A Bibliometric Analysis of Data Literacy Research Hotspots in China

Authors: Wei Jiansong

Date: 2023-07-28T00:00:00+00:00

Abstract

[Purpose/Significance] Data literacy has become an essential basic competency for individuals living in the digital age. [Method/Process] Reviewing published literature on data literacy helps clarify the research trajectory of data literacy-related studies and further optimize subsequent research. By collecting, organizing, and cleaning academic papers published in SCI, EI, CSSCI, CSCD, and Peking University Core source journals indexed in CNKI, this study conducts bibliometric analysis to systematically map the research trajectory of data literacy. [Results/Conclusion] Domestic scholars have conducted preliminary explorations in research on the connotation and value of data literacy, data literacy competency assessment, and data literacy cultivation pathways. Through CiteSpace analysis, it is found that "data-driven" is the latest burst term in data literacy-related research, which may represent the current research frontier in the field of data literacy.

Full Text

Analysis of Research Hotspots in Domestic Data Literacy Based on Bibliometrics

Nanchang Normal University, Nanchang 330000

Abstract

Data literacy has become an essential competency for individuals living in the digital age. Reviewing published literature on data literacy helps clarify the research trajectory and optimize future studies. This study conducts a bibliometric analysis of academic papers from SCI, EI, CSSCI, CSCD, and Peking University core journals indexed in CNKI, systematically 梳理 ing the research 脉络 of data literacy. Current domestic scholars have conducted preliminary explorations into the connotation and value of data literacy, capability evaluation,

and cultivation pathways. CiteSpace analysis reveals that “data-driven” is the latest emergent term in data literacy research, likely representing the current research frontier.

Keywords: data literacy; knowledge graph; bibliometrics

For a long time, decision-making processes based primarily on intuition and personal experience have been criticized for their arbitrariness and lack of objective foundation. With the continuous construction of digital infrastructure across industries, the difficulty of storing and transmitting various information carriers such as text and images in digital encoding formats has gradually decreased. The vigorous development of big data information technology and breakthroughs in computing and analytical technologies have enabled the aggregation, analysis, and application of big data containing fragmented information from socioeconomic, real-world, and management decision-making contexts. By interpreting this series of fragmented information, the productive 要素 and productive force functions of data have been strengthened, profoundly changing traditional human thinking modes and production-lifestyles. Data-driven development has gradually become an important trend in current social informatization development.

Accompanied by the deep transformation of various fields driven by new-generation information technology, the informatization environment across industries is undergoing unprecedented changes. The ability to acquire, store, and develop various types of data has become a fundamental capability for industry practitioners. Data literacy, as a comprehensive ability to reasonably acquire, understand, apply, evaluate, and manage data in a big data environment, represents an extension of concepts such as information literacy and statistical literacy. It is an essential competency for individuals to innovatively apply various types of data to carry out social production and life practices in an informatized environment.

This paper employs bibliometric research methods to conduct statistical analysis, topic mining, and comparative literature analysis of academic literature related to data literacy published in important domestic academic journals, aiming to understand the development trends of data literacy research in China and provide decision-making references for domestic scholars to capture research dynamics and topics.

This work is supported by the Jiangxi Provincial Higher Education Teaching Reform Research Project “Research on Teaching Quality Evaluation Reform Based on Value-Added Assessment” (Project No.: JXJG-22-23-30) and the Nanchang Normal University Education Assessment Project “Research on Evaluation and Improvement Pathways of Teachers’ Data Literacy Competence” (Project No.: 21XJZX30).

Author Information: Wei Jiansong, Deputy Director of Quality Management Section, Academic Affairs Office; Lecturer; Master’s Degree; E-mail: jason@ncnu.edu.cn

1 Research Methods and Data Sources

In recent years, the explosive growth of literature and continuous optimization of bibliometric research tools have attracted increasing numbers of scholars to the field of bibliometric research. Bibliometrics is a method that combines mathematical and statistical approaches to collect, organize, and analyze existing literature data for evaluation and prediction, enabling comprehensive identification from multiple dimensions regarding literature topics, scholars, institutions, and journals.

