# Research Hotspots and Frontier Evolution of Foreign Big Data Management—A Visual Analysis Based on Web of Science Database

Yan Zhang[1] and Hui Xu[2(✉)]

[1] Guangdong University of Finance and Economics, No. 21, Luntou Road, Guangzhou, Guangdong, China

[2] Guangzhou College of Applied Science and Technology, No. 20, Fengle Road, Lianhua Town, Guangzhou, Guangdong, China

gdsxyxuhui@126.com

**Abstract.** With the advent of the era of big data, how to manage and utilize data better has become a hot topic at present. Based on the resources in the Web of Science database, this paper sorts out and integrates relevant research literature on big data management, so as to analyze the research status and hot topics of big data management. This paper uses the information visualization software Citespace to analyze the co-occurrence of the subject words in the literature published in foreign countries from 2016 to 2021. At the same time, it explores the research hotspots and knowledge structures related to the topic of big data management, and then reveals the development trend of hot topics in this research field. It lays the foundation for the research and development of big data management research in different sub-discipline field.

**Keywords:** Big data · Data management · Knowledge graph · Visual analysis

## 1 Introduction

With the advancement of science and technology and the continuous development of Internet technology, big data is playing an increasingly important role in different fields. The transition from "data" to "big data" has undergone tremendous changes. Big data refers to the large-capacity data that is difficult to obtain, store, manage and analyze by ordinary database software. Big data comes from various channels, so the amount of data and data collected is huge, and it has the characteristics of diversification of information. The rapid development of the information society has led to the birth of big data, which also marks the gradual transition of society from informatization to intelligence.

Big data has gradually become an indispensable and important resource in various fields, but it is also a "double-edged sword". High-precision and rapidly growing data provide unprecedented opportunities for improving the country's macro-scientific decision-making and social supervision and services. The mining and analysis of big data heralds the arrival of a new wave of productivity growth, which will have a profound


Z. Zhan et al. (Eds.): ICBDIE 2022, AHCS 5, pp. 237–245, 2023.
https://doi.org/10.2991/978-94-6463-034-3_24

impact on future economic development. Whether it is the prediction of individual behavior, consumer choice, search behavior, or the selection of traffic patterns, or machine learning and network analysis of disease outbreaks, big data plays an irreplaceable role [8]. Because big data can not only analyze patterns, but also provide predictive possibilities for events. At the same time, how to effectively manage and make good use of these data has become a problem that needs to be solved at present.

As the information society gradually becomes intelligent, the living environment of data has undergone major changes, and various management problems and risks have followed. The collection, storage, management and application of large-scale data will continue to face challenges, and the management model of big data will also usher in a series of changes. In recent years, the research on big data management has continued to deepen, foreign countries have successively promulgated data management and sharing policies, and the exploration of data management has continued to increase. Scholars have many different opinions on the management, technology and application of big data, which need to be further explored. Therefore, it is of great significance to analyze and discuss the research hotspots in the field of big data management abroad.

## 2 Data Sources and Research Methods

This article uses the Web of Science as the source, and uses the advanced retrieval function of the database to retrieve relevant literature. The selected database is the core collection of Web of Science, and the citation index is limited to Science Citation Index Expanded (SCI-EXPANDED). The search term in this study is “big data management”, and the search is limited to “topic”. The time span for selecting the publication date of the literature is from January 1, 2016 to December 31, 2021, a total of 5 years. The language of the literature is limited to English, and the type of literature only selects papers, review papers and conference proceedings, excluding online publications and editorial materials, and then retrieves all literature on big data management research. The retrieval time was February 26, 2022. Through data cleaning and eliminating duplicate literatures, the final data sample for this study is 6040.

This paper uses Citespace to organize and analyze the literature. CiteSpace was developed by Dr. Chaomei Chen from Drexel University, USA. As a knowledge graph visualization analysis tool, CiteSpace can show the development trend of a certain subject area over a period of time, thus laying the foundation for the follow-up research and development direction.The applicability of Citespace is very strong, and this method can be mastered and used proficiently by a wide range of users, especially scientific researchers and graduate students, to detect and visualize new trends and fundamental changes in the future development of scientific disciplines [3]. CiteSpace can perform co-occurrence analysis based on different elements, including countries, institutions, authors and keywords, etc. Co-occurrence analysis can quantify various information carriers and provide help for in-depth knowledge mining. It reflects information and related knowledge and lays the foundation for subsequent research. CiteSpace visualizes the connections and relationships between different documents, allowing scholars to have a deeper understanding of previous research in this field, and can also stimulate vision and innovation for future research prospects.

