Journal on
Artificial Intelligence

Tech Science Press

Doi:10.32604/jai.2025.073792

ARTICLE

Topic Mining and Evolution Analysis of Domestic Smart Library Research Based on the BERTopic Model

Meile Li[1] **and Yinuo Jiang**[2,*]

[1]Faculty of Public Administration, Xiangtan University, Xiangtan, 411100, China
[2]School of Information Science and Engineering, Hunan University, Changsha, 410000, China

*Corresponding Author: Yinuo Jiang. Email: yinuo747153@163.com

Received: 25 September 2025; Accepted: 21 October 2025; Published: 28 November 2025

ABSTRACT: This paper conducts topic mining and analysis of research literature in the domestic smart library field based on the BERTopic model, aiming to reveal its topic development context and evolution trends. Journal literature in the smart library field collected by CNKI (China National Knowledge Infrastructure) from 2015 to 2024 was analyzed using the BERTopic model and dynamic topic modeling for topic mining and evolution trend analysis. The study found that the domestic smart library field involves multiple core topics, identifying a diversified topic structure centered around "data", "user", "5g", etc. The research results provide data support and practical reference for libraries to accurately identify key points of technology integration during smart transformation and to optimize smart service models.

KEYWORDS: Domestic smart library; BERTopic; topic mining; evolution analysis

1 Introduction

Libraries, on one hand, bear the fundamental responsibility of inheriting civilization and imparting knowledge, and on the other hand, serve as important engines driving modern social development. In recent years, with the rapid rise of high-tech such as the Internet of Things and artificial intelligence, traditional libraries have gradually embarked on the path of intelligent development [1]. This transformation is not merely a simple technical upgrade but involves changing service methods through big data, reorganizing resources to improve user experience, and thereby assuming a core role in the knowledge service domain of the entire information era. Therefore, sorting out the core issues in the smart library field helps the academic community fully grasp the development context and evolution laws of smart library research topics [2], and also provides empirical support and theoretical foundation for library workers to formulate technical strategies and improve service models.

Research topics in the domestic smart library field have become increasingly diverse in recent years [3]. Studies mainly employ bibliometric methods, while fewer use deep learning methods for topic mining and analysis. This makes existing analyses reveal topics with a coarse granularity, making it difficult to delve into the semantics and reveal fine-grained research topics. Therefore, this paper utilizes the deep learning-based BERTopic topic modeling to address the above issues. Compared to the LDA topic model: in terms of semantic representation, LDA relies on word frequency statistics, ignoring contextual information, whereas the BERTopic model is based on deep learning context, greatly improving the semantic accuracy of topic clustering; in terms of topic number specification, LDA requires researchers to specify the number of topics in advance, which is highly subjective, while BERTopic automatically identifies it from the data through

clustering algorithms, making it more objective; in terms of cross-lingual consistency, LDA heavily relies on a single language system, while BERTopic natively supports multiple languages; in terms of interpretability, LDA relies on topic high-frequency word lists, which are easily disturbed by common words, while BERTopic, based on the c-TF-IDF (class-based Term Frequency-Inverse Document Frequency) method, can accurately extract feature words for each topic.

Therefore, this study, based on the BERTopic model, conducts a deeper topic mining and evolution trend analysis of domestic smart library journal literature collected by CNKI from 2015 to 2024, aiming to more comprehensively elucidate the latest achievements in this field, thereby providing empirical support for improving service quality.

2 Research Data and Methods

2.1 Data Collection

This study utilized the CNKI platform for literature retrieval, ensuring the authority and completeness of the data source. The specific retrieval strategy was as follows: "Subject = Smart Library", with the time range set from 2015 to 2024, including various types of samples such as Chinese academic journal articles, dissertations, and conference proceedings to ensure diversity of data types. After retrieval according to the above scheme, 5336 relevant documents were obtained. They were systematically organized, including title, author, author affiliation, abstract, keywords, and publication year. This data provides a reliable source for subsequent topic modeling using the BERTopic model, ensuring that the research conclusions accurately reflect the development context and evolution characteristics of the domestic smart library field in the past decade.

