https://doi.org/10.51408/issi2025_210

# Fusing Multi-Source Data through a Multi-Layer Network for Technological Opportunity Identification

Jinzhu Zhang[1], Mingxia Lu[2], Haoyu Li[3]

[1]*zhangjinzhu@njust.edu.cn*, [2] *lmxluna@163.com*, [3] *lhaoyu@njust.edu.cn*
Nanjing University of Science and Technology, Department of Information Management, Xiaolinwei Street 200, Nanjing (China)

## Introduction

Predicting potential technology opportunities from vast data has been an indispensable research topic (Wang et al., 2023). With data volume and sources growing, identifying these opportunities has become much harder. Therefore, efficient methods for identifying technological opportunities are needed to recognize them accurately and comprehensively.

Previous research has shown that academic papers reveal foundational research topics (Jiang, Yang, & Gao, 2024), patents indicate emerging technology opportunities (Ba et al., 2024), and reviews reflect market-driven innovation potentials (Choi & Kwon, 2023). All of them are used in the research of technological opportunity identification. However, there are still problems with using multi-source data. Some scholars simply splice the texts of multi-source data, construct a single-layer network to extract combinations of knowledge units as technological opportunities. This approach fails to give sufficient consideration to the unique contributions of different data sources. While other scholars construct networks for each data source separately to identify technological opportunities, and then select the common knowledge unit pairs as the finally identified technological opportunities. This approach doesn't fully take into account the integration relationships among different data.

The emergence of multi-layer network theory offers a promising solution to this challenge. A multi-layer network can simultaneously display intra-layer and inter-layer relationships. By regarding each data source as a layer of the network and then constructing a three-layer network to combine the data from the three sources, we can easily integrate multi-source data with the help of the multi-layer network. This structure naturally links basic research, patented technologies, and market feedback, enabling each layer to contribute its unique information while facilitating cross-domain knowledge integration, thereby helping to identify more comprehensive technology opportunities.

This paper proposes a multi-layer network to integrate multi-source data for technological opportunity identification. Specifically, we extract a unified key phrase set from multi-source data, followed by constructing a multi-layer network based on the distinct co-occurrence patterns of these phrases within each data source. The unified phrase set ensures holistic utilization of multi-source information, while the multi-layer network preserves the inherent characteristics of individual data types. By integrating intra-layer relationships (reflecting domain-specific knowledge) and inter-layer connections (bridging cross-domain interactions), this method achieves synergistic integration of multi-source data and explicitly captures their technical interdependencies, thereby enabling comprehensive identification of technological opportunities.

## Data and method

We conduct experiments using data from patents, academic papers, and consumer reviews in the field of new energy vehicles. Firstly, key phrases representing technological elements are extracted from each data source using tailored methods, while semantic unification is performed to address differences in expression across data types, constructing a unified multi-source key phrase set. Secondly, a multi-layer network is built by analyzing the co-occurrence relationships of the multi-source key phrase set across

different network layers, followed by network analysis. Finally, GCN and link prediction are employed to identify technological opportunities within the constructed multi-layer network.

*Data description*

This study selects data from the new energy vehicle sector for the year 2023 as the research subject. Patent data, scientific papers, and user reviews, including titles and abstracts (with review text referring to the content of the reviews), were collected from the Derwent Innovations Index, the SCIE database in Web of Science, and the Edmunds website (Edmunds.com), respectively. The search query was set as TS = ("new energy vehicle*" or "NEV$"), covering the period from Jan 1, 2023, to Dec 31, 2023. A total of 2,437 patent records, 758 academic papers, and 1,790 consumer reviews were retrieved.

*Extraction and Fusion of Multi-Source Technical Knowledge Units*

For patents and papers, TF-IDF extracts high-frequency keywords from texts as knowledge elements. RAKE and KeyBERT respectively capture syntactic and semantic features, with merged results yielding technical elements combining frequency, syntax, and semantics. For product reviews, BERTopic performs topic modeling to extract topical keywords. Syntactically and semantically salient phrases from RAKE/KeyBERT are deduplicated and integrated, deriving topic-relevant core phrases with syntactic-semantic features.

