DOI: 10.53469/jssh.2025.7(11).05

Research on Academic Information Retrieval in the Age of Big Data and AI

Huaili Zheng*, Ting Jiang

School of Computer and Artificial Intelligence, Nanjing University of Finance & Economics, Nanjing, China
*Correspondence Author, zhenghuaili@nufe.edu.cn

Abstract: ***In the era of big data and artificial intelligence, academic information retrieval has undergone significant transformations. The data sources, retrieval methods and tools, as well as the presentation of search results, now differ substantially from traditional retrieval models. This paper analyzes how big data and AI technologies optimize academic information retrieval by expanding data sources and search scopes, enabling personalized search, semantic search, and associative search. It then delves into optimization strategies for traditional retrieval platforms to adapt to new technologies, including the integration of AI technologies into search engines and literature databases to achieve one-stop retrieval, enhancing network infrastructure improvements, and the need for users to develop big data literacy and AI literacy in this new technological landscape.***

Keywords: Artificial Intelligence, Big Data, Academic Information Retrieval, Big data literacy.

1. Introduction

Against the backdrop of the Internet's rapid advancement, both the quantity and quality of online information services have significantly improved. Research users leverage scientific information exchange systems to access intelligence in their fields of interest, thereby enhancing research efficiency and driving disciplinary progress. In the era of big data, nations, enterprises, and individuals alike are prioritizing the enhancement of information literacy.

The advancement of big data and artificial intelligence has enabled the integration of information retrieval with big data and AIGC technologies. This convergence is not merely a simple tool overlay but has sparked a profound transformation in the very essence of "search." It has completely shattered the limitations of traditional search in terms of data volume, diversity of types, retrieval speed, and cognitive depth. The scope of search has expanded from the finite, structured world of text to a nearly infinite, multimodal, and real-time flowing data space. This expansion not only dramatically increases the breadth and efficiency of information access but, more importantly, endows information retrieval with "insight" and "predictive power." It evolves search from a mere information retrieval tool into a core engine for knowledge discovery and intelligent decision-making.

2. The Technological Context of Big Data

The emergence of big data technology stems from the explosive growth of data and the acceleration of the informatization process. Since the late 1990s, with the widespread adoption of the internet and the rapid advancement of informatization, organizations and institutions began collecting, storing, and processing vast amounts of data. This includes both structured data (such as corporate transaction records and customer data) and unstructured data (such as social media comments, images, and videos). Traditional data processing technologies could no longer meet the demands of handling such massive volumes of data, leading to the emergence of big data technology.

In the era of big data, data has become a crucial element in academic information resource retrieval. Data sources now encompass not only traditional scientific and technical literature but also social media data, knowledge platform data, government data, commercial data, and scientific and technological data.

Big data and artificial intelligence technologies have transformed academic information retrieval from a passive, tool-based "search box" into an active, intelligent "research partner." AI empowers academic users to interact with vast databases and knowledge repositories in more efficient and insightful ways, accelerating the scientific research process. For researchers, mastering and leveraging these AI-driven retrieval tools has become an indispensable core competency in the new technological era.

3. Academic Information Resource Retrieval in the Context of Big Data and AI

3.1 Expansion of Data Sources and Retrieval Scope

Traditional information resources and data sources are typically acquired through methods such as questionnaires, interviews, web crawlers, and API interfaces. Generative AI, however, can be integrated into search engines, knowledge bases, or various social media platforms while ensuring compliance with data acquisition regulations. Users submit requests to the AI based on their information needs, specifying parameters like data scope and field types. The AI then collects and filters data, returning information resources in the requested format. For information science, generative AI is not merely a tool, the content it generates also becomes a subject of study, effectively expanding the scope of data sources.

The deep integration of big data technology and information retrieval has significantly expanded the scope of traditional information retrieval, achieving a paradigm shift from "limited search" to "unlimited exploration." Big data and information retrieval complement each other: big data primarily handles massive amounts of heterogeneous data,

while traditional retrieval systems can only process structured data. However, in the big data context, social media images and real-time sensor stream data have all become searchable objects.

3.2 Personalization of Search Results

By analyzing users' search histories and behavioral patterns through big data, search results can be tailored for each individual, delivering more personalized and precise information.