1.1 Research Methods

In the application of bibliometric methods, most scholars focus on the following aspects: 1) Using core journals in specific disciplines as data sources and applying bibliometric methods to 梳理 research 脉络, identify current hotspots, and predict research directions. For example, Ding Xiulian and colleagues used bibliometric methods to 梳理 research hotspots in the field of management science and engineering over the past decade based on papers published in 46 international authoritative journals in the field [1]. 2) Conducting academic journal evaluation research using bibliometric methods. Yu Liping and colleagues proposed an improved composite bibliometric indicator for evaluating academic journals by using the inverse of the low-cited paper ratio instead of citation concentration based on the z-index principle [2]. Additionally, some scholars have used bibliometric methods to study contributions from scholars in different regions to specific fields.

1.2 Data Sources

To enhance the quality of literature data for data literacy research and avoid interference from overlapping domestic and international databases, literature sources were limited to academic papers indexed in CNKI's SCI, EI, CSSCI, CSCD, Peking University Core, and AMI categories. Based on the characteristics of common literature retrieval methods such as "topic," "title," "keywords," "abstract," and "full text," and balancing recall and precision, this study selected "topic" for retrieval. Using CNKI's advanced search function, with the time range set to 2022 and earlier, the topic term set to "data literacy," and source categories limited to SCI, EI, CSSCI, CSCD, and Peking University Core, 565 literature records were initially obtained.

To improve research result accuracy, data cleaning was performed before analysis to further enhance precision by removing reviews, conference reports, achievement introductions, prefaces, entries without authors, and other irrelevant items, resulting in 502 valid papers.

2.1 Annual Publication Growth Trend

Typically, the growth of scientific knowledge is closely related to changes in publication volume on a given topic. When measuring research output and hotspots, scholars often use publication volume as an important indicator, which to some extent reflects the research 热度 and maturity of a topic.

By 梳理 and comparing the publication data, the first literature on the theme of "data literacy" appeared in 2010. As shown in Figure 1 [Figure 1: see original paper], which displays the annual trend of "data literacy" literature publication volume, the number of publications began to rise around 2015, but the growth trend remained relatively flat between 2016 and 2019.

2.2 Journal Distribution Analysis

According to information management theory, analyzing information distribution helps understand the concentration and dispersion characteristics and patterns of information. Hou Jianhua (2015) and other scholars found that the Matthew effect also exists in publishing journals and research directions through their study of Chinese journal evaluation [4]. After 统计 the distribution of published literature across journals, 130 journals were found to have published data literacy research papers. The top 10 journals by publication volume published a total of 206 papers, accounting for 41.4%; journals with more than 10 publications published a total of 251 papers, accounting for 50%. Specific source journals are shown in Table 1 . Further analysis reveals that in data literacy research, approximately 10% of journals published over 50% of all papers, demonstrating an enrichment phenomenon of data literacy research papers in a minority of journals. These top journals are all in the fields of library and information science and education research.

Table 1: Source Journals with More Than 10 Publications on Data Literacy Research [Note: Impact factor data in the table are from CNKI 2022 comprehensive impact factor.]

2.3 Author Collaboration Network Analysis

The number of publications by the same scholar on the same topic often represents deeper research in that field and demonstrates the scholar's potential leadership role, with their academic achievements possibly serving as important channels for ideas and perspectives in the research field. When using CiteSpace to analyze author collaboration networks, g-index was selected for data screening, yielding 269 nodes and 139 connections without network simplification, indicating that research on data literacy remains relatively fragmented. When the threshold was set to 3 for visual analysis of the author collaboration network, Figure 2 [Figure 2: see original paper] shows a relatively sparse network. From the publication statistics, only scholars such as Yang Xianmin, Huang Ruhua, Hu Hui, and Deng Lijun have published extensively on "data literacy."