## 3 Research Hotspot Analysis

The core part of the literature can be represented by keywords, which are highly condensed and generalized to the subject of the article. Therefore, we use co-occurrence analysis on the keywords of the paper to better grasp the gist of the article. Each paper generally has more than one keyword, and there are inevitably some connections between different keywords. The frequency of co-occurrence can well reflect the association between these subject terms. The more frequently it appears in the literature, the closer the relationship between the themes is. Co-word analysis determines the relationship between topics in a subject area based on the frequency of co-occurrence of different words. The co-word analysis method also helps to stimulate innovation consciousness, and further search for new breakthroughs through the hot research of previous research.

Citespace visualizes the high-frequency keywords in the selected literature and derives the final keyword co-occurrence map, as shown in Fig. 1. The node represents the keyword, and its size shows the frequency of the keyword. The connection lines between nodes represent the co-occurrence of hotspots, and the number of connection lines reflects the closeness of the connection between different keywords. Hot research issues in different fields and future research trends can be reflected by the co-occurrence time zone map. Using Citespace to visualize each cluster, many irregular polygon clusters were obtained. Closely related keywords will be automatically clustered, and will be classified and labeled according to the meaning of the keywords. The order of the labels represents the degree of research hotspots in the cluster. This paper further analyzes the visual background data, and obtains the frequency and data information of hot keywords in the field of big data management research. Table 1 is obtained after eliminating the keywords with unclear meaning.

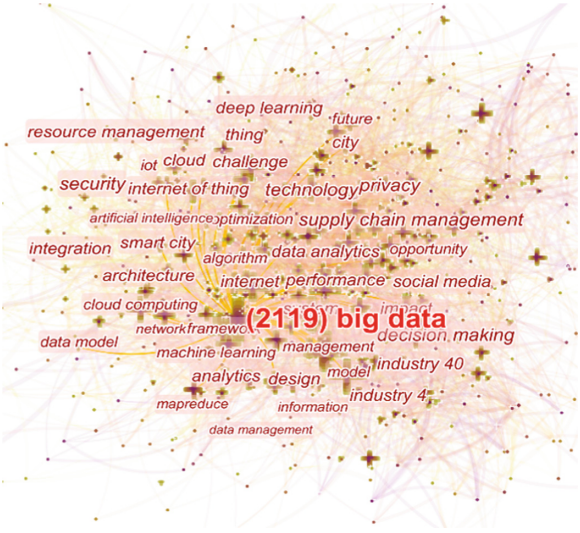


**Fig. 1.** Keyword co-occurrence map.

**Table 1.** Keyword Frequency Statistics.

| keywords | relevant information | |
|---|---|---|
| | number | frequency |
| big data | 1 | 2119 |
| management | 2 | 1248 |
| system | 3 | 599 |
| model | 4 | 537 |
| internet | 5 | 341 |
| challenge | 6 | 328 |
| framework | 7 | 322 |
| performance | 8 | 307 |
| big data analytics | 9 | 288 |
| machine learning | 10 | 282 |
| impact | 11 | 274 |
| cloud computing | 12 | 257 |
| internet of thing | 13 | 248 |
| technology | 14 | 226 |
| network | 15 | 206 |
| algorithm | 16 | 201 |
| design | 17 | 188 |
| information | 18 | 184 |
| data analytics | 19 | 180 |
| prediction | 20 | 175 |
| thing | 21 | 172 |
| optimization | 22 | 169 |
| artificial intelligence | 23 | 160 |
| analytics | 24 | 156 |
| smart city | 25 | 149 |
| supply chain management | 26 | 140 |
| neural network | 27 | 128 |
| future | 28 | 126 |
| classification | 29 | 126 |
| data mining | 30 | 123 |
| service | 31 | 120 |
| deep learning | 32 | 119 |

(*continued*)

**Table 1.** (*continued*)

| keywords | relevant information | |
|---|---|---|
| | number | frequency |
| architecture | 33 | 119 |
| risk | 34 | 117 |
| quality | 35 | 115 |
| strategy | 36 | 114 |
| decision making | 37 | 110 |
| supply chain | 38 | 107 |
| data management | 39 | 104 |
| cloud | 40 | 101 |
| data science | 41 | 101 |
| iot | 42 | 101 |
| resource management | 43 | 96 |
| predictive analytics | 44 | 94 |
| innovation | 45 | 91 |
| climate change | 46 | 88 |
| security | 47 | 87 |
| care | 48 | 86 |
| social media | 49 | 81 |
| city | 50 | 78 |

