Research on Data Privacy Protection Strategies Based on Artificial Intelligence

Jun Yang*, Qiukai Ye, Jiujiang Han, Ming Xian

College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China

* Corresponding author:779753251@qq.com

ABSTRACT. With the rapid development of big data and artificial intelligence, the importance and urgency of data privacy protection issues have become increasingly prominent, making it a hot research area worldwide. However, the future research directions and hot topics in this field are not yet clear. To this end, this article uses the Web of Science core collection as the data source and uses bibliometrics to visually analyze 1395 related literature on artificial intelligence and data privacy protection, including quantitative analysis of articles, co-citation analysis, and keyword co-occurrence analysis. The results show that although China started relatively late in this field, it has developed rapidly and has become the country with the highest number of publications. The latest research hotspots in the field of data privacy protection focus on blockchain, edge computing and federated learning.

Keywords: data privacy; WOS; artificial intelligence

1 Introduction

Artificial intelligence has become a key field for countries to seek dominance in a new round of technological competition. As data is the basic element driving the rise of artificial intelligence, data security has become a key factor affecting the safe development of artificial intelligence. How to ensure data security in the context of artificial intelligence has become a hot research field internationally.

Bibliometrics is a cross-disciplinary science that uses mathematical and statistical methods to analyze all forms of knowledge carriers. By statistically analyzing and quantifying literature, in-depth research can be conducted on the progress of research in the field, the cooperation between researchers, and the future development trends. There are many examples to use bibliometric methods to study and summarize the achievements in a field [1,2].

Artificial intelligence and data privacy have become important research areas that concern the future development of nations. Countries around the world have shown a high level of interest in this field, but research has been scattered across various industries and fields, and there is still no clear direction for future development and research trends. This study explores and analyze the number of publications, coopera-

G. Guan et al. (eds.), *Proceedings of the 2023 3rd International Conference on Education, Information Management and Service Science (EIMSS 2023)*, Atlantis Highlights in Computer Sciences 16,
https://doi.org/10.2991/978-94-6463-264-4_41

tion relationships between countries, keyword co-occurrence, and clustering in this field using 1395 relevant articles from the Web of Science (WOS) core collection through bibliometrics. It provides a review and summary of research achievements and development trends in this field, which can serve as a reference for future research.

2 Data sources and research methodology

In order to ensure the authority and scientific of research data, this article selects Science Citation Index-Expanded citation index database in Web of Science core literature database as data source, using "artificial intelligence or AI" and "data privacy" as search topics. The types of literature are limited to journals and reviews. Due to the relatively short emergence time in this field, the retrieval time is set from 2014 to 2023. Finally, 1395 articles are retrieved as data sources.

This article utilizes bibliometric analysis and visualization software such as CiteSpace, Bibliometric and VOSviewer to analyze the retrieved literature, obtaining insights into research progress, hotspots, and emerging trends. Firstly, a bibliometric analysis platform will be used to analyze the number of publications by country and the collaborative relationships between countries. Changes in the number of publications will be examined to analyze development trends in the field. Then importing the article information into CiteSpace, the pathfinder -er method is implemented and network pruning is performed both on sliced and merged networks to analyze co-citation and clustering. In VOSviewer co-occurrence analysis are performed to generate keyword density maps to display the co-occurrence relationships between keywords.

3 Analysis and Results

3.1 Publication Volume Analysis

The change in the number of published articles can indicate the research enthusiasm and progress of a field. Through a bibliometric online platform, we have conducted a statistical analysis of the total number of articles published in this field by various countries. It can be seen that the number of publications in this field is increasing year by year, indicating that the research enthusiasm of this field is constantly growing. The distribution of publication years is shown in Figure 1.

