trajectories. We use Timeflow to visualize vision (Classification_1= "vision") and trend (Classification_2= "trend & potential & opportunity") information along a timeline (Time) with classification (Classification_3). Node sizes represent weights, and colors indicate classification metrics (Classification_4). Temporal and weight calculation methods follow the demand analysis approach.

Route Analysis

To anticipate technology development directions in a domain, we integrate global future route planning information. Using Gephi [13], we present national technology selections and targets by time planning nodes, with keywords as node labels and intra-sentence co-occurrence relationships as edges. Since core information includes all terminological keywords from a sentence, we construct edge relationships using keyword and related term fields through Python programming:

1. **Technology Development Situation Analysis**
   - Filter a domain (Area) and integrate keywords (kwi), weights (wi), and related terms (rtj) for technology (Classification_3="technology") trends, potential, and opportunities (Classification_2= "trend & potential & opportunity") from national route (Classification_1= "pathway") planning across time stages (Time).
   - Construct keyword networks by iterating through each keyword kwi.
   - Output node weights to node_{trend}.txt and edge weights to edge_{trend}.txt.
   - Import files into Gephi, using node weights for label size/color intensity and edge weights for thickness/color intensity.
2. **Technology Development Direction Analysis**
   - Filter a domain and integrate keywords, weights, and related terms for technology (Classification_3= "technology") development targets (Classification_2= "need & target") from national route (Classification_1= "pathway") planning.

- Construct and visualize keyword networks using the same method.

3. **Technology Development Path Analysis**
   Filter a specific technology keyword and compare different national target (Classification_2= “target”) planning for that technology’ s development path (Classification_1= “pathway”) using Timeflow visualization along a timeline (“Time”) for different countries (“Object”).

Strategic Analysis

To compare national strategic planning in a domain—including visions, routes, and measures—we use Timeflow for comprehensive visualization of relevant strategic information for target countries. Presentations follow time series (x-axis: “time”) with classification (y-axis: “Classification_3”), using weights (W) for node sizes and colors for classification metrics (Classification_4).

Select a domain (Area) and target object (Object) to integrate information across policy, economy, technology, market, and other dimensions (Classification_3), analyzing the object’ s vision, route, and measures. Specific analysis targets and configurations are shown in .

Strategic Analysis Parameter Settings for Target Objects in a Domain

Visualization of Integrated Analysis Results: Renewable Energy Case Study

Using renewable energy as a case study, we integrated global STR information to analyze development status and trends, future key technologies, and different national technology choices and strategic measures.

Demand Analysis Example

Integrating national renewable energy status and demand information, Timeflow visualization ([Figure 3: see original paper]) reveals that development demands encompass technology, market, policy, environment, and economy—primarily technology, market, and policy needs: - **Technology:** Focuses on performance (orange) and maturity (magenta), such as improving solar collector performance and commercializing advanced biofuels. - **Market:** Focuses on production (red) and consumption (dark red), such as increasing power generation, renewable energy share, and energy demand. - **Policy:** Focuses on planning (green) and support (blue), such as strengthening photovoltaic research, formulating national energy plans, and increasing hydropower projects.

[Figure 3: see original paper] Development Demands in Renewable Energy (Parameters: Classification_1= “today” , Classification_2= “need” , Year$ $2015)

Trend Analysis Example

Integrating national renewable energy vision and trend information, Timeflow visualization ([Figure 4: see original paper]) shows trends spanning market, technology, environment, policy, economy, and social dimensions—primarily technology, market, and environment: - **Market:** Production (red) and consumption (dark red), such as increasing renewable energy share and power generation while reducing prices and incremental costs. - **Technology:** Performance (orange) and cost (brown), such as built-in thermal storage, photovoltaic power generation, plant efficiency, and reduced investment costs. - **Environment:** Emissions (cyan) and resources (purple), such as reducing greenhouse gas emissions and expanding renewable resource development.