Yang Xianmin focuses on data literacy in education and teaching, primarily on teacher data literacy, while Huang Ruhua, Hu Hui, and Deng Lijun focus on data literacy in library science.

2.4 Research Institution and Collaboration Network Analysis

Using CiteSpace software again with g-index selected for data screening and the threshold set to 5, the research institution collaboration network is shown in Figure 3 [Figure 3: see original paper]. The figure reveals that collaboration among domestic research institutions in this field is not strong, though some cooperation exists. The top five institutions by publication volume are Wuhan University School of Information Management, National Science Library of the Chinese Academy of Sciences, University of Chinese Academy of Sciences, Sun Yat-sen University School of Information Management, and Sichuan International Studies University Library.

3 Knowledge Graph Based on Keywords

Analyzing the keyword co-occurrence knowledge graph of data literacy research articles can reveal knowledge clustering and evolution patterns in the field, as well as co-citation trajectories and evolution networks, thereby helping to extract current research hotspots and clarify evolution trends.

3.1 Research Hotspot Analysis Based on Keyword Co-occurrence Network

Literature keywords provide a high-level summary of research content. In CiteSpace, node size indicates keyword frequency, with high-frequency terms representing current research hotspots. Connections between nodes indicate that the represented keywords appear in the same literature. Without network simplification, CiteSpace calculations yielded 323 nodes and 575 connections. With the threshold set to 24 and cluster analysis performed on keywords, Figure 4 [Figure 4: see original paper] shows the keyword co-occurrence and clustering map after adjusting the display effect.

The top ten keywords by frequency are data literacy, big data, information literacy, library, scientific data, artificial intelligence, data-driven, data journalism, data librarian, and data service. Keyword frequencies and centrality values are shown in Table 2 .

Table 2: High-Frequency Keywords in Data Literacy Research (2010-2022) [Note: The table would show serial number, frequency, and centrality values.]

3.2 Keyword-based Cluster Analysis

After cluster analysis of keywords, the clustering module value Q is 0.5819, and the clustering average silhouette value S is 0.907, both greater than 0.3 and 0.7 respectively, indicating that the clustering structure is reasonable and effective. In CiteSpace, clusters with fewer than 10 documents are not displayed by default, so Figure 4 shows only 10 cluster labels: #0 Big Data Research; #1 Data Literacy Research; #2 New Liberal Arts Research; #3 Influencing Factors Research; #4 Information Literacy Research; #5 Library Research; #6 Data Journalism Research; #7 Indicator System Research; #10 Self-Regulated Learning Theory Research; #11 Connotation and Characteristics Research. Converting Figure 4 to Timeline view yields Figure 5 [Figure 5: see original paper].

Figure 5: Timeline View of Keyword Co-occurrence in Data Literacy Research (2010-2022)

Based on keyword clustering and further analysis of published data literacy research papers, research hotspots can be refined as follows:

(1) Research on the Connotation and Value of Data Literacy

Regarding this theme, academic circles have mainly focused on teaching and research personnel and librarians. Data literacy is a core competency for librarians and researchers and a key concept in data management service research [5]. For teachers, those with data literacy can better adapt to data-intensive scientific research models, which also promotes research output [6] and facilitates the transformation of teacher professional development from "rough experience" to "evidence-based" paradigms [7]. For students, the dimensions of data literacy vary across different educational stages [8-9], but being adept at using big data to obtain needed information can help enhance professional knowledge reserves and better prepare them for society [9].