AI systems construct comprehensive user profiles by collecting multi-source heterogeneous data (such as search records, click behavior, dwell time, and social interactions). Cross-platform data integration technology resolves information silos, while the fusion of big data and artificial intelligence delivers personalized search experiences for academic information retrieval. By analyzing user search behavior and preferences, search engines can generate customized results for each individual. This approach helps meet diverse information needs, enhancing user satisfaction and loyalty.

In the field of user behavior research, AI can effectively capture user intent and emotional tendencies, learn patterns emerging from data, and generate new content. As a future development in information science retrieval, the key consideration lies in how to deeply integrate intelligence thinking with AI. By leveraging the expertise of retrieval specialists and intelligence thinking, AIGC can better adapt to complex research tasks involving disciplinary information and knowledge retrieval. This enables intelligent information resource retrieval and knowledge generation, empowering knowledge production capabilities.

3.3 Implementing Semantic Search

By leveraging natural language processing (NLP) technology, search engines can understand and analyze user intent to deliver results that better align with user needs. Breakthroughs in NLP form a crucial foundation for the widespread adoption of AIGC. Semantics is the science of meaning. Core semantic technologies include knowledge extraction, retrieval, modeling, and reasoning. In practical applications, NLP helps retrieval systems better understand users' actual needs by accurately parsing the raw text they input, thereby enhancing human-computer interaction efficiency. The system recognizes that "computer vision" and "image recognition" are highly related concepts. Thus, even if a user enters only one term, it can return relevant literature containing the other concept.

3.4 Implementing Related Search

In academic information retrieval, users often require access to additional research information related to their current search terms to further expand and refine their search needs. By analyzing associative relationships within big data, intrinsic connections between different academic resources can be uncovered, thereby delivering more comprehensive search results. Particularly driven by artificial intelligence (AI) technology, the efficiency and accuracy of academic information search have seen significant improvements. AI models automate processes by extracting key research findings, evaluating study quality, and identifying relevant data points across vast research landscapes. This generates content of higher professional quality, reducing manual labor while enhancing the reproducibility and transparency of information resources. Furthermore, AI technology analyzes trends and patterns within big data to predict research trajectories and deliver real-time solutions, empowering researchers to achieve breakthrough discoveries.

In summary, the technological backdrop of big data provides robust technical support and limitless possibilities for academic information retrieval. The application of big data technology in this field further propels the advancement and development of academic research. As big data and AI technologies continue to evolve, future academic information retrieval will become increasingly intelligent, personalized, and efficient.

4. Optimization and Enhancement of Traditional Retrieval Systems in the Age of AI

4.1 Strengthening Improvements to Search Engines and Scientific Literature Databases

The primary platforms for traditional online academic information retrieval are public search engines and specialized scientific literature databases, making improvements to these tools particularly crucial. While public search engines and scientific literature databases share similarities in supporting advanced search and Boolean logic queries, they exhibit significant differences in information organization and presentation methods.

Public search engines deliver diverse, high-volume results, yet this abundance complicates accurate information retrieval. In contrast, scientific literature databases offer clearly categorized, highly relevant, and uniformly formatted results. However, they typically require subscription access and demand advanced user search skills.

Search engines globally are integrating artificial intelligence to deliver more diverse results. AIGC analyzes user intent and search context to generate more targeted content. ChatGPT and Google's Gemini excel at producing human-like responses, solving complex problems, and engaging in multi-turn conversations. Baidu and Sogou have also made significant strides in optimizing retrieval algorithms and enhancing accuracy through machine learning and semantic understanding technologies.

Therefore, when improving search tools, emphasis should be placed on optimizing the search interface and functionality. The search interface should be clean, elegant, and clearly organized, with key content and buttons positioned prominently on the screen and utilizing easily understandable graphical icons. Simultaneously, search functions should be refined and customized as much as possible to meet the diverse needs of different users. Additionally, providing full-text download links and supporting multiple file formats are also crucial measures for enhancing user experience.

