Text Mining and Information Retrieval Optimization of Large Language Models in Digital Libraries

Xujie Wang[1], Yu Wang[2]

[1]Ganymede (Beijing) Technology Co., Ltd. Beijing 100000
[2]Beijing Datianmian White Sugar Technology Co., Ltd. Beijing 100000

Abstract: ***Digital libraries are key knowledge management and dissemination platforms in the information age, providing a large amount of literature resources that are crucial for academic research and daily information acquisition. But with the increase of data volume, effectively mining information and improving retrieval efficiency have become challenges. Although traditional text mining and information retrieval techniques have some effectiveness, they still have shortcomings in semantic understanding and complex query processing. In recent years, the development of natural language processing technology, especially large language models (LLMs), has significantly improved text processing capabilities. LLM performs excellently in semantic understanding, text generation, and knowledge extraction through deep learning and large-scale pre training, with strong generalization ability. This article explores how to apply big language models to improve text mining and information retrieval techniques in digital libraries, in order to enhance user retrieval experience and achieve intelligent and personalized information services.***

Keywords: Big language model; LLM; Text mining; Information retrieval; Intelligent Q&A; Intelligent digital library.

1. INTRODUCTION

1.1 Research Status

1.1.1 The main text mining and information retrieval technologies currently used in digital libraries

The digital library is a key platform for modern information storage and management, providing rich digital resources to support academic research, education, and daily queries. In order to effectively manage and retrieve large-scale and diverse literature resources, digital libraries have adopted technologies such as keyword matching (achieved through Boolean search and inverted indexing), TF-IDF (word frequency inverse document frequency), LDA (latent Dirichlet allocation), and sorting algorithms such as PageRank. These methods are somewhat efficient but lack the ability for deep semantic understanding and processing complex queries. Lu et al. (2025) developed NeuroDiff3D, a diffusion-based method for 3D generation that optimizes viewpoint consistency, enhancing the quality of synthesized digital objects [1]. Similarly, Hu (2025) focused on accessibility by proposing a low-cost 3D authoring pipeline driven by guided diffusion and graphical user interfaces (GUI) [2].

Concurrently, there is a growing emphasis on making AI systems more transparent and actionable for practical decision-making. Zhang (2025) addressed this in the financial sector by proposing an adaptive Explainable AI (XAI) framework designed to transform opaque models into actionable insights for proactive tax risk mitigation in small and medium enterprises [3]. This focus on practical application extends to business intelligence and digital advertising. Tian et al. (2025) introduced a cross-attention multi-task learning approach to innovate ad recall strategies [4]. Complementing this, Zhang (2025) proposed AdOptimizer, a self-supervised framework aimed at efficient ad delivery in low-resource markets [5]. For broader campaign management, Li, Wang, and Lin (2025) enhanced sequential recommendation for cross-platform ads using graph neural networks [6].

The robust perception and understanding of physical environments remain a critical challenge. Xie et al. (2025) advanced this area with MARNet, a multi-scale adaptive relational network for completing 3D point clouds via cross-modal fusion, improving environmental digitalization [7]. Beyond commerce, AI is also being leveraged for public and logistical planning. Xu (2025) explored generative modeling for public space development in the project CivicMorph [8], while Zhang (2024) applied cohesive hierarchical clustering to dynamically adapt the supply and demand of power emergency materials, optimizing critical resource logistics [9].

Finally, ensuring the reliability and efficiency of the underlying digital and network infrastructure is paramount. Tu (2025) developed AutoNetTest, a platform-aware framework for automating 5G network testing and issue diagnosis [10]. Xie and Liu (2025) optimized industrial monitoring systems for real-time analysis by integrating OpenCV and WebSocket in their InspectX project [11]. System reliability is further addressed by Zhu (2025) in REACTOR, a framework for automated causal tracking and observability reasoning in reliability engineering [12].