(2) Research on Data Literacy Capability Evaluation

As research on the connotation of data literacy deepens, evaluations of data literacy capabilities have gradually increased, with research subjects concentrated among university teachers and students. Data literacy represents a new societal requirement for individual competencies in the big data era, inheriting and developing traditional information literacy [10]. Different groups among university teachers, doctoral students, master's students, and undergraduates show significant differences in data literacy capabilities [11]. In evaluation indicator system research, scholars have conducted a series of studies from different dimensions. For example, Li Qing and Zhao Huanhuan (2018) used multiple research methods to 归纳 and analyze 筛选 ed literature on data literacy components, constructing a teacher data literacy evaluation indicator system from four aspects: data knowledge, data skills, teaching application, and awareness/ethics [12]. Ma Teng and Sun Ling (2019) built an evaluation indicator system from dimensions such as information, information environment, and information technology based on information ecology theory to evaluate student data literacy

[13]. For primary and secondary school students, capability models can be constructed from dimensions such as data knowledge and skills, data thinking, data awareness, and data ethics [14]. Combining high school mathematics teaching characteristics, some scholars have designed an evaluation framework for data analysis literacy from four dimensions [15]. Reviewing published research on data literacy evaluation reveals that most existing capability evaluation studies include dimensions such as data awareness, data processing capability, data communication, and data evaluation in their indicator systems.

(3) Research on Data Literacy Cultivation Pathways

To generate value from data across different fields, it is necessary to cultivate data literacy capabilities among practitioners in those fields. Compared with Harvard University's educational practices, data management services in Chinese universities are in a disordered state. In the big data era, universities should first formulate data literacy education and management policies [16]. For teachers (including normal school students), data literacy cultivation involves broad knowledge areas and is a systematic project requiring joint efforts from multiple parties [17]. For primary and secondary school teachers, a teaching model for cultivating data literacy based on self-regulated learning theory and regulatory scaffolding can be constructed [18]. For researchers, a data literacy cultivation framework based on data lifecycle and research project lifecycle theories can help researchers overcome capability deficiencies [19].

3.3 Frontier Analysis Based on Burst Detection Method

Burst detection analysis can identify key nodes in a research field to find current active or frontier topics. Since reference data are not included in CNKI-exported data, only keyword burst analysis was performed. With default parameters unchanged, calculations yielded two burst keywords: "data journalism" (strength 3.3) and "data-driven" (strength 3.71). The burst period for "data journalism" was 2014-2017, while for "data-driven" it was 2020-2022. The source literature for this keyword during the burst period is shown in Table 3 .

Table 3: Source Literature for the Keyword "Data-driven" [Note: The table lists authors, years, journals, and titles of papers.]

Based on the burst term "data-driven," analysis shows that source journals still mainly focus on education teaching and library/information science research. Since the burst period is 2020-2022, "data-driven" can be considered the current frontier in data literacy research.

3.3.1 "Data-driven" Research in Library and Information Science

Since the 21st century, big data has continuously been a research hotspot in academia. Across different disciplines, research dimensions on big data are gradually 细分 ing. In data-driven library and information science research, "data" serves as the development foundation, powerfully driving the development of library and information science [20]. However, implementing data-

driven approaches as the future development direction of libraries still requires macro-level efforts. He Yali (2020) and colleagues, after investigating 15 foreign libraries including the U.S. National Library of Medicine and the British Library, argued that in facing data-driven transformation, data should become the support point for library decision-making management, the 发力 point for business reconstruction, and the growth point for user services [21]. At the micro level, the rise of new research paradigms has prompted the emergence of a new generation of subject librarians characterized by data-driven approaches, supported by big data technology applications, and based on multi-source data services. Their functions are mainly reflected in knowledge services supporting scientific and technological decision-making, research management, and research processes [22].

