A Cross-source Topic Fusion and Multi-dimensional Synergistic Indicator Approach for Emerging Technology Identification

Presenter: Haiyun Xu[1,2]

1 Business School, Shandong University of Technology, Zibo 255000, China

2 Max Planck Institute for Solid State Research, Stuttgart 70569, Germany

Contact email: haiyunxu@fkf.mpg.de; xuhaiyunnemo@gmail.com

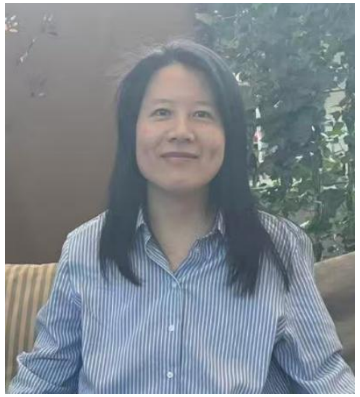
Authors: Xueli Yu[1], Haiyun Xu[1,2], Robin Haunschild[2], Zhengyin Hu[3], Zenghui Yue[4], Chunjiang Liu[3]

1 Business School, Shandong University of Technology, Zibo 255000, China

2 Max Planck Institute for Solid State Research, Stuttgart 70569, Germany

3 National Science Library, Chinese Academy of Science, Chengdu 610299, China

4 School of Medical Information Engineering, Jining Medical University, Rizhao 276826, China



Dr. Haiyun Xu

haiyunxu@fkf.mpg.de; xuhaiyunnemo@gmail.com

Professional Experience

- Dr. Haiyun Xu is a Professor at the Business School, Shandong University of Technology. She received her Ph.D. from the University of Chinese Academy of Sciences and has been a visiting scholar at CWTS, Leiden University, and IVS-CPT, Max Planck Institute for Solid State Research. Her research focuses on science, technology, and innovation management, particularly the identification of emerging and disruptive technologies, interdisciplinary topic discovery, and knowledge transfer.

Publications & Activities

Dr. Haiyun Xu has led multiple national and provincial projects, published over 90 papers, and contributed to several academic reports and books. Dr. Xu was listed among the World's Top 2% Scientists in 2022, 2023, and 2024.

MAX PLANCK INSTITUTE FOR
SOLID STATE RESEARCH

Outline

1. Introduction
2. Literature Review
3. Research Methods
4. Empirical Analysis
5. Validation
6. Conclusion

Introduction

Research Background

Emerging technologies (ETs) are key drivers of scientific innovation and industrial transformation. However, existing identification studies mainly rely on papers and patents, which primarily reflect formalized scientific and technological outputs and are difficult to capture broader signals from industry, policy, funding, and market activities. In particular, patents often exhibit significant time lag in reflecting technological developments, thereby limiting the timeliness and comprehensiveness of ET identification. Integrating multi-source heterogeneous data is essential for enabling earlier detection and more comprehensive identification of technological evolution.

Research Objective

Our study proposes an ET identification method integrating multi-source data fusion and multi-dimensional indicators. By combining BERTopic-based topic modeling and semantic similarity analysis, the study aims to improve the timeliness and comprehensiveness of ET identification, facilitating the early-stage identification of emerging technologies.

Research Significance

The study enriches ET identification research by integrating multi-source data into a unified framework and provides practical support for technology forecasting, policy formulation, and industrial strategic planning.