2.2 Research Methods

This study employed the BERTopic model to analyze the topic mining and evolution paths in the domestic smart library field. BERTopic is a topic modeling method based on deep learning [4]. By integrating pre-trained language models with traditional clustering techniques, it captures contextual information of words and documents, enabling deep semantic mining of text data. The core advantage of this model lies in its multi-stage processing pipeline: first, BERT-based models transform text into high-dimensional vectors; then, the UMAP (Uniform Manifold Approximation and Projection) algorithm reduces dimensionality; next, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) performs density-based clustering; finally, the c-TF-IDF (class-based Term Frequency-Inverse Document Frequency) method extracts topic keywords. This pipeline design allows BERTopic to automatically determine the number of topics, effectively handle noisy data, and adapt to text collections of varying densities.

To verify the effectiveness of the BERTopic model in this study, a comparative experiment was set up beforehand. On the same preprocessed corpus, both BERTopic and the traditional LDA model were run. By comparing manual evaluation of topic keywords and quantitative indicators, it was found that topics generated by BERTopic (such as "reading promotion", "user profile") were more semantically distinct, while LDA topics were more susceptible to interference from high-frequency words [5]. This comparative analysis strongly supports the advantage of using the BERTopic model in this study. The specific operational steps are: first, preprocess the original text data, including cleaning, word segmentation, and removal of stop words; then use the paraphrase-multilingual-mpnet-base-v2 pre-trained language model, embedding size 768 dimensions, choose the UMAP dimensionality reduction method with UMAP (n_neighbors = 15), choose the HDBSCAN clustering method with HDBSCAN (min_cluster_size = 28); finally, determine the core of the smart library research topics based on topic keywords and representative literature, thereby completing

the task of automatic identification and precise extraction, providing support for further investigation, all the experimental data are recorded in the supplementary materials.

3 Topic Mining in Domestic Smart Library Research

3.1 Topic Identification

Topic modeling methods were used to analyze the paper abstracts in the dataset, ultimately refining 14 key topics. An interactive visualization map was created based on these topics, as shown in Fig. 1. In this map, circular nodes represent individual topics. The size of the node corresponds to its frequency within the entire document collection—the larger the node, the more frequently it appears, and *vice versa*. Based on the spatial distribution of nodes within the coordinate system, the strength of connection or similarity between various topics can be intuitively seen. Adjacent nodes generally have higher connectivity, while distant nodes indicate greater differences between them. In-depth analysis of the interactive map reveals that among the extracted topic categories, Topic0 (data), Topic1 (user) [6], Topic2 (5g), and Topic3 (public) are the four most frequent and important topics.

Figure 1: Intertopic distance map

After a systematic analysis of the literature database using topic modeling techniques, a theoretical framework comprising 14 representative core topics was refined and constructed. To make the characteristic

attributes of each topic and the distribution patterns of keywords more visually apparent, a bar chart is used to display the key feature words and quantitative statistics for several typical topics, as shown in Fig. 2. By observing Fig. 2, it can be seen that research in the smart library field is characterized by diversity and a cutting-edge nature. The research scope covers not only the smart library itself but also multiple areas closely related to smart libraries, such as data, users, artificial intelligence, and the metaverse. Furthermore, the emergence of topics such as reading promotion, smart space, smart city [7], the 5G era, public library, blockchain, user profile, and ancient texts also indicates that the research system of smart libraries is continuously developing in a deeper and more comprehensive direction. The topics and their feature words displayed in Fig. 2 not only reflect the breadth and depth of research in the smart library field but also reveal the development trends and frontier dynamics in different directions within this field. Research on these topics not only helps promote theoretical and practical innovation in smart libraries but also provides important references and insights for the future development of the library industry.