3.3.2 "Data-driven" Research in Education and Teaching In the education and teaching field, the governance transformation brought by data-driven approaches plays an important role in accelerating the modernization of school governance [23]. Compared with other traditional industries, China's higher education field lacks innovation, and its education teaching and talent cultivation models have the drawback of batch production, leading to serious homogenization in talent cultivation. Therefore, universities should rely on data, follow value orientations such as learning personalization, management refinement, and teaching informatization, and explore effective methods to use big data to achieve transformation in university education teaching methods and forms [24]. By constructing new relationships among schools, government, and society, comprehensive and deep integration between education governance business and data can be promoted, enhancing education governance levels by leveraging the enthusiasm of various stakeholders [25]. In recent years, the term "precision" has been used with increasing frequency and broader application scope. In the field of sports training, "precision training" and "data-driven" have become high-frequency terms and main themes. The advantages of data-driven precision training enable athletes to obtain optimal stimulation based on individual differences, thereby producing optimal adaptation and improving training quality [26]. However, if teachers' data literacy is insufficient, data-driven empowerment of teacher leadership faces various problems such as insufficiently convenient and refined individualized teaching [27]. To achieve effective operation of "data-driven" teaching, teachers need to clearly recognize that the value embodiment of teaching data should always center on students, applying cold data to warm teaching practices [28].

4 Conclusion

Through multi-angle analysis of research literature related to data literacy published in SCI, EI, CSSCI, CSCD, and Peking University Core journals indexed in CNKI, the following conclusions are drawn:

First, data literacy research has become one of the continuously growing hotspot

topics in important journals of library science, information science, and education. Research on big data, information literacy, and scientific data applications already occupies an important position in library science, information science, and education journals, with related research advancing from data management toward data-driven approaches.

Second, from literature explicitly mentioning data literacy, early research mainly explored the connotation and value of data literacy in the big data era, which promoted subsequent research development. A large number of studies on data literacy components, evaluation, and education cultivation have become increasingly rich and diverse.

Third, among published research literature, most studies focus on student, teacher, and researcher groups. In the big data era, the digital economy has become a new driving force for economic growth. For people living in the digital age, data literacy will be one of the essential competencies for promoting human development in the 21st century.

This study inevitably has limitations. First, regarding literature data sources, although the selected academic research papers from SCI, EI, CSSCI, CSCD, and Peking University Core journals indexed in CNKI are representative, the overall coverage is insufficient. Second, foreign data literacy research was not included in the analysis, lacking examination of international research hotspots and frontier topics.

[1] Ding Xiulian, Wu Qiang, Zhang Peng, et al. Analysis of the Current Status of Chinese Management Science and Engineering Discipline Based on International Authoritative Journals in Management [J]. Chinese Journal of Management, 2022, 19(2): 159-168.

[2] Yu Liping, Wang Zuogong. Applicability of z-index in Evaluating Academic Journals and Its Improvement Research [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(11): 1132-1139.

[3] He Jiaxun, Ge Jiaye, Zhang Fan. Chinese Scholars' World Contributions to Management Research: International Cooperation, Frontier Hotspots, and Contribution Paths—A Quantitative Analysis Based on Papers from 1,000 World Management English Journals (2013-2019) [J]. Management World, 2021, 37(9): 36-67.

[4] Hou Jianhua, Liu Bo. The Matthew Effect in Journal Evaluation Research Output [J]. Chinese Journal of Scientific and Technical Periodicals, 2015, 26(9):

[5] Meng Xiangbao, Chang E, Ye Lan. Data Literacy Research: Origin, Status, and Prospects [J]. Journal of Library Science in China, 2016, 42(2): 109-126.

[6] Zhang Jinliang, Li Baozhen. Connotation, Value, and Development Path of Teacher Data Literacy in the Big Data Era [J]. e-Education Research, 2015, 36(7): 14-19, 34.

[7] Zhao Hongyuan. Teacher Professional Development Based on Data Literacy: Connotation and Path [J]. Continuing Education Research, 2017, (10): 77-80.

[8] Jia Pu, Song Naiqing. Data Literacy of Middle School Students in the Big Data Era: Connotation, Value, and Constituent Dimensions [J]. e-Education Research, 2020, 41(12): 28-34, 58.

[9] Zhang Minghai, Zhou Yanhong. Target Positioning and System Construction of College Student Data Literacy Education in the Big Data Era [J]. Library, 2016, (10): 84-88.

[10] Deng Lijun, Yang Wenjian. Research Progress on Individual Data Literacy Evaluation Systems and Related Indicator Connotations [J]. Library and Information Service, 2017, 61(3): 140-147.