With the development of natural language processing and big language modeling technology, it is expected to further improve the retrieval efficiency and user experience of digital libraries. The big language model uses deep learning to understand the grammar, semantics, and contextual associations of language, optimizing text mining and information retrieval, especially in semantic retrieval and intelligent document processing, showing great potential to promote the development of digital libraries towards a more intelligent and personalized direction.

1.1.2 Limitations of Traditional Technologies

Traditional text retrieval techniques face challenges in digital libraries:

(1) Low retrieval efficiency: The increase in literature volume leads to low efficiency in keyword matching and Boolean retrieval, returning a large amount of redundant information, making it difficult to quickly and accurately meet the needs.

(2) Limited text comprehension ability: can only recognize surface keywords, unable to understand deep semantics, resulting in relevant literature being overlooked or results not meeting user needs.

(3) Weak ability to handle complex language: unable to deal with language phenomena such as synonyms, ambiguity, and metaphors, and unable to recognize semantic equivalence of different expressions.
(4) Insufficient personalized recommendations: Lack of deep learning and behavioral analysis capabilities, difficulty in understanding users' long-term interests and needs, and poor effectiveness of personalized recommendations.

These challenges indicate that the application of advanced natural language processing and large language modeling technologies is necessary in digital libraries to improve retrieval efficiency and semantic understanding, and provide better personalized services.

1.1.3 The successful application of big language models brings new opportunities for digital libraries

In recent years, the rise of Large Language Models (LLMs) has sparked a revolution in the field of Natural Language Processing (NLP). By pre training based on large-scale datasets, these models have demonstrated excellent language understanding and generation abilities, and have shown outstanding performance in many tasks, bringing new opportunities for text mining and information retrieval in digital libraries.

(1) Machine translation: Large language models such as Alibaba's Tongyi Qianwen and OpenAI's GPT have demonstrated extremely high accuracy and fluency in machine translation tasks. These models can not only handle common language translations, but also capture subtle differences between languages in multilingual translation, which is very beneficial for digital libraries with abundant multilingual resources.

(2) Sentiment analysis: In sentiment analysis tasks, models such as BERT can accurately determine the emotional orientation of text through a deep understanding of the context. This provides important support for user evaluation, book recommendation, and other services in digital libraries, enhancing the library's ability to respond to users' emotional needs.

(3) Question answering system: Large language models have been widely used in question answering systems, and models such as GPT-3 can provide coherent and accurate answers based on context. Compared to traditional information retrieval, big language models can not only directly generate answers, but also handle complex problems and reasoning tasks, providing digital library users with a more intelligent way of interaction.

(4) Information generation and document summarization: The ability of big language models in text generation enables them to quickly generate document summaries or summaries, helping users extract core information from massive literature and enhance the user experience of digital libraries.

2. CHALLENGES OF TEXT MINING AND INFORMATION RETRIEVAL IN DIGITAL LIBRARIES

With digital libraries becoming an important platform for knowledge acquisition, text mining and information retrieval face challenges of large-scale, diverse literature, and complex user needs.

2.1 Literature Diversity and Language Complexity

The types of literature in digital libraries are diverse, including structured and unstructured data such as academic papers, books, datasets, etc. These documents have language complexity, especially in cross lingual documents (such as Chinese, English, Arabic) where there are differences in semantic understanding. Traditional retrieval methods are difficult to handle the subtle semantic differences in multilingual documents.

2.2 Difficulties in Large scale Literature Processing

Faced with massive literature, traditional retrieval methods, although using indexing techniques to improve query speed, have limitations in handling complex queries and deep semantic understanding. Especially when dealing with large amounts of data or complex queries, ensuring system response speed and stability becomes a challenge.

2.3 Limitations of Traditional Information Retrieval Technologies

Traditional information retrieval methods such as keyword matching and Boolean retrieval rely on user input keywords, making it difficult to understand the semantics behind the query. For example, when users search for "climate change", the system may not be able to retrieve literature related to "global warming". Meanwhile, although reverse indexing and other technologies accelerate query speed, their semantic understanding ability is limited and they cannot handle synonyms and contextual changes.