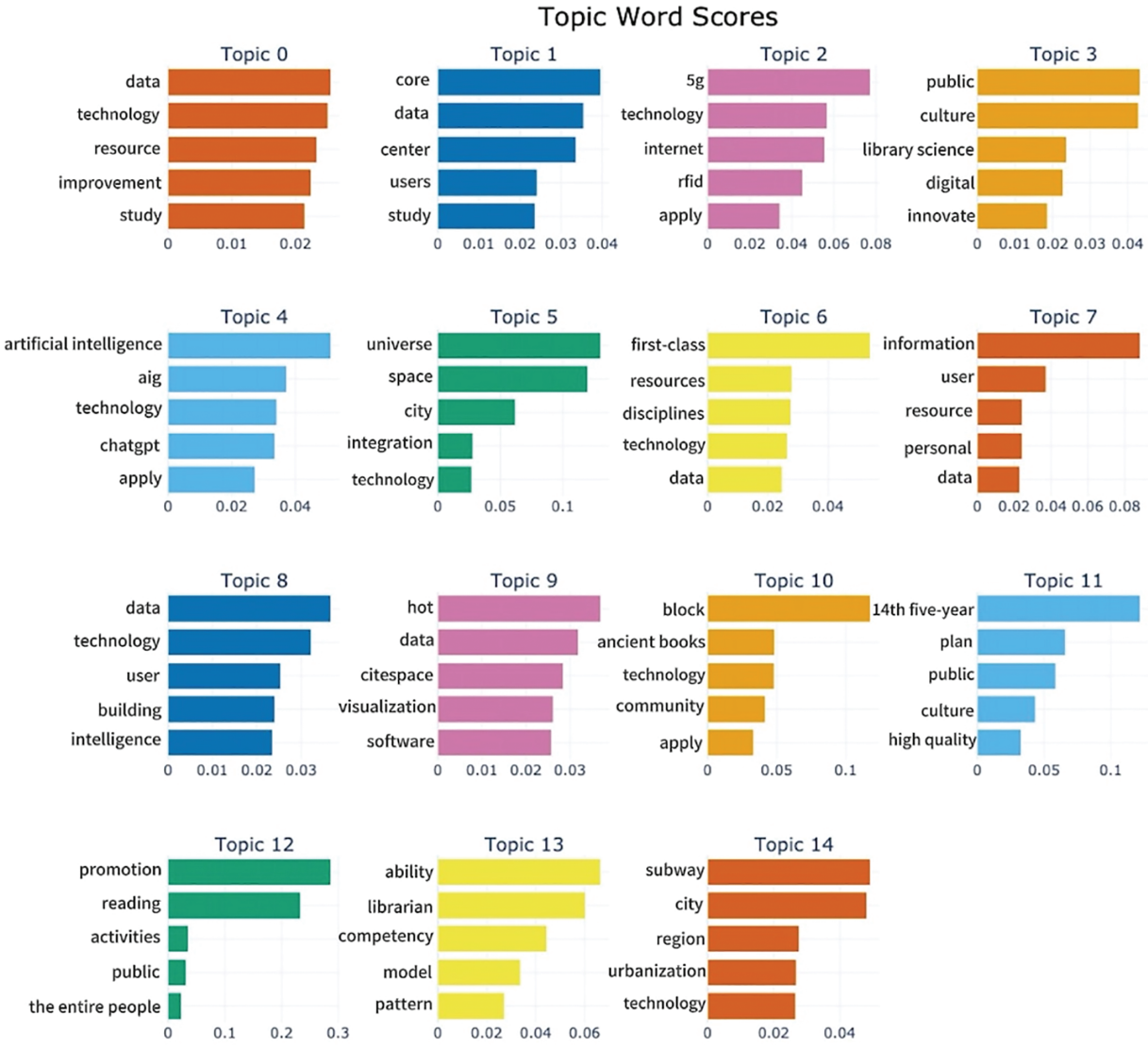

Figure 2: Topic word scores

3.2 Topic Consolidation

To gain an in-depth understanding of the logical relationships and connections within each research topic, this study utilized the topic hierarchy functionality of the BERTopic model to visually present the complex network relationships among various topics, as illustrated in Fig. 3. Based on research topics from the past decade, it has been found that studies in the field of smart libraries are not only extensive but also deeply focused, covering multiple hot topics such as artificial intelligence, smart cities, reading promotion, smart spaces, the 5G era, public libraries, blockchain, and user profiles. Simultaneously, Fig. 3 intuitively and systematically reveals that the current smart library research domain has formed three core modules that are relatively independent yet interconnected: first, the technology-driven module, which focuses on library intelligent upgrades with AI and big data at its core; second, the strategy and space module, which combines national strategies and cutting-edge technologies such as the metaverse and blockchain to plan the future form and development direction of libraries [8]; and finally, the public service and practice module, which concentrates on public services and talent cultivation to ensure that technological development and strategic planning ultimately translate into services for the public. This figure clearly outlines the overall knowledge map and internal logical structure of smart library research in China from a macro perspective.

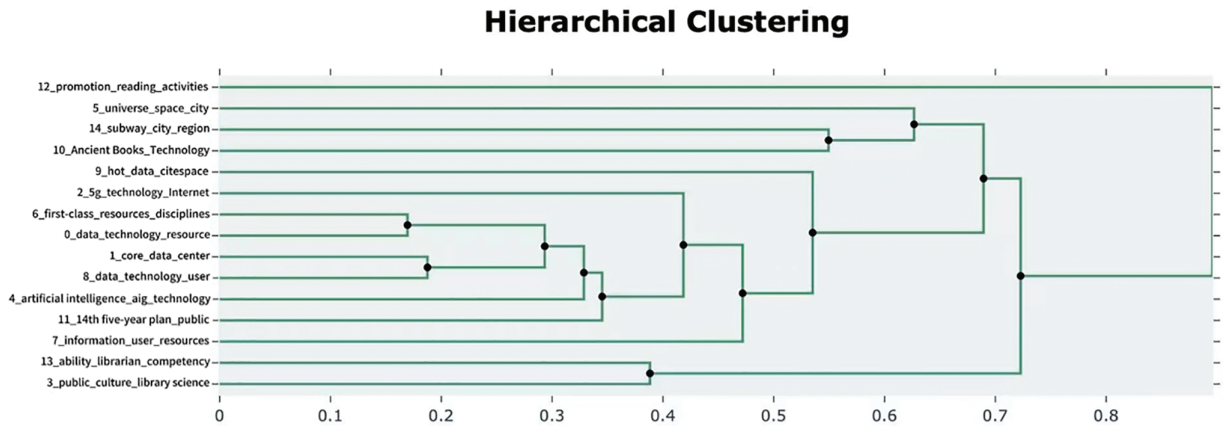

Figure 3: Hierarchical clustering