[11] Long Qian. Construction of Data Literacy Capability Indicator System and Investigation and Analysis of Current Data Literacy Capabilities of University Teachers and Students [J]. Library, 2015, (12): 51-56, 62.

[12] Li Qing, Zhao Huanhuan. Research on Teacher Data Literacy Evaluation Indicator System [J]. e-Education Research, 2018, 39(10):

[13] Ma Teng, Sun Ling. Research on College Student Data Literacy Evaluation from the Perspective of Information Ecology [J]. Information Science, 2019, 37(8): 120-126.

[14] Hui Gongjian, Zeng Lei. Data Literacy in the Intelligent Era: Model Construction, Indicator System, and Cultivation Path—Based on Comparative Analysis of Domestic and Foreign Models [J]. Journal of Distance Education, 2021, 39(4): 52-61.

[15] Chen Jianming, Sun Xiaojun, Yang Bodi. Research on Evaluation Framework and Implementation Path of Data Analysis Literacy [J]. Journal of Mathematics Education, 2022, 31(2): 8-12, 57.

[16] Hao Yuanling, Shen Tingting, Gao Shan. Reflections and Suggestions on Data Literacy Education Practice in Universities—Based on Analysis of Harvard University Cases and Interviews with Chinese Library and Information Personnel [J]. Library and Information Service, 2015, 59(12): 44-51.

[17] Zhang Bin, Liu Sannüya, Liu Zhi, et al. Research on Cultivation Strategies for Normal College Students' Data Literacy Based on Big Data [J]. e-Education Research, 2017, 38(12): 86-91, 120.

[18] Hu Yiling, Zhang Qidi, Sun Ke, et al. Research on Cultivation Models and Applications of Primary and Secondary School Teachers' Data Literacy [J]. Chinese Journal of Distance Education, 2022, (3): 51-60.

[19] Zhang Jun. Research on Cultivating Researchers' Data Literacy for the Fourth Paradigm of Scientific Research [J]. Library and Information, 2016, (2):

[20] Wang Shiwei. Data-Driven Library and Information Science—A Bird's-Eye View of Hotspots in Library and Information Science in 2019 [J]. Information and Documentation Services, 2020, 41(1): 39-44.

[21] He Yali, Zhao Qingxiang, Xiao Peng. Library Strategic Planning and Implementation Strategies in the Data-Driven Era [J]. Library Tribune, 2020, 40(11): 98-104.

[22] Zhao Yanqiang, Zhou Bozhu. Construction of Subject Librarian 3.0 and Its Service System [J]. Library Science Research, 2021, (14):

[23] Gu Jiani, Yang Xianmin, Zheng Xudong, et al. Logical Framework and Practical Exploration of Data-Driven School Governance Modernization [J]. Modern Distance Education Research, 2020, 32(5): 25-34.

[24] Chen Xi. Development Path of Data-Driven Teaching Reform in Universities [J]. Theory and Practice of Education, 2020, 40(33):

[25] Yang Xianmin, Guo Liming, Wang Dongli, et al. Data-Driven Education Governance Modernization: Practical Framework, Realistic Challenges, and Implementation Path [J]. Modern Distance Education Research, 2020, 32(2): 73-84.

[26] Zhong Yaping, Wu Zhangzhong, Chen Xiaoping. Data-Driven Precision Training: Theoretical Connotation, Implementation Framework, and Promotion Path [J]. China Sport Science, 2021, 41(12): 48-61.

[27] Yuan Li, Zhang Jinhua. Reflections on Data-Driven Empowerment of Teacher Leadership in the New Era [J]. Contemporary Education Forum, 2021, (1):

[28] Wei Yali, Zhang Liang. From "Experience-Based" to "Data-Driven": New Teaching Paradigms in the Big Data Era [J]. Contemporary Education Science, 2022, (02): 50-56.

Note: Figure translations are in progress. See original paper for figures.