[Figure 4: see original paper] Development Trends in Renewable Energy (Parameters: Classification_1= "vision" , Classification_2= "trend & potential & opportunity" , Year$ $2015)

Route Analysis Example

Technology Development Situation

Integrating global renewable energy technology trends, potential, and opportunities, Gephi visualization ([Figure 5: see original paper]) presents the technology development situation from 2015 to 2050. The development focuses on concentrated solar power (CSP), biomass, photovoltaic (PV), and wind power technologies, including performance improvements in cost, CO_2 emissions, and power generation.

Technology Development Direction

Integrating national renewable energy technology target information, Gephi visualization ([Figure 6: see original paper]) shows the technology development direction from 2015 to 2050. The direction concentrates on CSP, PV, hydropower, wind power, biomass, and biofuel applications in transportation, building, and heating sectors.

Technology Development Path

Selecting specific technology keywords, we compare different national target planning for technology development paths using Timeflow. [Figure 7: see original paper] presents development path planning for four key technologies—CSP, carbon capture and storage (CCS), wind power, and PV—across different countries. For CSP, hovering over nodes reveals development status information at each time point, enabling cross-country path comparison.

Strategic Analysis Example

Comparing US and Chinese renewable energy strategies, Timeflow provides comprehensive visualization ([Figure 8: see original paper]):

Vision Comparison

- **US:** Encompasses market, policy, technology, environment, and economy—

primarily market, policy, and technology. Market focuses on production (red): increasing renewable energy power generation and share. Policy focuses on planning (green): sustainable energy development plans and renewable power plant construction. Technology focuses on performance (orange): improving solar and wind efficiency. - **China:** Encompasses market, policy, technology, economy, society, and environment—also primarily market, policy, and technology. Market focuses on production (red) and consumption (dark red): increasing renewable energy generation and capacity, with wind power priced below coal. Policy focuses on planning (green) and support (blue): wind development strategies and incentives for grid companies. Technology focuses on maturity (magenta): advancing wind power technology, large wind farms, and offshore wind.

Route Comparison

- **US:** Focuses on technology performance (orange) and environmental emissions (cyan), such as improving wind/solar efficiency and reducing CO_2 emissions by 80% by 2050 (vs. 2005). - **China:** Focuses on technology performance (orange) and maturity (magenta), such as wind capacity targets (15GW in 2010-2015 $\rightarrow$ 200GW by 2020 $\rightarrow$ 400GW by 2030 $\rightarrow$ 1TW by 2050) and transitioning from onshore to offshore wind and micro-siting technologies.

Measures Comparison

- **US:** Primarily policy support (blue), economic investment (brown: $86 billion annually), and technology performance improvements. - **China:** Primarily policy (planning, regulation, evaluation, support) and market measures (market mechanisms, power market reform, pricing marketization).

Conclusion

This study applied our proposed text mining framework and information extraction methodology to construct an STR information database, using renewable energy as a case study to integrate and analyze global development demands, long-term trends, national technology routes, and strategic measures. Visualization using Timeflow, Gephi, and other open-source tools demonstrates the feasibility of our approach. The method enables rapid grasp of global development status and trends, comprehensive understanding of national long-term routes and measures, and provides efficient strategic intelligence services for decision-makers.

The research remains in a methodological exploration phase with limitations: integration of multiple programming languages and tools requires improved automation, and visualization relies on existing open-source tool functions. Future work should customize functionalities, such as enabling click-through from keywords to associated paragraphs and documents.

References

[3] Amer M, Daim T U, Jetter A. Technology Roadmap Through Fuzzy Cog-

nitive Map-Based Scenarios: The Case of Wind Energy Sector of a Developing Country [J]. Technology Analysis & Strategic Management, 2016, 28(2): 131-155.

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.

[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.

[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 Statement

All authors declare no conflict of interest.

Supporting Data

[1] Xie Xiufang, Zhang Xiaolin. Keywords-network.py. Keyword network construction code.
[2] Xie Xiufang, Zhang Xiaolin. data-20160608-timesplit.csv. Supporting datasets for demand analysis (Figure 3), trend analysis (Figure 4), path 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 situation 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).

Supporting data are stored by the authors, E-mail: xiexiufang@mail.las.ac.cn.

Received: 2016-09-30
Revised: 2016-11-02

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