WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n2.02
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Privacy Protection Technology Based on Big Data Technology

Rundong Xu

Faculty of Automation, Nanjing University of Information Science and Technology, Nanjing, 210044, China

202283250033@Nuist.edu.cn

ABSTRACT

In the era of big data, with the development of the Internet industry, privacy protection has become a non-negligible part of network security. At the same time, big data technology has also been integrated into various industries, which is of great significance to the development of all walks of life. The inability of security technology to meet the rapidly evolving needs of this field has become an urgent problem in many fields. Network technology, cloud computing technology, and artificial intelligence have all taken the lead in the development of the front-end of the industry, making them play a role on the Internet platform in the era of big data. At this time, the lack of privacy and security protection has become a huge loophole that endangers the security of personal information and personal property. In order to maintain normal development, it is necessary to invest technologies such as big data in the research of privacy protection. Big data technology has significant advantages in information analysis insight, information integration, and data mining capabilities. The development and utilization of new privacy-preserving technologies have been developed and utilized to address the privacy concerns of the new era. This article will provide several mainstream and relatively complete privacy protection technologies based on the main problems faced by privacy protection. Through the different integration of several mainstream technologies, a variety of privacy protection technologies that can deal with problems in different directions have been formed, acting in different fields, so that users' privacy and security can be more perfectly protected.

KEYWORDS

Big data; Privacy protection; Computer technology; Algorithms

1. INTRODUCTION

As an emerging Internet technology, big data technology has its unique advantages: huge amount of data, diverse data types, and rapid data processing speed. Therefore, big data technology has been developed and has played an important role in the fields of computer and Internet, and has become the mainstay of data science and technology today. With the help of the above advantages, the big data industry has developed rapidly, and data has become an important strategic resource for countries and enterprises. In the hands of some people, the hot big data technology has become a sharp blade to tear open network vulnerabilities, using it to steal users' personal information and enterprise confidential data, posing a threat to network security. At the same time, privacy protection technology is also continuously investing in the application of big data, making big data the key to maintaining and ensuring privacy security. Therefore, the application of big data technology in privacy protection can be summarized

Its features are easy to classify and summarize, and provide reference for privacy protection in other emerging fields.

2. BACKGROUND OF THE ERA OF BIG DATA

2.1. Opportunities in the Context of Big Data

First of all, big data provides unprecedented information resources and provides a broad space for research and application in various fields. Deep mining and analysis of big data can lead to more knowledge and a deeper understanding of the world. Second, big data provides new business models and service models. For example, through big data analytics, businesses can more accurately understand the needs of consumers and provide more personalized services. Another example is that the government can optimize public services and improve the efficiency and effectiveness of public management through big data analysis. In addition, big data has also provided a new impetus for the development of technology. In order to process and analyze big data, new computing models, storage technologies, and analytical tools need to be developed. This will promote the advancement of computer software technology and trigger technological innovation. Finally, the impact of big data on society cannot be ignored. The application of big data is profoundly changing the way we live, work and think. From individuals to society, from business to government, the impact of big data is everywhere.

2.2. The Challenges Posed By Big Data

A large amount of private data is leaked into the Internet, which will not only cause data security problems for individuals and enterprises, but also affect national security. The collection and use of data is often without the consent of the data owner, which leads to data abuse and privacy leakage, and the arbitrary use of data by data owners and data collectors has brought certain difficulties to privacy protection.

Big data is a double-edged sword. While it helps us obtain information with great efficiency, it will also be maliciously used by some criminals and become a tool for them to steal personal privacy. Not only that, but many third-party software will also try to pry the permissions of our devices and try to get our personal information. Big data is like a sharp knife, and there is no place where it goes. Therefore, we need to explore how to use the assistance of other technologies to make big data a shield for us to protect privacy when others face the same sharp spear.