Literature Review

☑ Concepts and Connotations of Emerging Technologies

There is no unified definition of emerging technologies (ETs). Existing studies generally regard ETs as technologies rooted in scientific discoveries with significant growth potential and transformative impacts on industries (Wharton School, University of Pennsylvania, 2010; Day et al., 2011–2013). **Compared with frontier and core technologies, ETs emphasize early-stage scientific foundations and future potential.**

☑ Feature Framework and Identification Dimensions of Emerging Technologies

Existing ET identification studies mainly focus on scientometric methods and feature frameworks. Traditional scientometric approaches include citation analysis, co-citation analysis, and co-word analysis (Garfield, 1979; Small, 1973; Callon et al., 1983)**. With the advancement of text mining techniques, topic modeling has gradually become a mainstream method for ET identification** (Blei et al., 2003; Mimno et al., 2011; Grootendorst, 2022)**. a representative five-feature framework for ET identification has been proposed, including novelty, growth, coherence, impact, and uncertainty** (Rotolo et al., 2015) **.**

Literature Review

☑ Research Gaps and Proposed Approach

- **Most existing studies rely on single data sources such as papers and patents, which mainly represent formal scientific and technological activities while overlooking signals from industrial deployment, policy orientation, funding support, and market dynamics. Moreover, patent data often lag behind actual technological development. These limitations constrain the breadth and promptness of ET identification.**

- **Therefore, integrating multi-source heterogeneous data has become an important direction for supporting earlier discovery and more holistic understanding of emerging technologies, thereby strengthening decision support**(Rafols et al., 2019)**.**

Research Method

☑ Research Framework

Part I : Data acquisition and preprocessing

Focuses on collecting and preprocessing multi source heterogeneous data, including papers, patents, funding data, industry reports, industrial market data, and policy documents. We conduct query construction, data cleaning, and text standardization to build a unified dataset.

Part II : Topic integration and indicator system construction

We use the BERTopic model to identify technology topics and apply semantic similarity analysis to integrate topics across different data sources, thereby constructing candidate technology topics and a topic co occurrence network.

Part III: ETs identification and validation

Focuses on emerging technology identification. Based on the ET identification indicator system, we evaluate, screen, and verify candidate topics to finally identify emerging technologies.

This framework enables the systematic identification of emerging technologies from multi-source heterogeneous data.