In summary, Fig. 3 not only provides a comprehensive overview of the research topics in the field of smart libraries but also, through hierarchical clustering and visual analysis, reveals the complex and profound intrinsic connections among the various topics.

4 Evolutionary Analysis of Domestic Smart Library Research Themes

The evolution of domestic smart library research over the past decade demonstrates clear patterns [9]. Fig. 4 shows the changes in research popularity of the 14 topics over time. From the figure, the overall trend of domestic smart library research topics transitioning from stability to explosion, and then to deepening, over the last ten years can be clearly observed. The period from 2014 to 2018 was a developmental phase, during which the frequency of most topics remained below 60, indicating that the overall scale of research during this time was relatively small and still in the exploratory stage. From 2019 to 2022, an explosive period occurred, where the frequency of research topics rose significantly, showing explosive growth especially after 2020 [10], which is related to the maturation of technologies such as big data and artificial intelligence. From 2023 to the present, a differentiation phase has emerged, where the growth momentum of most topics has slowed, and some topics have even shown a declining trend, such as

Topic 2 (5G) and Topic 5 (Metaverse). This indicates that researchers are shifting from technology introduction to more specific and in-depth technology integration, moving from single technology applications to the integrated development of technologies such as 5G, big data, blockchain, the Internet of Things, and artificial intelligence.