Although generative AI has seen initial application in information retrieval, it remains a novel technology for most search engine users, who may be unfamiliar with its operations and conversational interaction patterns. To address this, it is recommended that intelligent search engines provide usage guidance and instructions for new users, such as interactive tutorials, smart prompts, and guide documentation — to help them quickly and effectively adapt to and fully leverage the various features of generative AI search engines.

4.2 Advancing One-Stop Searching of Domestic and International Scientific Literature Databases

Academic resources are widely distributed online. Beyond formally published academic information resources such as online journal databases, electronic journals, and e-books—which undergo systematic review by scientific literature systems—a vast amount of semi-formal information published by various academic organizations, universities, research institutions, and others on their official websites or information platforms has become a significant source of academic information. Simultaneously, the dissemination of cutting-edge research perspectives and ideas through academic forums, discussion groups, personal websites, public accounts, microblogs, and other channels or platforms has emerged as a new mode of academic information exchange for an increasing number of researchers across disciplines. This informal mode of information exchange is characterized by rapid dissemination, novel perspectives, and diverse formats. It offers researchers significant convenience in communicating with peers, sharing knowledge, and staying abreast of research trends, developments, and hot topics within their fields. Consequently, it plays an increasingly vital role in the scientific research process.

With the increasing abundance of academic resources both domestically and internationally, users often need to switch between multiple databases and platforms when conducting literature research. This not only impacts retrieval efficiency but also diminishes users' motivation to search. Science and technology information retrieval service platforms primarily leverage advancements in internet big data and artificial intelligence technologies. The emergence of data-intensive research paradigms has propelled the development of science and technology information systems into an interconnected big data era.

One-stop services have emerged as a key trend in AI search. AI can deliver services independently or integrate with diverse tools and platforms to provide users with more powerful, comprehensive solutions. This integrated approach helps users efficiently fulfill professional work and task requirements by reducing platform switching and simplifying complex search processes. It not only enhances retrieval efficiency but also significantly boosts user satisfaction.

Integrating domestic and international scientific and technological literature resources to achieve one-stop search functionality can significantly enhance the efficiency and quality of academic users' searches. Search capabilities in the era of artificial intelligence are becoming increasingly intelligent. When training AI algorithms, search systems must utilize diverse datasets to reduce negative biases. Simultaneously, the evaluation of algorithm development processes must be strengthened to minimize algorithmic bias, ensuring comprehensive and objective search results. This necessitates enhanced collaboration and resource sharing among database providers to collectively advance the optimized allocation and efficient utilization of academic information resources.

4.3 Fostering a Healthy Online Environment and Improving Network Infrastructure

While big data and artificial intelligence technologies have enhanced academic resource retrieval by offering users efficiency, speed, and convenience, they have also given rise to issues such as algorithmic bias and personal privacy protection. Stricter regulatory mechanisms and review systems must be established in the future. Robust legal frameworks and industry standards are equally crucial. To foster a better online environment and enhance the utilization of academic information resources, national security agencies should enact clear regulations to ensure cybersecurity and elevate the quality of online information resources. Administrative bodies should mandate regular maintenance by network administrators to remove junk software and eliminate viruses, thereby safeguarding personal privacy.

Additionally, improving network infrastructure is equally important. Relevant network management departments should implement reasonable regulations to reduce internet fees and enhance network connectivity and speed, providing users with a more convenient environment for retrieving academic information online. By fostering a favorable online environment and upgrading network facilities, the efficiency and satisfaction of online academic information retrieval can be effectively enhanced.

5. Enhancing User Competency in the Age of AI

5.1 Expanding Knowledge Reserves

During online academic information retrieval, users' knowledge frameworks and proficiency levels significantly influence their search behaviors and outcomes. This is because their knowledge reserves and practical application skills permeate the entire process of information literacy, cognition, and behavior. Research findings indicate that the influence of academic users' knowledge frameworks and proficiency levels remains consistent throughout the entire process, from the task-driven and information-needs stages to the information retrieval, screening, acquisition, and feedback stages. Therefore, whether they are research users or university faculty and students, they should actively focus on knowledge accumulation, enhance their English proficiency, and improve their mastery of utilizing high-quality foreign-language databases. Simultaneously, they should adeptly leverage online platforms for research innovation, familiarize themselves with common online academic resources to broaden their exposure, and continuously enhance theoretical literacy. Through extensive practice, exploration, engagement with novel concepts, and enrichment of experiences, they can broaden their horizons, thereby enriching knowledge reserves and optimizing knowledge

structures.

