https://doi.org/10.51408/issi2025_178

Recent Advance of Text Mining in LIS: A bibliometric review

Siqi Hong[1], Guo Chen[2]

[1] 3402202918@qq.com, [2] delphi1987@qq.com
Nanjing University of Science and Technology, No. 200 Xiao Ling Wei, Nanjing, Jiangsu (China)

Introduction

Text mining has become an essential tool in Library and Information Science (LIS), yet systematic reviews remain scarce. Early reviews mainly provided technical overviews, while recent research has expanded into specific application areas.

Based on this gap, this study analyzes text mining research in LIS from 2022 to 2024. We first apply topic modeling to identify key research directions, then focus on three core questions:

1) What types of texts are studied?
2) What technologies are used?
3) What are the main application scenarios?

To answer these, we propose a "Text–Technology–Scenario" three-dimensional framework that examines LIS text mining from the perspectives of research objects (what), methodologies (how), and application value (why), offering a structured view of its current landscape and future trends.

Data and methods

This study analyzed 1,726 text mining-related papers published between 2022 and 2024 in 39 authoritative LIS journals (23 Chinese, 16 English). A Python-based tool was used to collect bibliographic data, followed by keyword screening and manual verification. Chinese papers were automatically translated using a LLM API.

To identify research trends, we applied BERTopic and resulting in 17 major topics (Figure 1).

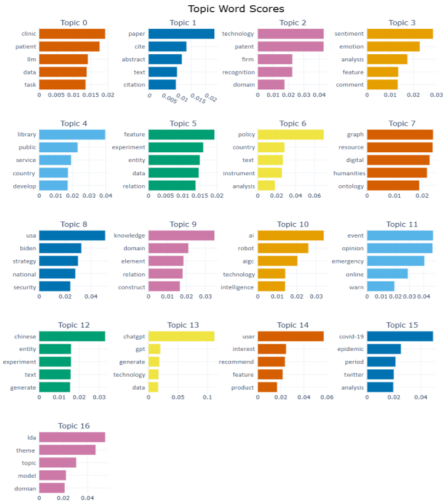

Figure 1. Visualization of topic word distribution.

We then conducted content analysis to interpret each topic. Based on this, we developed a three-dimensional framework (Figure 2) to categorize the findings into three analytical perspectives.

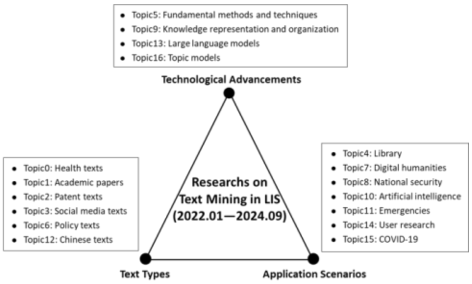
Topic5: Fundamental methods and techniques
Topic9: Knowledge representation and organization
Topic13: Large language models
Topic16: Topic models
Technological Advancements
Topic0: Health texts
Topic1: Academic papers
Topic2: Patent texts
Topic3: Social media texts
Topic6: Policy texts
Topic12: Chinese texts
Researchs on Text Mining in LIS (2022.01—2024.09)
Topic4: Library
Topic7: Digital humanities
Topic8: National security
Topic10: Artificial intelligence
Topic11: Emergencies
Topic14: User research
Topic15: COVID-19
Text Types
Application Scenarios

Figure 2. A Three-Dimensional Framework of Text Mining Research in LIS.

Results and discussion

Text Type Perspective

Different types of texts present distinct challenges and research priorities in LIS text mining. Table 1 summarizes major topics associated with six representative text types.

Table 1. Summary of Topics by Text Type.

Topic	Text Type	Key Focus
T0	Health Texts	Clinical decision support, disease prediction, knowledge services
T1	Academic Papers	Word/citation/topic-level mining, scientific evaluation
T2	Patent Texts	Tech evolution, opportunity detection, entity mapping
T3	Social Media	Sentiment analysis, misinformation detection
T6	Policy Texts	Cross-national analysis, coordination issues
T12	Chinese Texts	Classical/ancient texts, cultural NLP tasks

In health texts, Chinese research emphasizes knowledge services, while English studies focus on clinical applications—with LLMs widely applied in both. For academic papers, the focus has shifted from feature extraction to semantic understanding, with growing use of bibliometric methods for scientific evaluation. Patent research highlights cross-domain opportunities and merges patent with social media data for trend prediction. Social media research has moved from public opinion tracking to sentiment analysis and misinformation detection. Policy text mining is more active in Chinese but remains methodologically limited. Research on Chinese texts focuses on linguistic heritage such as classical Chinese and minority languages.