Combining the frequency statistics in Table 1, it can be seen that the nodes corresponding to "big data" and "management" are the largest and have the most occurrences, 2119 and 1248 times respectively. They have the most connections and relationships with other nodes, which is a core theme of research in this field. Secondly, the frequency of system appears 599 times, the frequency of model appears 537 times, the frequency of internet appears 341 times, and the frequency of challenge appears 328 times. The nodes of these subject headings are also relatively large, indicating that in the past five years, hot research on big data management has mainly focused on these areas.

### 3.1 Big Data Analysis and Management

As more and more public and private organizations are keen to gather information, the size of the datasets known as "big data" is far beyond the capabilities of conventional software tools and storage systems. This capability includes the ability to acquire, store, manage and process data within an acceptable time frame [11]. Therefore, the traditional data analysis and management operation mode will face major reforms, and the research on big data analysis methods and management system upgrades will gradually increase.

The sources of big data information and data are diverse and can be obtained in a variety of ways, including social media platforms and sensors. This information may contain useful information on issues such as national intelligence, cybersecurity, fraud detection, marketing, and healthcare informatics [13]. How to efficiently process a large amount of high-precision data is an urgent problem that needs to be solved.

Big data, with its huge sample size and high dimensionality, poses unique computational and statistical challenges, including scalability and storage bottlenecks, noise accumulation, spurious correlations, accidental endogeneity, and measurement errors [7]. For the complex problems of data capture, data storage, data analysis and data visualization of big data, it is necessary to establish a new management and service model to improve utilization efficiency, improve the traditional knowledge management model, and add intelligent design to the decision-making system. Gradually enhance the ability to systematically analyze data. At the same time, big data modeling, management systems, and visualization will enhance data management strategies [9]. Some foreign governments and institutions use distributed processing technology to manage and analyze big data through wireless network and cloud computing analysis.

Distributed storage databases based on Hadoop or Spark are also gradually entering the public eye. The management of massive amounts of data and the efficiency of analysis may be enhanced by the correlation between efficient management and big data aggregation. Data modeling is also an innovative approach. In the big data environment, data management needs to solve data complexity modeling and architecture problems through primitive methods, using modeling and development system technology to deal with the management mode of big data [1].

### 3.2 Big Data Technology

The improvement of data storage capacity and the improvement of computing processing power also promotes further breakthroughs in big data technology, which also brings more data to organizations than they have the computing resources and technology [13].

The process of processing big data is divided into: data collection, data preprocessing (data cleaning, data integration, data transformation, etc.), data storage, data mining and analysis, data presentation and application (data visualization, data security and privacy, etc.). The basis for processing analytical data lies in data acquisition. Obtain data from diverse data sources (including structured, semi-structured and unstructured types such as databases, text, images, videos, etc.), and then preprocess and store the data subject. A large amount of data is cleaned, integrated and transformed in this link to become "standard" data, which can be further analyzed and utilized. [12]. With the development of big data and related applications, researchers also need to collaborate with each other to jointly manage and develop databases [16].

The driving force of big data information technology mainly comes from practice and application. Because the storage capacity of traditional technologies is not large enough, management tools are not flexible enough to use to provide the environment required for big data [14]. At the same time, processing massive amounts of data with traditional technologies is expensive. Therefore, according to the needs of big data management, new methods and powerful technologies have emerged one after another. MapReduce, a distributed processing technology first proposed by Google, can be used in mobile

devices to analyze system failures and prevent the processing of applications using mobile devices [5]. The operation of this parallel mode quickly attracted widespread attention, and became a technical representative in the continuous system upgrade. After the big data technology is improved, the efficiency of data management and analysis can also be greatly improved.

### 3.3 Big Data Security and Privacy

The rapid development of technology is a "double-edged sword". On the one hand, people are enjoying the convenience provided by the Internet, and on the other hand, they are also threatened by privacy leakage. The extensive use of big data must take into account the privacy and security of users. However, due to the massive amount of data generated every day, it is inevitable that some security vulnerabilities will inevitably appear. Therefore, with the advent of the era of big data, people pay more and more attention to data security and privacy management. Data security involves various fields such as economy, society and culture. At present, people's words and deeds on the Internet are controlled by Internet merchants, including shopping habits, contact with friends, reading habits, retrieval habits and so on. A number of practical cases illustrate that even when harmless data is collected in large quantities, personal privacy can be exposed [6].