2.3. The Status Quo of User Privacy in the Context of Big Data

Some companies exceed their own authority when using customer information, and even sell customers' private data to make huge profits, resulting in the leakage of users' private information. In addition, due to the socialization and sharing of big data, hackers take data as the primary target of attacks, and some hackers even reach out to national security agencies to threaten national security. In terms of information security, big data will become a source of frequent security problems. In the era of big data, it is no longer surprising that various online platforms and software collect user data, and as long as the data is not leaked, it will not pose a threat to the security of personal data. Data privacy is analyzed from multiple perspectives, with many attributes and classification rules that include relevant records such as user behavior and traces on the platform, in addition to the customer's personally identifiable information. Privacy generally refers to sensitive information that the party in possession of the data does not want to be known to the outside world, such as personally identifiable information, behavioral trace information, and web browsing information. From the perspective of the owner of privacy, privacy can be divided into personal privacy and ordinary privacy. Personal privacy refers to information that has obvious individual characteristics or is related to an identifiable individual, but the individual is unwilling to disclose it. Ordinary privacy refers to information that can reveal connections or commonalities between multiple individuals and is not willing to be revealed. The classification varies from different perspectives.

3. PRIVACY PROTECTION TECHNOLOGY

3.1. Overview of Privacy Protection Technologies

Privacy is a kind of user's information, and in the era of the Internet, the object of privacy protection for users is a huge amount of user data. In addition to the user's personally identifiable information, the user's browsing and behavior traces on the platform, potential preferences and other information, and even a large amount of sensitive information and data of the company or platform may be invaded and stolen by criminals. Among them, big data has become the source of frequent security problems, and it will also be an important tool to solve problems and improve protection.

Privacy protection technologies can be broadly divided into four directions, namely, data encryption, data desensitization, access control, and intrusion detection. A perfect privacy protection technology often needs to be processed from multiple directions in order to achieve effective protection. This chapter introduces and analyzes the current mainstream privacy protection technologies, including differential privacy, homomorphic encryption, and blocking Chain data privacy, text intelligent encryption method, k-anonymity, etc.

3.2. Privacy-preserving Technology for Recommendation Algorithms

(1) Overview of recommendation algorithms

The recommendation algorithm is an algorithm designed in the recommendation system that can balance the accuracy of recommendation and the intensity of privacy protection. Because the user's data in the recommendation system is plaintext and not desensitized, it is easy for the user's privacy to be inferred, snooped on or intercepted by Internet attackers, resulting in more and more users worrying about the leakage of personal privacy information. Therefore, this algorithm is used to strengthen privacy protection security. This section lists the common privacy protection techniques in recommendation algorithms, including differential privacy technology, homomorphic encryption technology, secure multi-party computing technology, and federated learning technology. Table 1 summarizes various privacy-preserving technologies from three aspects: advantages, disadvantages, and typical applications [1].

(2) Differential privacy

Under the differential privacy protection framework, noise is inserted into the return value of the query function to distort the data, and at the same time, the data features are not deviated due to noise insertion. Therefore, when the source data is attacked, differential privacy technology can minimize the risk of individual data leakage, so even if the attacker knows all the background knowledge, it is impossible to infer a certain attribute of a single person in the final recommendation result, and the recommendation system can still achieve a safe recommendation with the same prediction data trend.

(3) Homomorphic encryption

The homomorphic encryption technology allows users to directly calculate the ciphertext data, so that the decrypted data participating in the model training is consistent with the output results obtained by the original data after the same processing, which realizes the coexistence of data privacy protection and data processing, ensures the security of information in the process of data processing, and fundamentally solves the problem of user privacy leakage. In the recommendation system under privacy protection, homomorphic encryption technology is often used to encrypt user characteristics or intermediate results, so as to realize the recommendation system to carry out the relevant operations of the recommendation algorithm without seeing the specific information of the user.

(4) Secure multi-party computation

Secure multi-party computation (SMPC) is a privacy-preserving technology based on cryptography, which enables multiple parties who do not trust each other to operate together without a trusted third party, ensuring that each participant cannot infer any valuable additional information from the interaction data in the calculation process, and can only obtain its own input data and the final calculation result, ensuring the security of user information.

3.3. Semantic-based Privacy-Preserving Technology

(1) Overview of LBS Privacy Protection

LBS stands for Location-based Service, which relies on a large amount of diverse and accurate location data, and once stolen, it will pose a great threat to the security of users. Therefore, a privacy protection technology based on semantics is designed to give different degrees of encryption and anonymity protection through the sensitivity of personalized semantics, so as to reduce the risk of privacy data leakage. This section lists the common privacy-preserving techniques in recommendation algorithms, including k-anonymity, differential privacy, and pseudonym [2].

(2) k- Anonymous

k-anonymity is the basis of many privacy protection methods, and its principle is to ensure that each individual record in the anonymity set sent to the LBS server is indistinguishable from other k-1 individuals in terms of sensitive attributes, so the probability of being attacked is 1/k. In order to further protect user information, many technologies will also combine query probability and time characteristics for anonymous protection.