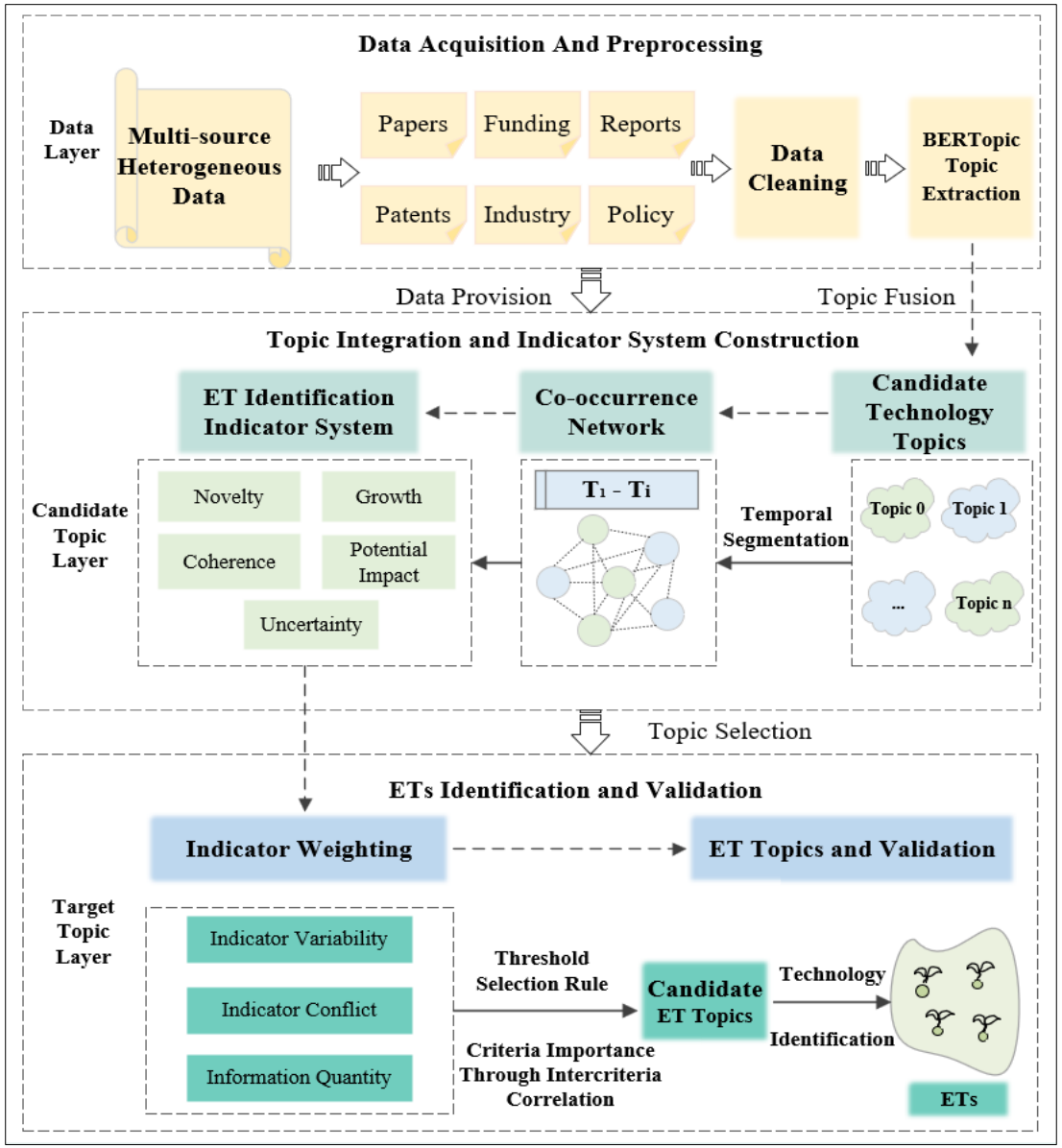


Figure 1 Research Framework

Research Method

☑ Functional Roles of Multi-source Data

- **Different data sources perform distinct functions:** academic papers and patents help locate research frontiers, emerging technologies, core technological trajectories and competitive patterns; funding projects reflect strategic priorities and potential technologies; industrial data and reports enable the evaluation of industrialization and market application prospects; policy documents capture policy-driven technologies; **and the combined use of all sources validates overall technological development trends.**
- Overall, the integration of multi-source data enables a more comprehensive identification of emerging technologies. The functional roles of different data sources are summarized in **Table 1.**

Table 1 Functional mapping of multi-source data sources in emerging technology identification

Data type	Role in technology identification
Papers	Identify research frontiers and emerging technologies
Patents	Identify core technological trajectories and competition
Grants	Identify strategic priorities and potential technologies
Reports	Evaluate industrialization and market application
Industry	Validate technological development trends
Policy	Identify policy-driven technologies

Research Method

Step 1: Technology Topic Extraction

☑ Topic Modeling Method

- **We adopts the BERTopic model.** During the topic modeling process, the model parameters were specified as shown in **Table 2**.
- In particular, we adopted **an adaptive approach** by using the default **"auto" parameter** in BERTopic to automatically determine **the optimal number of topics** according to the semantic density and distribution characteristics of the corpus, thereby **avoiding biases caused by manual specification.**

Table 2 BERTopic Model Parameters

Parameter Name	Value	Description
Language	English	Language type of model training corpus
embedding_model	all-MiniLM-L6-v2	Use the default embedding model
calculate_probabilities	True	Calculate the probability of each topic in each document
top_n_words	10	The number of topic keywords

Research Method

Step 1: Technology Topic Extraction

☑ Topic Evaluation

- We employs the **BERTopic model** to conduct topic modeling on multi-source textual data and evaluates topic identification performance using the **Intertopic Distance Map**.
- **Figure 2** illustrate the comparison of topic structures before and after optimization.
- Based on the UMAP-reduced topic vector space, the map visually presents semantic distances and clustering relationships among topics, where the **relative distance** between topic points reflects **semantic similarity** and the bubble size represents topic weight.
- We selects topic modeling results with **clearer topic boundaries, tighter semantic clusters,** and **less overlap**, thereby improving the interpretability and stability of the topic structure.