As the analysis and processing technology of big data continues to improve, it is gradually able to predict people's preferences and behaviors. And if this information is not properly kept or handled, it is easy to leak personal privacy. The large amount of data traces people leave on the Internet can easily provide opportunities for criminals. After they illegally collect information on the Internet, they carry out illegal activities, such as resale, fraud, etc., which will bring economic losses and various troubles to people's lives, and will also seriously affect social stability and harmony. Therefore, in the context of the prevalence of big data, how to ensure the privacy and security of the public is an urgent problem that needs to be solved at present.

With the continuous update and iteration of systems and technologies, a series of security mechanisms have been developed and used one after another. Big data has a complete life cycle, and the mechanisms corresponding to different cycles are different. These include stages such as data generation, data storage, and data processing [10]. For structured data, data release anonymity protection technology is the key technology to ensure that personal privacy is not leaked. Due to the prevalence of social networks now, the massive data generated by social networks also contain a large number of clients' private content. Its typical anonymity protection requirement is point anonymity, which can ensure that the user's identity and attributes are hidden when the data is released.

## 4 Conclusions

Through the visualization research on hot topics in the field of big data management abroad, the hot keywords in this field are obtained, including big data, management, system, model, internet, challenge, framework, performance, big data analytics, machine learning, impact, cloud computing, internet of thing, technology, network, algorithm, data analytics, etc. The big data management system is constantly

being updated and upgraded, while focusing on the combination of data analysis and knowledge management.

Big data technology is an indispensable new "weapon" for the development of social informatization to intelligence. Big data management will also continuously improve data management methods to meet the needs of different levels. Big data analysis and management will open up new application spaces in different fields and provide better data management and service platforms for various fields. The new challenges brought about by big data are different from those of the past, and because of the great value it contains, it is vulnerable to attack. Therefore, big data security and privacy management should not be underestimated.

Nowadays, big data is prevalent, and the processing methods and analysis technologies for big data are becoming more and more mature, and at the same time, they are also bringing innovations to traditional methods. Knowledge and information management systems are constantly being upgraded, and related management theories are constantly being integrated and innovated. Technologies in related fields such as artificial intelligence, Internet of Things, cloud computing, machine learning, and deep learning are also showing a trend of integrated applications. However, although scholars have made great progress in the related research of big data management, there are still many potential crises and challenges, and there are still many new fields that scholars need to continue to explore in the future.

**Acknowledgements.** Thanks to the predecessors for their contributions to this research area, and to the websites that provide digital platform support for this research.

## References

1. Alelaiwi, A. (2017). A collaborative resource management for big IoT data processing in Cloud. *Cluster Computing*, *20*(2), 1791-1799.
2. Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics*, 22(1), 155-205.
3. Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, *57*(3), 359-377.
4. Chen, X., & Liu, Y. (2020). Visualization analysis of high-speed railway research based on CiteSpace. *Transport Policy*, 85, 1-17.
5. Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.
6. Espinal, Y. (2013). Viktor Mayer-Schonberger and Kenneth Cukier, Big Data: A Revolution That Will Transform How We Live, Work and Think. *International Journal of Communication*, *7*, 3.
7. Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293-314.
8. George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of management Journal*, *57*(2), 321-326.
9. Gil, D., & Song, I. Y. (2016). Modeling and management of big data: challenges and opportunities. *Future Generation Computer Systems*, *63*, 96-99.

10. Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. *Journal of Big Data*, *3*(1), 1-25.
11. Kubick, W. R. (2012). Big data, information and meaning. Applied Clinical Trials, 21(2), 26.
12. Liao, J. (2015). Big data technology: current applications and prospects. Telecommunications science, 31(7), 1.
13. Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of big data*, *2*(1), 1-21.
14. Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, *30*(4), 431-448.
15. Sledgianowski, D., Gomaa, M., & Tan, C. (2017). Toward integration of Big Data, technology and information systems competencies into the accounting curriculum. *Journal of Accounting Education*, *38*, 81-93.
16. Storey, V. C., & Song, I. Y. (2017). Big data technologies and management: What conceptual modeling can do. *Data & Knowledge Engineering*, *108*, 50-67.