5.2 Enhancing Digital Literacy and AI Literacy

The concept of digital literacy was first proposed in 1994, defined by Israeli scholars as essential survival skills for the digital age. With the advent of the big data era, increasing attention has been directed toward digital literacy. In 2018, UNESCO released the Global Digital Literacy Framework. In 2021, China's Central Cyberspace Affairs Commission issued the Action Plan for Enhancing Digital Literacy and Skills Among All Citizens.

AI literacy builds upon digital literacy, encompassing the understanding and application of artificial intelligence technologies, interaction with AI systems, and the selection and utilization of AI-generated search results.

With the rapid advancement of AI technology, artificial intelligence education will gradually become widespread. Users must master the fundamental principles and applications of AI, such as machine learning and deep learning, to cultivate comprehensive competencies. Integrating AI with specialized disciplines enables professional information retrieval and analysis, social network analysis, and other applications.

5.3 Enhancing Users' Information Retrieval Skills

5.3.1Strengthening Retrieval Course Education to Improve Retrieval Skills

While universities have widely introduced courses such as "Information Retrieval" and "Online Information Resource Retrieval and Utilization," actual feedback indicates that course effectiveness has fallen short of expectations. Graduate students generally demonstrate superior retrieval skills compared to undergraduates, and there exists a significant gap between the retrieval skills of lower-year undergraduates and those of upper-year undergraduates and graduate students. Students must master traditional retrieval approaches, methods, and strategies as the foundation for skill development. In the era of big data and AI, institutions should prioritize cultivating students' information literacy, data literacy, and AI literacy within retrieval instruction. This requires strengthening information literacy education, aligning practical retrieval training with academic disciplines and grade levels, and ensuring faculty continuously learn new technologies—particularly the integration of big data, AI, and information retrieval—to enhance their own information literacy. University students should recognize that information retrieval skills are integral components of information literacy, big data literacy, and artificial intelligence literacy. They must understand the critical importance of these skills for future learning, work, and scientific research. Students should learn to design, implement, and optimize retrieval strategies. At the initial stage of a retrieval task, they need to clearly define and fully understand the requirements to facilitate the implementation of retrieval techniques. After clarifying fundamental concepts, they should conduct in-depth analysis of the subject matter's essence, utilize specialized terminology, and understand synonym alternatives. When constructing search queries, the users learn search operator knowledge and apply Boolean logic with advanced search techniques. develop the ability to select high-quality search results and optimize search strategies based on outcomes.

5.3.2 Actively Participate in Search Practices to Gain Experience

"Knowledge gained from books alone is always superficial." While theoretical knowledge can be acquired through study and accumulation, transforming it into practical skills requires continuous hands-on experience. In retrieval practice, some users face challenges in selecting appropriate platforms for different types of online academic information and may lack proficiency in using specialized retrieval platforms. These factors can significantly impact retrieval effectiveness. Therefore, academic users should actively engage in information retrieval activities. Institutions should also regularly organize relevant events to provide students with learning and practice opportunities. Under traditional retrieval models, users typically refine search strategies through methods like broadening or narrowing queries to enhance results; In contrast, AI-driven modes enable users to express search queries using natural language. Intelligent search systems, grounded in pre-trained large language models, generate natural language responses and deliver search results directly based on an understanding of user intent. These systems achieve more accurate comprehension of document content beyond mere surface-level keyword matching, evolving from "retrieval" to 'understanding' and "discovery." This places higher demands on users' information evaluation skills, requiring them to assess the credibility, applicability, and quality of search results based on their own information literacy.