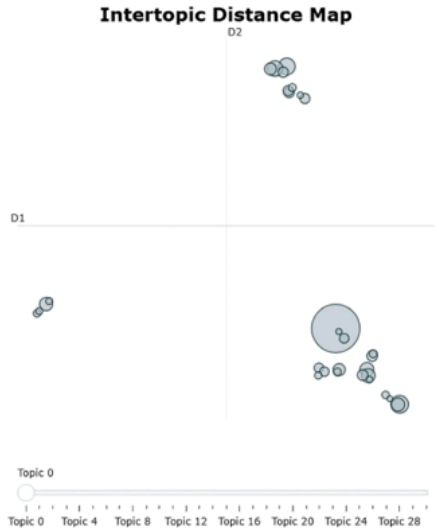


Fig 2-1 Topic clustering results before adjustment

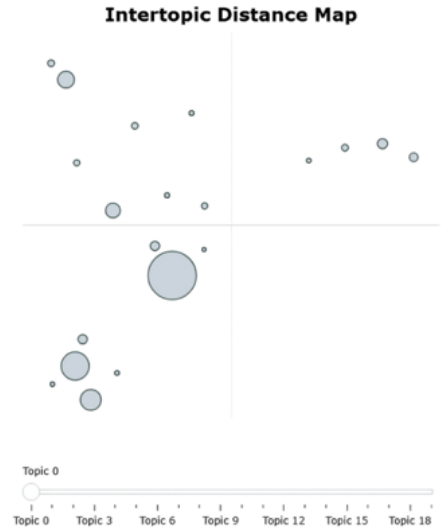


Fig 2-2 Topic clustering results after adjustment

Figure 2 Comparison of Topic Clustering Results

Research Method

Step 2: Technology Topic Fusion

☑ Multi-Source Topic Fusion

- **A two-stage fusion strategy based on cosine similarity is adopted.**
- **A cosine similarity-based strategy is adopted for topic fusion.**
- Paper and patent data can be accurately classified into specific technological directions through refined search strategies, whereas data such as industrial information, funding projects, and policy documents are difficult to categorize with the same level of precision.
- As a result, their topic distributions are often cross-domain, mixed, or semantically ambiguous, making direct alignment with paper and patent topics difficult.
- Therefore, to unify the semantic space and enhance the completeness of technological topics, a two-stage topic fusion process is designed in this study.

Research Method

Step 2: Technology Topic Fusion

☑ Multi-Source Topic Fusion

- **In the first stage**, **paper and patent topics** within the same technological subdomain are fused. Cosine similarity between topic vectors is calculated to perform topic matching and merging, and the mean vector of the matched topics is adopted to maintain semantic balance. This process generates a preliminarily fused topic set within each technological direction.
- **In the second stage**, the directional topics obtained from the first stage are further fused with the shared topics derived from **funding, reports, industrial, and policy data.** Similarity between topic vectors is calculated and topics are merged according to the same screening principle. This process ultimately generates a **global candidate technology topic set** with unified semantics and comprehensive directional coverage across multi-source datasets.

Research Method

Step 3: Indicator System Construction and Topic Identification

☑ Technology Topic Identification

- **TF-IDF** is used to extract keywords and construct **topic co-occurrence networks**.
- **A multi-dimensional indicator system** is then applied, with weights determined by the **CRITIC method** to identify representative technology topics. **(Table 3)**

Table 3 A Multi-dimensional Technological Identification Indicator system

Technology category	Indicator dimension	Specific content
ETs	Novelty	Temporal and semantic novelty.(Xu et al., 2020)
	Growth	Annual average growth rate of topic-supporting documents. (Xu et al., 2020)
	Coherence	Jaccard index and topic autocorrelation coefficient. (Xu et al., 2020)
	Potential impact	The number of multi-source data types.
	Uncertainty	Structural Entropy. (Xu et al., 2022)