Finally, cosine similarity is used to unify synonymous key phrases. A higher similarity threshold (0.8) is set for patent and paper key phrases, while a lower threshold (0.5) is set for user review key phrases to capture more diverse and colloquial expressions.

*Construction of Multi-Layer Networks*

Technical elements in multi-source data exhibit diverse relationships. A three-layer network is constructed based on the co-occurrence relationships of key phrase sets in the three data types. This network is a multiplex network. The nodes in each layer are the same, all being multi-source key phrase sets, but the edges are different, representing different co-occurrence relationships in each data source.

Next, the edge overlap ratio metric is established to analyze the multi-layer network. The edge overlap ratio refers to the probability of overlapping connections between two network layers, indicating the inter-layer correlation between them. A higher overlap ratio suggests greater similarity in network structure, reflecting stronger inter-layer relationships. This metric measures the proportion of overlapping edges shared by two network layers, thereby reflecting their inter-layer structural similarity.

*Identification of Technological Opportunities in Multi-Layer Networks*

First, node embeddings are generated by integrating node features and local structural information through GCN, capturing semantic features and technical relationships of key phrases. Second, cross-source interaction is modeled by calculating association strength between data sources via edge overlap ratio, with weighted embeddings from support layers. Finally, link prediction is performed using optimized node embeddings to evaluate potential technical opportunities between key phrases.

In this study, positive samples are real key phrase pairs from the dataset, while negative samples are generated through random sampling. The dataset is split into training, validation, and test sets at an 8:1:1 ratio. Using the patent layer as the target domain and integrating papers and reviews as support layers, the model outputs link prediction scores reflecting technical association strength.

**Result**

We employed the accuracy, F1 score and AUC as evaluation metrics to assess the effectiveness of link prediction using only patent data and using multi-source data in a multi-layer network. The results are shown in Table 1, indicating that link prediction using the multi-source data multi-layer network has certain improvements in various indicators. The increase in the AUC value indicates that this method is better in the ability to distinguish between positive and negative samples; the rise in accuracy reflects the enhanced accuracy and reliability of the

prediction results; the improvement of the F1 value further proves the enhancement of the comprehensive performance of this method.

**Table 1. The effectiveness of different methods.**

| | Accuracy | F1 | AUC |
|---|---|---|---|
| Patent Data | 56.67% | 69.77 % | 95.36 % |
| MSML | 62.04% | 72.46 % | 96.00 % |

## Conclusion

This paper enriches the data sources for technological opportunity identification and applies the methods of deep learning and complex networks to make full use of multi-source information through a multi-layer network approach. Compared with single-data-source methods, this method performs better. In the next step, we plan to apply other deep-learning methods, which may perform more outstandingly in the semantic representation of multi-source phrases.


## Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 72374103, 71974095), China Society of Indexers (Grant No. CSI24C10), and Jiangsu Provincial Federation of Philosophy and Social Sciences (Grant No. 24SYC-023).


## References


Wang, J., Zhang, Z., Feng, L., Lin, K.-Y., & Liu, P. (2023). Development of technology opportunity analysis based on technology landscape by extending technology elements with BERT and TRIZ. *Technological Forecasting and Social Change*, 191, 122481.

Jiang, M., Yang, S., & Gao, Q. (2024). Multidimensional indicators to identify emerging technologies: Perspective of technological knowledge flow. *Journal of Informetrics*, 18(1), 101483.

Ba, Z., Meng, K., Ma, Y., & Xia, Y. (2024). Discovering technological opportunities by identifying dynamic structure-coupling patterns and lead-lag distance between science and technology. *Technological Forecasting and Social Change*, 200, 123147.

Choi, K. H., & Kwon, G. H. (2023). Strategies for sensing innovation opportunities in smart grids: In the perspective of interactive relationships between science, technology, and business. *Technological Forecasting and Social Change*, 187, 122210.