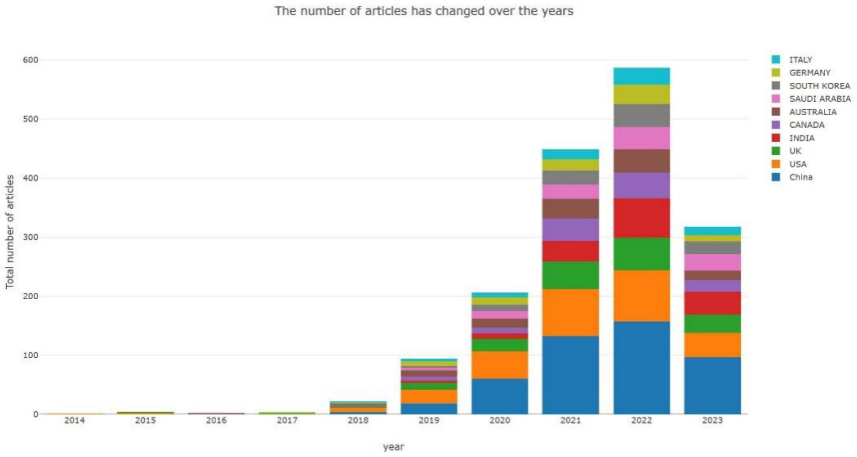

Fig. 1. The number of articles in each year

Based on the distribution and changes in the number of papers over time, we can see that research in the field of data privacy protection began to rise in 2018. Prior to 2018, there were only sporadic studies, with the United States having an earlier start in this field. From 2018 to 2019, the field entered a stage of rapid growth, with the United States being the main country for article publication. After 2019, the research in this field showed an explosive growth trend, with China's growth being particularly rapid. In terms of annual publication volume, China has surpassed the United States, indicating that China attaches increasing importance to research in the field of data privacy protection and the research is becoming more in-depth.

By conducting a literary analysis, we can gain insight into the international exchange situation in this field. The figure 2 illustrates the cooperation between countries. As evident from the graph, there is a high level of exchange between nations during the research process of this field, and oftentimes, new research results require collaboration among researchers from diverse backgrounds. Literature from China and the United States has been cited multiple times by other countries and exerts significant influence in this research domain. Research has shown that Americans are more concerned about privacy issues related to artificial intelligence and are more focused on privacy disclosure of AI applications. In contrast, Chinese people are more optimistic about the role of AI in promoting privacy protection.[3]

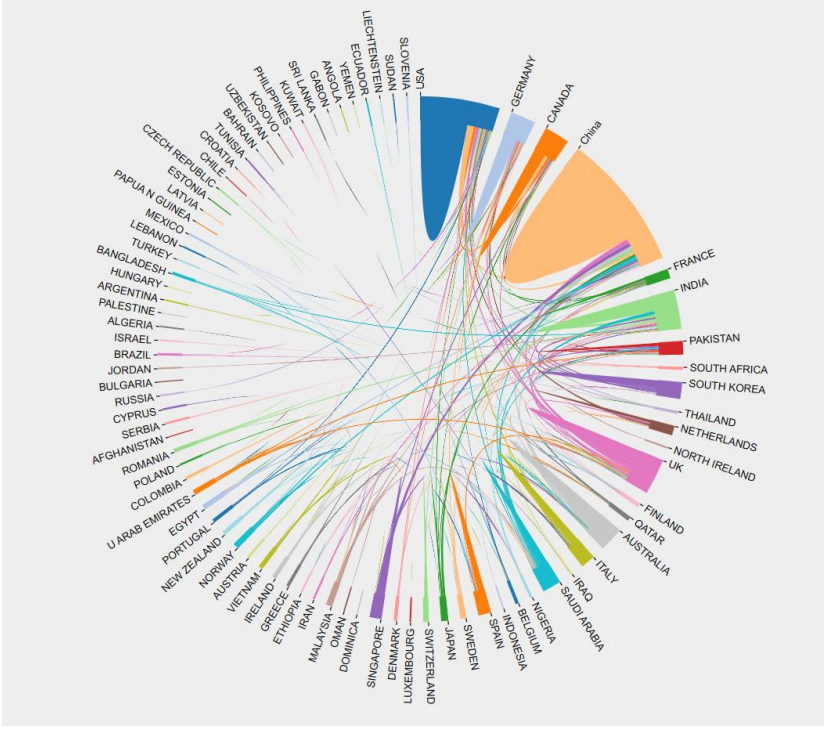

Fig. 2. Bibliographic citations between countries

3.2 Co-citation Analysis

Co-citation, also known as co-cited, refers to two articles that are simultaneously cited by a third article. It is generally believed that co-cited articles have a certain similarity in terms of their topics, so the co-citation frequency can measure the correlation of articles in terms of content. Co-citation analysis can reveal the research focus of scholars in a particular field. The frequency with which a paper is cited indicates its academic value within that domain. By using the LLR document clustering function in Citespace, the literature in the field can be clustered, and the clustering results are shown in Figure 3, where we can find that the central documents in the clustering are focused on "federated learning" , "internet of things" , "Unmanned vehicle driving," "blockchain" .