2.4 Insufficient semantic understanding and text generation ability

Traditional retrieval techniques perform poorly in semantic understanding and cannot deeply analyze the intentions behind user queries. This results in the system being unable to provide personalized recommendations or automatically generate relevant summaries. Meanwhile, traditional technologies lack the ability to generate text and are unable to provide users with intelligent and dynamic literature recommendations and explanations.

2.5 Challenges of User Experience and Personalized Needs

The retrieval needs of digital library users are diverse, and researchers, students, enthusiasts, and others have different goals and expectations. Traditional systems are difficult to meet these diverse needs, usually only returning results based on keywords and unable to dynamically adjust to meet personalized recommendations and complex semantic requirements.

2.6 Information overload and long tail literature issues

The increase in the number of literature has brought about the problem of information overload, making it difficult for users to quickly obtain valuable information from massive literature. At the same time, the system usually prioritizes recommending popular or latest literature, ignoring long tail literature that has less demand but is important. The effective presentation and recommendation of long tail literature remains an important challenge in information retrieval.

3. THE APPLICATION OF LARGE LANGUAGE MODELS IN DIGITAL LIBRARIES

3.1 Optimizing Text Mining

(1) Semantic Understanding and Text Classification

The Large Language Model (LLM), with its deep semantic understanding ability, can accurately identify complex language patterns and contextual relationships in text, significantly improving the accuracy of text classification. In digital libraries, LLM can automatically classify literature into different disciplines, not only identifying obvious themes, but also inferring implicit themes from text details, providing fine classification dimensions based on themes, purposes, methods, or conclusions.

(2) Exploring themes, emotions, and author style

The Large Language Model (LLM) performs excellently in topic mining, extracting core topics from a large number of literature, especially identifying cross disciplinary topics in a multidisciplinary environment, and providing researchers with literature clustering in specific directions. LLM can also analyze the emotional tendency of text, determine the stance (support, opposition, or neutrality) in literature, and infer the author's personal style or identify multiple authors of collaborative articles based on writing style.

(3) Information Extraction: Entity Recognition and Relationship Extraction

Information extraction is a key task in text mining, especially when researchers need to extract structured information from unstructured data. LLM has significant advantages in Named Entity Recognition (NER), as it can accurately identify various entities such as people's names, place names, organizational names, and chemical substances from literature. Furthermore, LLM can also perform relationship extraction by identifying relationships between entities, such as collaborations between authors and connections between experimental results and conclusions. This ability can greatly facilitate researchers in obtaining the necessary information from a vast literature.

3.2 Improve the accuracy and efficiency of information retrieval

(1) Accurate retrieval under semantic understanding

Traditional information retrieval is mostly based on keyword matching, but this approach is prone to missing synonyms, variant words, and the semantics behind complex query intentions. The big language model can understand the semantics behind user queries, thereby providing more accurate search results. For example, if a user inputs "the latest methods to improve image recognition accuracy," LLM can not only match keywords such as "image recognition" and "accuracy," but also understand the essence of the problem and provide relevant academic literature, rather than just literature containing these keywords.

(2) Literature search and sorting optimization

Pre trained large language models can more accurately measure the semantic similarity between queries and literature content. In this way, LLM can not only find literature related to user queries, but also perform semantic priority sorting on literature. This search and sorting optimization ensures that users not only obtain relevant literature, but also prioritize viewing high-quality literature that is most likely to answer their questions or meet their needs.

(3) Retrieval and recommendation of long tail literature

Long tail literature, which refers to those that are not very popular and have low citation rates, often finds it difficult to obtain recommendations through traditional information retrieval systems. The big language model has the ability to process and understand such literature, even if the keyword matching of long tail literature is not high, the model can still identify its potential value through deep semantic analysis. The recommendation ability of this long tail literature can help users discover hidden research treasures and enhance opportunities for academic innovation.