6. Conclusion

In the field of information retrieval, AI enables more effective natural language search and facilitates the transition from "fuzzy search" to "precise delivery" through multilingual interaction and continuous dialogue. This renders cross-language retrieval under traditional information retrieval models, along with research on retrieval strategies and safeguards aimed at improving recall and precision rates, less critical. However, artificial intelligence cannot absolutely guarantee the authenticity and accuracy of generated content, and it is challenging to trace the origins of retrieval results. Therefore, integrating generative AI into traditional intelligence or knowledge retrieval systems to achieve more efficient, accurate, and contextualized searches, along with intelligence traceability, particularly for erroneous and false intelligence remains a key future research direction. Furthermore, while traditional retrieval provides lists of search results, AI delivers direct answers, shifting the focus from relevance issues to the more critical concerns of intelligence reliability and credibility.

In the era of big data, technological information resources have diversified, and retrieval methods and tools have become increasingly varied. Users increasingly emphasize the fulfillment of personalized needs. Leveraging its technological strengths, AIGC innovates service models by prioritizing a thorough understanding of user requirements,

thereby enhancing user reliance on retrieval platforms and technologies. AIGC and big data technologies drive high-quality development in modern retrieval techniques, provide robust information support for scientific innovation and cultural advancement, and strengthen national technological capabilities.

For academic information resource retrieval, big data and artificial intelligence technologies present both opportunities and challenges. Traditional retrieval techniques must actively embrace AI's strengths, exploring new frontiers in information science research and practice that integrate AIGC. yet must maintain rigorous diligence to maintain a clear and comprehensive understanding of both the benefits and potential risks of AI applications. Despite facing technical and societal challenges such as data governance, privacy security, and algorithmic fairness, the transformative potential of this expansion is immense and irreversible, profoundly reshaping how we perceive the world and solve problems. Search technology can leverage AI to explore and research more complex, deep-level information science problems, driving sustainable development in specialized fields. Future search will evolve into a "cognitive partner" that is smarter, more contextual, better integrated with multimodal data, and capable of proactively providing insights.

Acknowledgments

This work was supported by Department of Education Funded Project for Philosophy and Social Sciences in Colleges and Universities of Jiangsu Provincial, Grant 2020SJA0288.

References

[1] Dong Huanqing, Cao Gaohui, Tuo Pei. Formation path analysis of user information search behavior in generative AI tools—From the perspective of digital literacy and artificial intelligence literacy. https://link.cnki.net/urlid/44.1306.G2.20241129.0846.002

[2] Hu K. ChatGPT Sets Record for Fastest-growing User Base - Analyst Note [EB/OL]. [2023-02-09]. https://www.reuters.com/technology/chatgptsets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[3] Kay G. Bill Gates Calls ChatGPT 'Every bit as Important as the PC' or the Internet [EB/OL]. [2022-02-23]. https://www.businessinsider.com/bill-gates-chatgpt-ai-artificial-ntelligenct-as-important-pc-internet-2023-2.

[4] Inthiran A, Alhashmi S M, Ahmed P K. A user study on the information search behaviour of medical students [J]. Malaysian Journal of Library and Information Science, 2015, 20(1).

[5] WANG S, SCELLS H, KOOPMAN B, et al. Can ChatGPT write a good boolean query for systematic review literature search?[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.2023:1426-1436.

[6] ESHET Y. Digital literacy: A conceptual framework for survival skills in the digital era[J]. Journal of educational multimedia and hypermedia, 2004, 13(1): 93-106.

[7] Wang Ruojia, Fan Keming, Liu Zhifeng, et al. A Study on User Querying Behavior in Generative Artificial Intelligence Environment[J]. Data Analysis and Knowledge Discovery, 2024, 8(8/9): 20-30.

[8] Al Shboul M K I, Alwreikat A, Alotaibi F A. Investigating the Use of ChatGPT as a Novel Method for Seeking Health Information: A Qualitative Approach [J]. Science & Technology Libraries, 2023: 1-10.

[9] Ahmed I, Roy A, Kajol M, et al. ChatGPT vs. Bard: A comparative study[J]. Engineering Reports, 2023(1): 1-18.

[10] Siegle D. A role for ChatGPT and AI in gifted education[J]. Gifted Child Today, 2023, 46(3): 211-219.

[11] Schultz C D, Koch C, Olbrich R. Dark sides of artificial intelligence: The dangers of automated decision-making in search engine advertising[J]. Journal of the Association for Information Science and Technology, 2024, 75(5): 550-566.