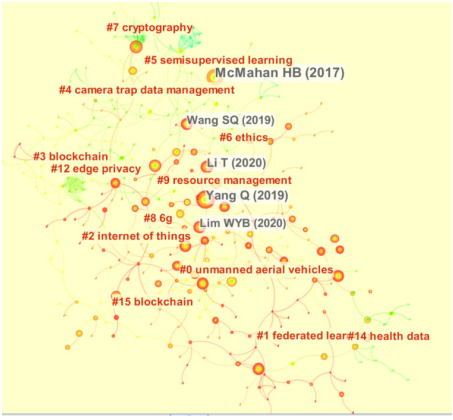

Fig. 3. The results of co-citation cluster analysis

After analysis, we obtained the most cited literature in this field, as shown in Table 1. From the table, it can be seen that the most cited articles are related to machine learning algorithms and federated learning, including the application of federated learning in edge computing and federated learning-based privacy enhancement techniques for machine learning. This indicates that the future development of this field is closely related to these algorithms, and many scholars are still continuing research in this field.[4,5]

Table 1. The most co-cited literatures

Literature name	Time	Co-cited	The paper source
A Survey of Algorithms and Analysis for Adaptive Online Learning	2017	107	JOURNAL OF MACHINE LEARNING RESEARCH
Federated Machine Learning: Concept and Applications	2019	86	ACM Transactions on Intelligent Systems and Technology
Federated Learning: Challenges, Methods, and Future Directions	2020	62	IEEE Signal Processing Magazine
Federated Learning in Mobile Edge Networks: A Comprehensive Survey	2020	56	IEEE Communications Surveys & Tutorials
Adaptive Federated Learning in Resource Constrained Edge Computing Systems	2019	55	IEEE Journal on Selected Areas in Communications

3.3 Keywords co-occurrence analysis

Repeatedly appearing hot keywords in literature indicate the popular topics within a particular field. To extract more representative hot keywords, VOSviewer software was used to perform keyword co-occurrence analysis on literature with a filter thresh-

old set at 10. This resulted in the filtering of 178 keywords and the generation of a density plot of their occurrence frequency, as shown in Figure 4.

Fig. 4. The results of keywords co-occurrence analysis

Fig. 5. Keyword cluster analysis

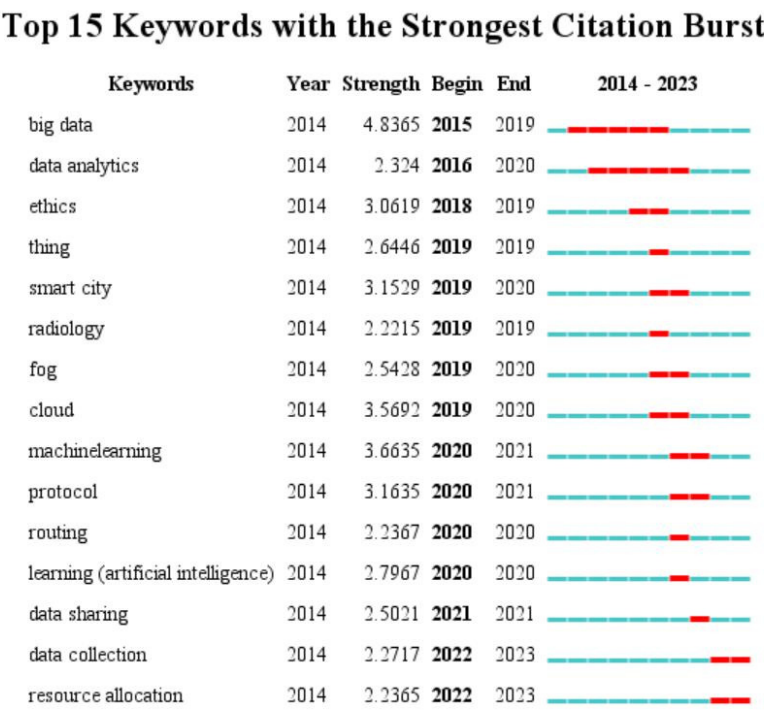

Fig. 6. Keyword outbreak record

By using the keyword clustering and keyword outbreak functions in CiteSpace, we can understand the development and changes of research hotspots, trends, and frontiers over a certain period of time. The situation of keyword clustering and keyword outbreak is shown in Figure5 and Figure6.

According to the analysis above, we can see that the keywords that appeared most frequently are deep learning, machine learning, federated learning, blockchain, security, and the internet. From the clustering results, it can be seen that areas such as blockchain, edge computing, and deep learning have higher attention. In medical and industrial IoT applications, protecting data privacy is particularly important.