3.3 Personalized recommendation system

(1) Personalized Literature Recommendation Based on Large Language Model

The personalized recommendation system in digital libraries can provide literature recommendations based on users' historical behavior, interests, and needs, while the large language model can provide more targeted recommendations through semantic analysis of users' historical search and browsing records. For example, if a user continues to follow a research direction, LLM can identify this tendency and prioritize recommending the latest research results in that field. At the same time, LLM can explore potential related fields in user interests and recommend research that users may not have explored but may be interested in.

(2) Improve the accuracy and coverage of recommendations

Traditional recommendation systems rely on mechanisms such as collaborative filtering and content filtering, while large language models can analyze the semantic content of user reading literature through deep learning methods, automatically construct user interest maps, and generate personalized recommendation lists. This method not only improves the accuracy of recommendation results, but also greatly expands the coverage of the recommendation system. For example, LLM can recommend literature with similar styles, cite relevant research, and even discover potential connections between different disciplines based on the content of the article the user is currently reading, providing interdisciplinary recommendations.

3.4 Question and Answer System and Intelligent Interaction

(1) Intelligent Assistant Based on Large Language Model

The question answering system in digital libraries can provide users with an intelligent interactive experience. The big language model has significant advantages in answering natural language questions, as it can provide direct and concise answers based on user questions without requiring users to browse a large amount of literature. For example, users can ask questions such as "What are the latest advances in modern deep learning?" LLM can synthesize multiple literature and provide accurate summaries or recommendations. Compared to traditional keyword search, question answering systems can provide users with information more efficiently.

(2) The combination of question answering system and retrieval system

LLM can be combined with information retrieval in question answering systems to enhance the system's responsiveness. When users ask complex questions, the question answering system will call LLM to perform semantic analysis on the user's query, quickly retrieve and generate answers from the massive literature in the digital library. This method not only saves users' reading time, but also ensures the accuracy and depth of the answers. For example, when users search for cutting-edge technologies in a certain field, LLM can generate short and clear summaries based on comprehensive information from multiple literature, helping users quickly understand the latest developments.

4. CHALLENGES AND FUTURE PROSPECTS

4.1 Future research directions

4.1.1 Improve the efficiency and low-power operation of large language models

With the expansion of big language models in digital library applications, their high computational costs and energy consumption issues are becoming increasingly apparent. Future research will focus on improving model computational efficiency and optimizing energy consumption, using techniques such as model compression, parameter sharing, and knowledge distillation to reduce the number of parameters while maintaining efficient reasoning ability. Optimizing computational complexity is crucial for improving the efficiency of large-scale data processing in information retrieval and text mining tasks. Model compression reduces hardware requirements and energy consumption by removing redundant calculations and pruning methods, while quantization techniques simplify the calculation process to adapt to low resource environments, ensuring efficient mining and rapid response to large-scale text data in digital libraries.

4.1.2 Expanding the Retrieval Capability of Multimodal Data

With the diversification of data formats in digital libraries, big language models will be extended to multimodal data processing and retrieval, integrating pre trained models of visual, language, and other modalities to achieve comprehensive retrieval and mining of text, images, and videos. This will enhance the user retrieval experience and advance fields such as image recognition, audio transcription, and text matching, meeting a wider range of research needs.

4.1.3 Enhance the performance of the model in specific domain knowledge processing

The big language model performs well in general domains, but needs to improve its understanding and generalization ability of specialized knowledge when dealing with literature in specific domains. Future research will combine expert knowledge bases and domain specific datasets to enhance the model's processing capabilities on complex corpora and long tail knowledge, and improve its performance in professional domains by fine-tuning pre trained models. The research on cross domain knowledge transfer will enhance the performance and generalization ability of the model in multi domain literature, thereby improving the accuracy and efficiency of text mining and information retrieval in digital libraries.