Empirical Analysis

☑ Data Collection

- We takes the ***smart grid domain*** as the empirical context.
- We first constructed a preliminary search strategy based on authoritative reports, technical guidelines, and industrial literature related to the smart grid field. Subsequently, professional terminologies were standardized using the terminology system of *the International Electrotechnical Commission and specialized databases*, while search terms were further optimized according to the characteristics of patent and academic databases.
- Through **multiple rounds of pilot retrieval, IPC classification restriction, and noise-data elimination**, the final search strategy was established.

Empirical Analysis

☑ Data Collection

- **As an example, the paper retrieval strategy for the power market subfield is presented as follows:**

TS=((((("electric power" or "electric force" or "electricity") and ("mart*" or "bazaar") and ("market" or "deal" or "Trad*" or "transaction")) or (("electric power" or "electric force" or "electricity") and ((("distributed" or "Distributive") and P2P) or (("completeness" or "quite" or "complete") and P2P) or (("mix*" and "blend*" or "hybrid" or "compound" or "intermixing") and P2P) or "community"))) or ((("Green Financ*“ or "Green Economy") and "new energy") or (("electric power" or "electric force" or "electricity") and "CO2 emissions" and ((("green energy resources" or "green energy") and "consum*" or "depletion" or "expenditure" or "decrement") or "energy efficiency"))) or (("electric system" or "electrical power system*" or "electric power system *" or "power system*") and ("Carbon" and (("mart*" or "bazaar") or ("market" or "deal“ or "Trad*" or "transaction") or "emissions reduction" or "Emission stream"))) or (("electric power" or "electric force" or "electricity") and ("family“ or "household*") and "internet" and (("forecast*" or "predict*") or ("marketing" or "load*" or "charge“ or "weight") or ("trading platform" or "Business platform*") or ("automatization" or "robotization" or "automation"))) or (("electric power" or "electric force" or "electricity") and ("mobile app*" or ("fee" or "expense" or "cost*" or "charge") or ("price" or "prix") or ("Business Hall" or "selling area"))))

And TMIC=("6.115.880 Renewable Energy Transition" or "7.70.919 Energy Systems Optimization" or "4.18.296 Energy Forecasting" or "6.115.234 Carbon Mitigation"or"4.18.788 Electric Vehicles"or"7.139.835 CO2 Capture" or "6.10.80 Market Interdependencies" or "6.10.502 Data Envelopment Analysis" or "4.84.1014 Optimization under Uncertainty" or "2.62.76 Electrocatalysis" or "6.122.1087 Contingent Valuation" or "4.18.472 Power System Stability" or "4.187.2766 Blockchain" or "4.18.101 Power Quality" or "6.115.1554 Circular Economy" or "6.153.558 Climate Change Adaptation" or "2.62.2102 CO2 Electroreduction" or "4.18.754 Doubly Fed Induction Generator" or "4.183.495 Urban Mobility" or "4.18.2795 Real-Time Power Simulation" or "6.115.2292 Energy Security" or "6.115.2022 Energy Economics") and **PY=**(2014-2023) and **DT=**(Article) **NOT TS=**(oscillating or healthcare or Saline Aquifers or catering or helical pile or High-resolution or Aerodynamic or acid or stream or amine or Potato or extraction)

Empirical Analysis

☑ Data Collection

Based on this strategy, multi-source data from **eight smart grid subfields** during **2014–2023** were systematically collected, including **power systems and planning, power generation side, grid side, load side, energy storage side, digital technologies, electricity market and carbon economy, and power materials**. **Table 4** presents the number of multi-source data records collected across the eight smart grid subfields.