Currently, there are many research achievements in artificial intelligence algorithms for data privacy protection. Such as Private Deep Learning based on collaborative deep learning [6], privacy-preserving machine learning under multiple keys [7]. Federated learning and blockchain as emerging technologies have demonstrated their enormous potential in data privacy protection, and have been widely applied to many fields of data privacy protection, such as privacy protection in sensor communication based on federated learning in the Internet of Things [8], and data privacy trustworthy sharing using blockchain and federated learning [9]. The increasing awareness of public data privacy protection has prompted research into new data collection paradigms [10].

4 Conclusion

In this study, we downloaded a total of 1395 articles from the Web of Science core collection. We conducted article count analysis, co-citation analysis, and keyword analysis using Bibliometric, Citespace, and VOSviewer tools. The following findings are preliminary conclusions drawn from our analysis:

(1) According to the number of papers published, the number of papers published by countries around the world has been continuously increasing in recent years. Based on the development trend, it can be predicted that the development of artificial intelligence in the field of data privacy protection will continue to deepen.

(2) From the results of co-citation analysis on literature, it can be seen that federated learning, blockchain, and machine learning algorithms have a high level of attention in the field of data privacy protection. This indicates that the development of the field of data privacy protection relies on the upgrading of traditional algorithms and the emergence of new algorithms.

(3) With the widespread use of the Internet of Things and the surge in the use of artificial intelligence, the advantages of edge computing are gradually emerging. How to protect data privacy security in edge computing will become an important research direction in the future.

(4) The results from the co-occurrence analysis of keywords indicate that artificial intelligence algorithms based on data privacy and security are constantly emerging. In the future, edge computing will become an important research direction in the field of data privacy protection. Technologies such as federated learning and blockchain have great potential in the field of data privacy protection. The fields of Medical, and Internet of Things are highly related to data privacy security.

Overall, research in the field of data privacy protection is currently in a rapid development stage, with various protection strategies constantly emerging, each with its own advantages and disadvantages [11]. As people's awareness of data privacy protection continues to increase, data privacy and security will surely become a new pillar and a new field that will drive future development. The development and research of AI privacy and security tools will also become a hotspot [12]. In the next step, we will conduct more in-depth and practical research, such as Enhancing Data Privacy in User Behavior Analysis.

References

1. Han J, Xian M, Liu J, Wang H. (2021) Research on the Application of Artificial Intelligence in Cyber Range. Journal of physics. Conference series 2030:12084
2. Li Z, Xian M, Liu J, Wang H. (2020) The Development Trend of Artificial Intelligence in Cyberspace Security: A Brief Survey. Journal of Physics: Conference Series; 1486:22047.
3. Xing Y, He W, Zhang JZ, Cao G. (2023) AI Privacy Opinions between US and Chinese People. Journal of Computer Information Systems :492-506.
4. Soykan EU, Karaçay L, Karakoç F, Tomur E. (2022) A Survey and Guideline on Privacy Enhancing Technologies for Collaborative Machine Learning. IEEE Access :97495-97519.

5. Le K, Le-Minh K, Thai H. (2021) BrainyEdge: An AI-enabled framework for IoT edge computing. ICT Express :211-221.
6. Zhao Q, Zhao C, Cui S, Jing S, Chen Z. (2020) PrivateDL: Privacy-preserving collaborative deep learning against leakage from gradient sharing. International Journal of Intelligent Systems; 35:1262-1279
7. Li PLP, Li TLT, Ye HYH, Li JLJ, Chen XCX, Xiang YXY. (2018) Privacy-preserving machine learning with multiple data providers (Article). Future Generation Computer Systems :341-350.
8. Manzoor SI, Jain S, Singh Y, Singh H (2023) Federated Learning Based Privacy Ensured Sensor Communication in IoT Networks: A Taxonomy, Threats and Attacks. IEEE Access; 11:42248-42275
9. Guo S, Zhang K, Gong B, Chen L, Ren Y, Qi F, Qiu X. (2022) Sandbox Computing: A Data Privacy Trusted Sharing Paradigm via Blockchain and Federated Learning. IEEE Transactions on Computers :1-12.
10. Ding J, Ding B. (2022) Interval Privacy: A Framework for Privacy-Preserving Data Collection. IEEE Transactions on Signal Processing :2443-2459.
11. Yin X, Zhu Y, Hu J. (2022) A Comprehensive Survey of Privacy-preserving Federated Learning. ACM Computing Surveys; 54:1-36.
12. Goldsteen A, Saadi O, Shmelkin R, Shachor S, Razinkov N. (2023) AI privacy toolkit. SoftwareX: 101352.