(3) Differential privacy

Differential privacy uses the Laplace mechanism and the exponential mechanism to protect the privacy of numeric and non-numeric data. For example, the Hilbert curve is used to map the user's position to one-dimensional space, and the Laplace mechanism is used to perturbate the location information [3].

(4) Pseudonymous technology

Pseudonymization breaks the association between the user and the query by giving the user a temporary pseudonym, and reduces the chance of attackers inferring the user's identity by changing the pseudonym frequently.

3.4. Privacy Protection Based on Explicit and Implicit Feedback

Implicit feedback refers to behaviors that do not clearly reflect the user's preferences, such as purchases, clicks, favorites, etc. Nowadays, implicit feedback has gradually replaced explicit feedback as an important feedback method that can reflect user interaction behavior. Explicit feedback refers to the behavior of clearly expressing the user's interests, and the combination of these two types of feedback mechanisms is put into the application of actual network information. In privacy protection based on this feedback method, the differential privacy algorithm is still applied, and the principle and mechanism are not much different from the first two [4].

4. DISCUSSION OF BIG DATA TECHNOLOGY IN PRIVACY PROTECTION

4.1. Summary of Big Data Technologies in the Field of Privacy Protection

In the previous section, recommendation systems, LBS systems, and personal information systems based on explicit and implicit feedback were mentioned, and this section summarizes two main information points from the common privacy protection technologies mentioned above:

First, the privacy protection technology that can apply big data technology must also be related to the application of big data. Big data is used in most fields due to its wide applicability and efficient execution. From computer science, which is closely related, to traditional agriculture, industry and medicine in the new era, as long as the amount of data is huge, it needs the support of big data, and when it comes to the information collection and data processing of big data, its privacy protection and security also need big data technology to maintain. [5] However, for example, in the fields of finance and industry involving blockchain, due to the structure of the blockchain itself and the high data security, most of its privacy protection methods use technologies such as machine learning and homomorphic encryption, and the application of big data technology is less [6].

Therefore, big data technology and privacy protection technology can be said to be very closely related, and the application of big data can be seen in the main privacy protection technologies at this stage. This is not only because in the context of big data and the Internet, privacy is presented as a kind of information, in the form of a large amount of data, and its protection and application are naturally closely related to big data technology. It is also because the object of privacy protection is a large amount of user information or private data of companies and institutions, rather than a simple individual information, so huge data and complex processing are required to be supported by more mature big data technology.

Second, the application of big data technology in privacy protection is mainly reflected in the three steps of data encryption, data masking, and access control, and the most commonly used method is data desensitization using differential privacy protection technology. There are many ways to encrypt data, such as isomorphic encryption, k-anonymity, and pseudonymization, all of which are involved in big data technology. The semantic-based privacy protection and differential privacy methods can play a role in access control. In different fields, according to the different needs and working principles of the system, different privacy protection methods will be selected to combine, so as not to affect the use of the system, but also to achieve the purpose of privacy protection. For example, in order to ensure the accuracy of the recommendation data, the recommendation algorithm used in the recommendation system adopts a combination of data encryption and data desensitization to ensure the accuracy of the recommendation and the security of private data.

Therefore, in the next section, we will focus on these mainstream and mature privacy protection technologies, analyze the applications of big data technologies, and summarize their similar components.

4.2. Summary of Big Data Technology in Privacy Protection Technology

(1) Differential privacy protection

Differential privacy is a privacy protection technology used in most fields, and its principle is based on data perturbation technology to balance data availability and privacy protection, and achieve the purpose of data desensitization and access control. It protects a single data in the database by adding noise that conforms to a specific probability distribution to the real data, and the Laplace mechanism maps the determined query and analysis results to an uncertain value range. [7] Differential privacy

has a solid mathematical foundation, and in the process of its implementation, there is no lack of application to obtaining and manipulating datasets [8].

Taking Sun Daozhu's experiment of differential privacy protection algorithm based on explicit and implicit feedback collaborative filtering as an example, in the experiment, the experimenter uses the MovieLens dataset, which contains the rating data of multiple users for multiple movies, as well as the movie data information and user attribute information. [7] Through the Laplace mechanism, the mean perturbation and gradient perturbation are added to the step of calculating the mean value to obtain the parameters of the gradient descent solving model, which are denoted as the DPEifSVD algorithm. Finally, by introducing a large amount of data, the result of satisfying ε-differential privacy is obtained.