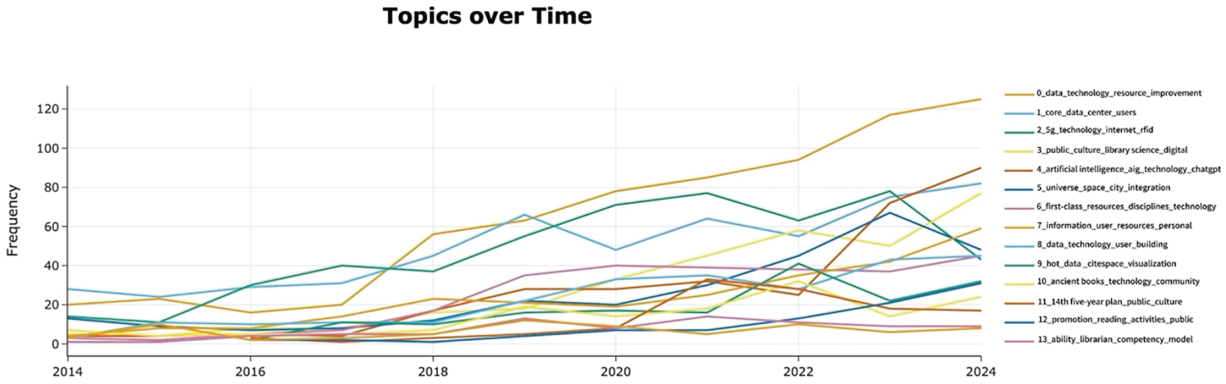

Figure 4: Topics over time

Overall, domestic smart library research is a dynamically evolving and active field that continuously incorporates cutting-edge technologies. The evolutionary trajectory over the past decade fully reflects the profound impact of technological innovation on the development of smart libraries.

5 Conclusion and Implications

Research on smart libraries demonstrates distinct interdisciplinary characteristics. Its main topics include smart services, smart librarians, and smart cities, among others. These topics are interconnected and influence each other, reflecting that the development of smart libraries is not a linear process in a single dimension but rather a dynamic development process involving multi-dimensional coordination and systematic integration. Comprehensive exploration and overall planning should be carried out from the perspectives of technological innovation, functional improvement, and management model renewal. With the deep integration of 5G communication, artificial intelligence algorithms [11], and blockchain technology, the construction of smart libraries is being infused with unprecedented momentum [12]. The interweaving and application of these technologies [13] have significantly enhanced the level of intelligence in libraries, leading to qualitative leaps in areas such as data processing, resource management, and security protection. This has given rise to service forms such as intelligent navigation, virtual reference services, and self-service borrowing terminals, greatly enriching readers' reading experiences and learning pathways. This fully demonstrates the core role of technology in promoting the entire smart library creation process and indicates that the future development trend of smart libraries should focus more on strengthening technological innovation and fostering close integration between technology and operations. The creation of smart libraries should be user-demand-oriented, with the fundamental goals of improving service quality and user experience. By leveraging advanced technological means such as big data analysis and artificial intelligence algorithms, personalized recommendations and precise retrieval can be achieved, better exploring and meeting the diverse information acquisition needs of readers [14]. Consequently, an intelligent, humanized service system centered on users can be established.