Table 4 Distribution of multi-source data records across eight smart grid subfields

Gomain	Subfields	Paper	Patents	Grants	Reports	Industry	Policy
Smart Grid	Power Systems and Planning	10,468	2,193	375	126	496	725
	Power Generation Side	6,034	978				
	Grid Side	3,645	396				
	Load Side	5,848	1,854				
	Energy Storage Side	4,886	2,392				
	Digital Technologies	4,028	2,764				
	Electricity Market and Carbon Economy	5,725	701				
	Power Materials	4,008	593				

Empirical Analysis

☑ Data Processing

- **Table 5** summarizes the **sources** of the multi-source datasets used in this study and their **post-processing sizes**, providing a foundational basis for subsequent analysis and topic modeling.
- Overall, the integration of multi-source data enables a more comprehensive identification of emerging technologies.

Table 5 Data sources and preprocessing details

Data type	Source	Preprocessed data volume
Papers	Web of Science	44642
Patents	Incopat Patent Database	11871
Grants	NSF Database	375
Reports	Conference Reports	126
Industry	Kehui Network, Cigre Database	496
Policy	Peking University Law Database	725

Empirical Analysis

☑ Technical Topic Extraction

- Based on **309 initial topics**, a two-stage cosine similarity fusion is applied to obtain **candidate topics**; combined with the indicator system and **CRITIC weighting** with normalization, **candidate topics** are identified. The process and results are shown in **Table 6**.

Table 6 Data sources and preprocessing details

Step	Method / Strategy	Number of publications / topics
Multi-source data acquisition	Data retrieval and collection	40,248 publications
C and idate technical topics	BERTopic topic modeling	309 topics
Integrated technical topics	Cosine similarity-based fusion	223 topics
Target technical topics	Weight determination using the CRITIC method	117 ETs

Empirical Analysis

☑ Emerging Technology Topics

- **The top 50% of topics** within each of the eight subfields were identified as emerging technology (ET) topics (Bai et al., 2020; Zhou & Min, 2022), resulting in **117 target topics**.
- **Table 7** presents a subset of these topics.

Table 7 Partial list of emerging technology topics

Subfield	Topic ID	Topic name	Document count	Representative keywords
Generation side	Topic 2	heat_efficiency_temperature_thermal	340	heat, efficiency, temperature, thermal, fuel, cell, solar, performance, cycle, exergy
Grid side	Topic 0	energy_microgrid_model_optimal	413	energy, microgrid, model, optimal, storage, microgrids, optimization, proposed, operation, cost
Power materials	Topic 11	gas_natural_co_electricity	21	gas, natural, co-power , electricity, model, power, integrated, fired, systems, technologies
.....				

Validation

☑ Technology Topic Validation

- The identified technology topics show high consistency with authoritative technology foresight reports and policy documents.
- Key Emerging technologies include **solid-state batteries, sodium-ion batteries, structural battery composites, green carbon capture, and osmotic power generation systems**.
- The results verify the effectiveness and practical value of the proposed technology identification framework.

Conclusions

- **First,** this study integrates multi source heterogeneous data to identify emerging technologies in the smart grid field, revealing key development trends such as low carbonization, energy storage, and interdisciplinary integration.
- **Second,** the proposed multi-source data fusion and multi-dimensional indicator framework improves the **comprehensiveness and timeliness** of **early-stage ET identification**, thereby providing stronger support for technology forecasting and policy making.
- **Finally,** the study still has limitations in dynamic evolution analysis and external factor consideration. Future work will incorporate **time series analysis**, **richer data sources**, and **cross field empirical validation** to improve the robustness and adaptability of the framework.

许海云

四川 成都

扫一扫上面的二维码图案，加我为朋友。

Thanks for your attention!

Haiyun Xu

haiyunxu@fkf.mpg.de

xuhaiyunnemo@gmail.com

Max Planck Institute for Solid State Research, Stuttgart 70569, Germany

Business School, Shandong University of Technology, Zibo 255000, China