Technological Perspective

Recent LIS research has actively explored new methods to enhance semantic understanding and task performance. Table 2 highlights four representative topics from technological standpoint.

Table 2. Summary of Topics by Technological Focus.

Topic	Technology Focus	Key Themes
T5	Core Methods	Prompt learning, multimodal fusion, NER, classification
T9	Knowledge Representation	Entity/tuple/document-level representation for organization and application, interdisciplinary knowledge mining
T13	Large Language Models (LLMs)	ChatGPT applications, opinion mining
T16	Topic Modeling	LDA+BERT, trend analysis, user modeling

LIS text mining shows dual momentum: performance breakthroughs and knowledge-centered exploration. Core methods are enhanced via deep learning and multimodal fusion, especially in low-resource settings. Meanwhile, knowledge representation is advancing from carrier-level to semantic-level modeling.

LLMs like ChatGPT empower tasks such as summarization and entity extraction, while also raising concerns around ethics and hallucinations. Topic models continue evolving through integration with BERT and transfer learning, expanding to new domains like policy and culture. Overall, LIS research is shifting toward intelligent, multimodal, and domain-adaptive methods.

Application Scenario Perspective

LIS researchers are applying text mining to a wide range of practical domains with diverse goals and methods. Table 3 outlines seven prominent application areas identified in the literature.

Table 3. Summary of Topics by Application Scenario.

Topic	Application Scenario	Key Focus
T4	Libraries	Resource organization, service design
T7	Digital Humanities	Cultural heritage mining, ontology construction
T8	National Security	Policy analysis, strategic insight
T10	Artificial Intelligence	Chatbots, AIGC applications, ethical issues
T11	Emergencies	Opinion evolution, knowledge graphs for events
T14	User Research	Recommendation, demand mining, satisfaction analysis
T15	COVID-19	Sentiment evolution, multilingual health texts mining

In libraries, it supports the smart organization of digital resources; in digital humanities, it aids the analysis of cultural heritage; and in national security, it enhances policy and intelligence research. Emergency-related studies widely adopt ontologies and knowledge graphs to enable semantic understanding and causal reasoning, improving event inference and decision-making. User research has shifted from global to short-term interest modeling, with applications expanding beyond e-commerce to areas such as academic citation and community Q&A. In the context of AI, the rise of generative models has brought growing attention to ethical risks, social impact, and governance issues. COVID-19 research highlights multilingual analysis of public sentiment, health information, and pandemic trends, with increasing focus on vaccine safety, drug efficacy, and mental health.

Conclusions

Overall, text mining research in LIS exhibits several notable trends:

(1) **From intelligence-centered to interdisciplinary integration.** Research has expanded from scientific texts to policy, culture, and health domains, aligning LIS with public administration, digital humanities, and health informatics.

(2) **Large Language Models and Generative AI as new drivers.** These technologies enhance core tasks(Li, Peng & Li, 2024) and introduce new research directions like hallucination detection and content authenticity assessment.

(3) **Text mining and bibliometrics: an evolving synergy.** Their integration enables efficient processing of unstructured data and robust scientific evaluation and trend forecasting (Luo, Lu & He, 2022).

Acknowledgments

This work is supported by Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No.SJCX24_0178).

References

Li, Y., Peng, X., Li, J., Zuo, X., Peng, S., Pei, D., ... & Hong, N. (2024). Relation extraction using large language models: a case study on acupuncture point locations. Journal of the American Medical Informatics Association, 31(11), 2622-2631.

Luo, Z., Lu, W., He, J., & Wang, Y. (2022). Combination of research questions and methods: A new measurement of scientific novelty. Journal of Informetrics, 16(2), 101282.