(2) k- Anonymous

The principle of k-anonymity is to define several attribute identifiers, as well as an anonymity algorithm, through generalization and anonymity techniques, to publish data with low precision, so that there are at least k records of the same quasi-identifier, so that the observer cannot connect the records through the quasi-identifier. The core idea is to generalize the specific location of the user into an ambiguous area (also an anonymous area), which contains at least k-1 other users, and the probability of each user being identified is 1/k, so that the observer cannot identify the user by a quasi-identifier with a confidence level higher than 1/k. When the user needs the location service, the k-anonymity algorithm calculates the anonymity area and sends the anonymity area to the LBS server instead of the user's real location for the corresponding query service [9].

Taking Yao Jitao's research and implementation experiment of Spark-based user privacy protection query optimization algorithm for spatiotemporal data as an example, in the experiment, based on the Hadoop YARN platform, using MapReduce, Spark and other computing frameworks, the experimenter found the KB-CPIR algorithm with higher optimization degree by means of comparative experiments, and analyzed the advantages and application prospects of the privacy protection technology [10].

(3) Analysis and summary

Through the above two mainstream privacy protection technologies and two experiments related to them, it can be seen that the main ways and fields of privacy protection technology are related to big data technology. The former uses datasets for data collection and data storage, while the latter chooses computing frameworks such as the Hadoop platform and Spark for experiments. And both use different algorithms, visualization techniques, and data management methods in data analysis and processing. In addition to the specific analysis of these two technologies, homomorphic encryption, secure multi-party computing, and blockchain applications also involve and help big data technology.

5. SUMMARY

To sum up, from big data processing technologies, such as MapReduce programming models, Hadoop and Spark platforms, NoSQL and NewSQL database technologies, to big data analysis technologies, such as machine learning and deep learning, data mining and data visualization, to big data storage and management technologies, such as distributed file systems, Data lakes and data warehouses, as well as big data security and privacy protection technologies, we can see that the application of computer software technology in the context of big data is diversified and comprehensive.

The advent of the era of big data has indeed brought a lot of convenience to people's lives. The various privacy protection methods it extends to do provide great protection for companies and individuals. But at the same time, we cannot ignore the threat brought by the continuous upgrading and development of computer technology to the privacy protection of society and individuals. Technology

is just an emotionless tool, and what kind of role it plays depends on what kind of person the user is. In today's ever-expanding power of algorithms, users, platforms, and society should face up to the issue of privacy protection in the era of algorithms, make trade-offs and balances in contradictions and dilemmas, and take corresponding measures to reduce the risks brought by algorithms to privacy protection. In the new challenges of "scanning data" and "fragmented data" that may arise in the future, users, platforms and society should also adhere to the bottom line of privacy, put the interests of society and the public first, and avoid privacy becoming "open and private".

REFERENCES

[1] Feng, H., Yi, H. (2023) A review of privacy protection research on recommender systems. Computer Science and Exploration, 17

[2] Li, W., Wu, H. (2023) A review of location privacy protection based on semantics. Computer applications, 43

[3] Lei, C., Zhang, L. (2023) Personalized semantically sensitive trajectory data publishing algorithms. Small and micro computer systems, 9

[4] Yang, S. (2023) A personal information privacy protection method based on explicit and implicit feedback. Information Science, 21(11)

[5] Xiang, Y., Yang, L. (2023) Research on power line loss data sharing based on differential privacy protection. Computer applications and software, 40

[6] Sun, G., Wan, M. (2023) Analysis of Blockchain Transaction Privacy Protection. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition)

[7] Sun, D., Li, N. (2023) A differential privacy protection algorithm that integrates explicit and implicit feedback collaborative filtering. Computer Applied Research., 38(08)

[8] DWORK C, MCSHERRY F, NISSIM K, R.A. (2006) Calibrating noise to sensitivity in private data analysis. The 21st International Symposium on Computer and Information Sciences, Istanbul, 265-284

[9] Li, Z. (2023) Design and implementation of medical information release system based on global K-anonymity. Thesis of China University of Mining and Technology

[10] Yao, J. (2023) Research and implementation of user privacy protection query optimization algorithm for spatiotemporal data based on Spark. Thesis of Northeastern University