In the future, continuous attention should be paid to the application of advanced technologies such as 5G, artificial intelligence [15], and blockchain in the field of smart libraries, encouraging technological and model innovation. Through innovation, smart libraries can significantly enhance service efficiency and

provide users with higher-quality and more convenient means of accessing information. Smart libraries are key carriers of knowledge services, and their long-term development relies on high-quality professionals. Emphasis should be placed on conducting vocational skills training for practitioners, improving their professional standards and practical operational capabilities, thereby providing talent support for the industry's development. Governments and relevant institutions should formulate supportive policies, improve the corresponding institutional environment, and create favorable external conditions to promote the establishment of smart libraries and encourage their sustainable development. Meanwhile, this study only used the CNKI database to present the domestic research landscape. Future research will incorporate datasets from international databases, such as Web of Science, and utilize cross-lingual BERTopic models to enable a comparative analysis of research trends between China and other regions.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Meile Li, Yinuo Jiang; data collection: Meile Li; analysis and interpretation of results: Meile Li, Yinuo Jiang; draft manuscript preparation: Meile Li, Yinuo Jiang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All data generated or analysed during this study are included in this published article and its supplementary materials.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Supplementary Materials: The supplementary material is available online at https://www.techscience.com/doi/10.32604/jai.2025.073792/s1.

References

1. Gul S, Bano S. Smart libraries: an emerging and innovative technological habitat of 21st century. Electron Libr. 2019;37(5):764–83. doi:10.1108/el-02-2019-0052.
2. Pan YH. Important developments for the digital library: data ocean and smart library. J Zhejiang Univ Sci C. 2010;11(11):835–6. doi:10.1631/jzus.C1001000.
3. Yang SL, Yu YH. Topic mining and evolution analysis of information resource management research based on BERTopic model. Inf Sci. 2024;42(8):12–21. doi:10.13833/j.issn.1007-7634.2024.08.002.
4. Yang AR, Chae J, Choi E. Analysis of peatland research trends based on BERTopic. Land. 2024;13(5):628. doi:10.3390/land13050628.
5. Zhu Y, Liu Y. Gibbs-BERTopic: a hybrid approach for short text topic modeling. IEEE Access. 2025;13(18):49162–73. doi:10.1109/ACCESS.2025.3552221.
6. Peng L, Wei W, Gong Y, Jia R. University library space renovation based on the user learning experience in two Wuhan universities. Int J Environ Res Public Health. 2022;19(16):10395. doi:10.3390/ijerph191610395.
7. Zanella A, Bui N, Castellani A, Vangelista L, Zorzi M. Internet of Things for smart cities. IEEE Internet Things J. 2014;1(1):22–32. doi:10.1109/jiot.2014.2306328.
8. Buyannemekh B, Gasco-Hernandez M, Gil-Garcia JR. Fostering smart citizens: the role of public libraries in smart city development. Sustainability. 2024;16(5):1750. doi:10.3390/su16051750.
9. Wang C, Blei D, Heckerman D. Continuous time dynamic topic models. Commun ACM. 2012;55(10):95–103.
10. Temiz S, Salelkar LP. Innovation during crisis: exploring reaction of Swedish university libraries to COVID-19. Digit Libr Perspect. 2020;36(4):365–75. doi:10.1108/dlp-05-2020-0029.

11. Asemi A, Ko A, Nowkarizi M. Intelligent libraries: a review on expert systems, artificial intelligence, and robot. Libr Hi Tech. 2021;39(2):412–34. doi:10.1108/lht-02-2020-0038.
12. Raman R, Pattnaik D, Hughes L, Nedungadi P. Unveiling the dynamics of AI applications: a review of reviews using scientometrics and BERTopic modeling. J Innov Knowl. 2024;9(3):100517. doi:10.1016/j.jik.2024.100517.
13. Bi S, Wang C, Zhang J, Huang W, Wu B, Gong Y, et al. A survey on artificial intelligence aided Internet-of-things technologies in emerging smart libraries. Sensors. 2022;22(8):2991. doi:10.3390/s22082991.
14. Zhao L, Zhang S, Bai Z. Research on the digital literacy evaluation system for college students in smart libraries. J Imaging Sci Technol. 2025;69(1):1–14. doi:10.2352/j.imagingsci.technol.2025.69.1.010409.
15. Xu SW. Library service revolution: future development trends of smart libraries based on the large language model. J Libr Sci. 2024;46(7):1–5. doi:10.14037/j.cnki.tsgxk.2024.07.020.