4.2 Prospects for the Development of Digital Libraries

4.2.1 Enhancing the Intelligence Level of Digital Libraries through Large Language Models

The future digital library will transform into an intelligent knowledge exploration platform, and the big language model will revolutionize the way literature is processed, optimizing multiple aspects such as information retrieval, user interaction, literature classification, and annotation. For example, a question answering system combined with a large language model will allow users to acquire knowledge through natural language, while a personalized recommendation system will improve the efficiency of literature discovery. Artificial intelligence will also drive automated functions such as literature abstract generation and topic analysis, improving library management and maintenance efficiency. With the advancement of automation technology, librarians can better manage large amounts of literature, and users can more easily find literature that meets their needs.

4.2.2 Intelligent Knowledge Exploration and Research Platform

The future digital library will not only be a space for information storage, but also a platform to support academic research and knowledge creation. By integrating big language models with natural language processing technology, the library will provide scholars with powerful research assistance tools, such as automatically generating literature reviews, intelligent recommendation of relevant research paths, and so on. This not only improves research efficiency, but may also trigger new academic discoveries and innovations.

5. CONCLUSION

The big language model is rapidly changing the face of digital libraries, and its powerful natural language processing and text generation capabilities have brought new optimization paths for text mining and information retrieval. Although there are still some technical and application challenges, with the continuous evolution of models and the upgrading of digital library functions, the combination of the two will inevitably promote libraries to move from traditional knowledge storage to intelligent knowledge sharing and generation platforms. In the future, we can look forward to a more intelligent, convenient, and personalized digital library ecosystem that will facilitate seamless knowledge transfer and innovative development.

REFERENCES

[1] Lu, K., Sui, Q., Chen, X., & Wang, Z. (2025). NeuroDiff3D: a 3D generation method optimizing viewpoint consistency through diffusion modeling. Scientific Reports, 15(1), 41084.

[2] Zhang, T. (2025). From Black Box to Actionable Insights: An Adaptive Explainable AI Framework for Proactive Tax Risk Mitigation in Small and Medium Enterprises.

[3] Xie, J., Zhang, L., Cheng, L., Yao, J., Qian, P., Zhu, B., & Liu, G. (2025). MARNet: Multi-scale adaptive relational network for robust point cloud completion via cross-modal fusion. Information Fusion, 103505.

[4] Q. Tian, D. Zou, Y. Han and X. Li, "A Business Intelligence Innovative Approach to Ad Recall: Cross-Attention Multi-Task Learning for Digital Advertising," 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Shenzhen, China, 2025, pp. 1249-1253, doi: 10.1109/AINIT65432.2025.11035473.

[5] Zhang, Yuhan. "AdOptimizer: A Self-Supervised Framework for Efficient Ad Delivery in Low-Resource Markets." (2025).

[6] Hu, Xiao. "Low-Cost 3D Authoring via Guided Diffusion in GUI-Driven Pipeline." (2025).

[7] Zhang, X. (2024). Research on Dynamic Adaptation of Supply and Demand of Power Emergency Materials based on Cohesive Hierarchical Clustering. Innovation & Technology Advances, 2(2), 59–75. https://doi.org/10.61187/ita.v2i2.135

[8] Li, X., Wang, X., & Lin, Y. (2025). Graph Neural Network Enhanced Sequential Recommendation Method for Cross-Platform Ad Campaign. arXiv preprint arXiv:2507.08959.

[9] Xu, Haoran. "CivicMorph: Generative Modeling for Public Space Form Development." (2025).

[10] Tu, Tongwei. "AutoNetTest: A Platform-Aware Framework for Intelligent 5G Network Test Automation and Issue Diagnosis." (2025).

[11] Xie, Minhui, and Boyan Liu. "InspectX: Optimizing Industrial Monitoring Systems via OpenCV and WebSocket for Real-Time Analysis." (2025).

[12] Zhu, Bingxin. "REACTOR: Reliability Engineering with Automated Causal Tracking and Observability Reasoning." (2025).