https://doi.org/10.51408/issi2025_183

Research on Scientific Frontier Topics Based on Citation Analysis and Content Analysis - Taking the Structural Analysis of Nature Index as an Example

Tan Xiao[1], Li Hui[2], Xu Haiyun[3], Li Jiayu[4], Jin Xiaohong[5], Xi Guiquan[6], Zhang Ting[7], Chen Shu[8]

[1]*tantan46227@163.com*, [2]*tantan83_1@163.com*, [4]*meganli328@hotmail.com*, [5]*531277608@qq.com*, [6]*81289113@qq.com*, [7]*tanx@bjast.ac.cn*, [8]*c.s.luck@163.com*
Institute of Scientific and Technical Information, Beijing Academy of Science and Technology, No.140, Xizhimenwai street, Xicheng District, Beijing,100044 (China)

[3] *xuhaiyunnemo@gmail.com*
School of Management, Shandong University of Technology, No. 266, West Xincun Road, Zhangdian District, Zibo City, Shandong Province, 255000 (China)

Introduction

In the era of global scientific and technological revolution, interdisciplinary integration prevails. Scientific frontier research, emerging after 2005, focuses on identifying technological trends. The Nature Index (NI) measures research output but lacks content analysis. This paper uses NI data, integrating entity-relationship and semantic-structure, to explore frontier topics via text modeling and automatic identification, aiding researchers in spotting hotspots and directions.

Literature Review

Research frontiers, rooted in novel discoveries, exhibit bibliometric patterns like term surges and citation network shifts (Chen, 2009). Detection methods, including Delphi and ML-based ones, fall into citation-clustering and indicator-driven categories.Metrological methods for scientific frontiers rely on high co-occurrence/clustering, using disciplinary knowledge networks (Chen, 2010). Traditional methods lack semantic analysis; this paper introduces semantic-structure approaches (Blei, 2003).

Research Steps for Topic Identification and Structural Detection

This study selects 68 NI journal literatures from 2018-2020, preprocesses keywords, vectorizes text to obtain matrices, constructs an LDA model, builds and standardizes a co-citation matrix, calculates cosine similarity, fuses matrices linearly per Janssens' idea, and uses community-discovery algorithms to identify research frontiers.

Description of the Experimental Dataset Construction

We selected 170,000 papers from 68 NI journals (2018-2020), used citation coupling to build relationships, standardized them, and stored data in SQL Server with ESI_IDs and normalized cocitation degrees.

Construction of Similarity Matrices Based on LDA Modeling

Using LDA model, we map documents to topic vectors, extract keywords from titles/abstracts, apply Gibbs sampling with α=50/K, β=0.01, set K=50 via perplexity, iterate 2000 times, and compute doc similarity by cosine.The topic vector mapped by the article is $d_i = (t_1, t_2, \ldots\ldots, t_K)$, the similarity between two articles is calculated using

$$Sim(d_i, d_j) = \cos\theta = \frac{d_i * d_j}{|d_i| * |d_j|} = \frac{t_1 * t_1' + t_2 * t_2' + \ldots + t_k * t_k'}{\sqrt{t_1^2 + t_2^2 + \ldots + t_k^2} * \sqrt{t_1'^2 + t_2'^2 + \ldots + t_k'^2}}$$

(Formula 1) to obtain the document-document similarity matrix.

Integration of Citation Relationships and Semantic Content

It is necessary to further refine the text knowledge units. The integration method applied in this part is a comprehensive method of topic mining and Fisher relationship fusion algorithm. The integration method applied in this part is a comprehensive method of topic mining and Fisher relationship fusion algorithm.

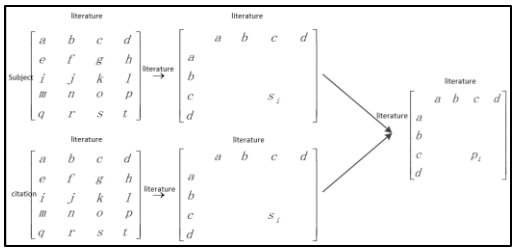

Figure 1. Framework of the Relationship Fusion Method Based on Fisher.

according to the semantic information, the correlation relationships in the network will be adjusted, and their weights will be adjusted:

$$Sim(i,j) = \begin{cases} \lambda \cdot Sim_{LDA} + (1-\lambda) \cdot cite_{couple}, \text{cite}(i,j) = 0 \text{ and } sim_{LDA}(i,j) > a; \text{cite}(i,j) > 0 \\ \quad \text{and } sim_{LDA} \geq b \quad \lambda \epsilon (0,1) \\ 0, cite(i,j) > 0 \text{ and } sim_{LDA}(i,j) < b\ ;\ \text{cite}(i,j) = 0 \text{ and } sim_{LDA}(i,j) \leq a \end{cases}$$

(Formula 2)

Community Detection and Topic Extraction

Community detection differs from traditional clustering by focusing on network node-link relationships. This study fused standardized citation networks and LDA-based similarity, then applied the Louvain algorithm—optimizing modularity via iterative community division and aggregation—for efficient large-scale community detection.

For the literature related to NI from 2018 to 2020, the steps in 3.3 were used for analysis, and secondary clustering of the communities was performed to form the following figure.

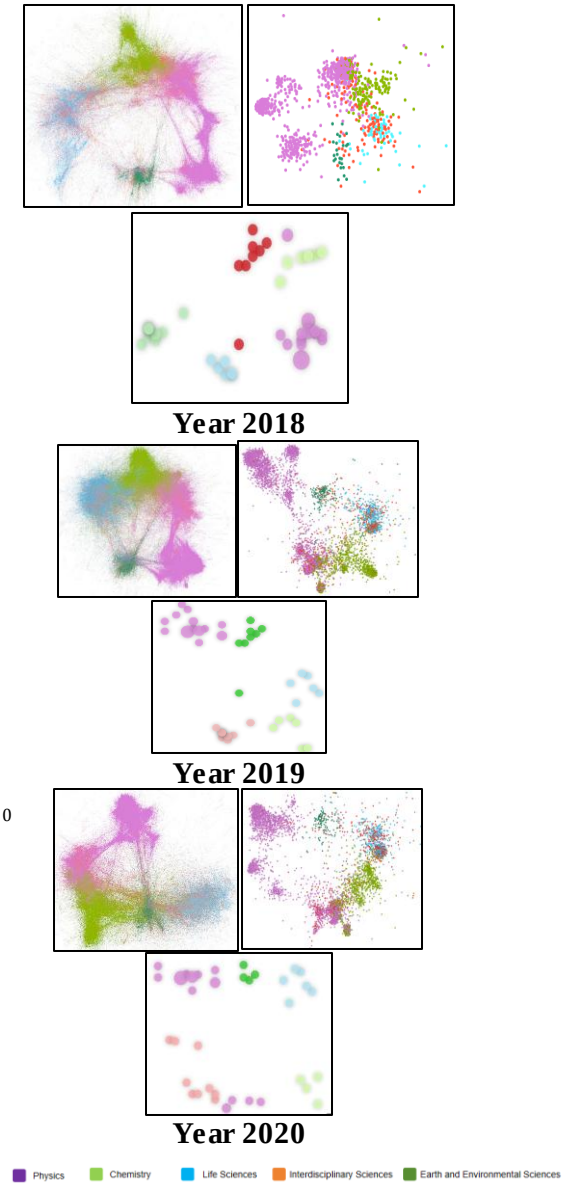

Figure 2. Clustering Diagram Formed by Applying Community Detection Algorithm from 2018 to 2020.

Each circle represents a research topic and consists of several pieces of literature, and its size is positively correlated with the number of documents that constitute the community. In the clusters formed by re-clustering, the keywords representing each cluster are extracted to form the current domain topics and the cross-topic relationships between disciplinary fields.

From an overall analysis, it is found from the community clustering diagrams of the three years from 2018 to 2020 that the highly

interdisciplinary research fields are showing an upward trend; the interdisciplinary integration is becoming wider and wider, showing a spreading trend.

Conclusion

This paper detects frontier field dimensions via "entity-relationship" and "semantics-structure" integration, using 3-year NI journal data with citation-topic combination to identify frontier topics, aiding interdisciplinary research. Limited by space, it only shows overall structural characteristics. Future research will incorporate diverse S&T documents (plans, patents, projects) to construct dynamic knowledge networks, integrate multi-entity relationships, strengthen entity attributes, classify frontier topic types, and enhance identification accuracy by linking topics to classification attributes.

Acknowledgments

This article is the outcome of the projects,"Research on the emergence mechanism and identification of disruptive technologies from the perspective of innovation ecological network" (No.9242006) supported by Beijing Natural Science Foundation, "Early Recognition Method of Transformative Scientific and Technological Innovation Topics based on Weak Signal Temporal Network Evolution analysis" (No.72274113) supported by the National Natural Science Foundation of China and the Taishan Scholar Foundation of Shandong province of China (tsqn202103069).

References

Chen, S. J. (2009). Survey of Approaches to Research Front Detection. *New Technology of Library and Information Service*, 28 - 33.

Chen, L. J. (2010). Research on the Three Evolution Stages of Knowledge Linking Theory and Practice. *Library and Information Service*, 54(12), 46 - 49+63.

Wang, W. B., Cheng, H. M., & Wang, L. J. (2013). Analysis of the Research Situation of China's Strategic Emerging Industries Based on Co-word Analysis. *Science & Technology Progress and Policy*, 30(21), 57 - 60.

Li, G., & Li, Y. (2011). A Co-word Analysis Method Based on Keyword Weighting. *Information Science*, 29(03), 321 - 324+332.

Bai, R. J., Zhang, Y. H., Zhang, Y. J., Ju, Z. H., & Feng, M. Y. (2024). Research on the Identification of Interdisciplinary Frontiers Based on Dual Measurement of Citation and Theme. *Journal of Modern Information*, 44(10), 27 - 40+63.