Overview of big data information privacy protection technology and challenges

Hanze Wan

Jingling High School, Hexi Campus, Nanjing, Jiangsu, 210019, China
Email: ericwanhz199@gmail.com

Abstract:

With the rapid development of advanced technologies like the mobile internet, cloud computing, and the Internet of Things, the era of big data is gradually upon us. In this era, big data has been widely utilized in various fields like business, finance, social networking, and healthcare, and the value of data has become increasingly prominent. Data has become a new factor of production, the digital economy has become a new engine of development, and an important driving force for promoting social progress and economic development. However, with the advancement of big data technology, the risks of data leakage and misuse have increased. Personal data information leakage or misuse can lead to personal privacy infringement and fraud, such as the Sina Weibo data leak in March 2020 where attackers breached part of the Sina Weibo database, obtaining sensitive data including real names, website usernames, genders, locations, and phone numbers, impacting 538 million Weibo users. Therefore, there is an increasing demand for data information privacy protection. This paper will explore the core technologies and application scenarios in data information privacy protection and investigate data privacy protection in real-life settings.

Keywords: Big data, Data privacy, Differential privacy, Data anonymity, Encryption technology

1 Introduction

1.1 The challenge of data and information privacy protection

Data privacy protection aims to maximize the utilization of data while minimizing the risk of data leakage. The main objective of data privacy protection is to safeguard sensitive information from being infringed upon by others [1]. With the advent of the era of big data, the protection of data and information privacy is facing unprecedented challenges.

Firstly, with the in-depth application of smart phones and Internet of Things technology, data collection is becoming increasingly diverse, making data collection more convenient. However, many users lack awareness of personal privacy protection.

Secondly, in the era of big data, the use and analysis of data has become more profound, and the requirements for data collection have become more and more precise, covering more extensive categories, leading to an increase in the risk of data leakage and misuse.

Thirdly, the disclosure of private data will likely have a huge impact on individuals, organizations, and even society as a whole. For instance, the risk of financial fraud increases.

2 Core technology of data and information privacy protection

Under the background of the new era, how to in the development of big data technology, facing the challenge of data privacy protection, strengthen the research and application of large data information privacy protection technology, establish a more reliable, safe, credible big data environment, provide a solid foundation for the sustainable development of society and support, become an important problem we explore.

At present, the commonly used data and information privacy protection technologies mainly include differential privacy, data anonymity, data encryption and other privacy protection technologies.

2.1 Differential privacy and privacy protection technology

Differential privacy is based on data distortion, first proposed by Microsoft distinguished scientist Cynthia Dwork in 2006[2]. It is mainly achieved by adding noise to the original data to achieve privacy protection. The purpose is to make the query results of the data set insensitive to the changes of a single record in the data set, so as to hide the real data and avoid the attackers with background knowledge to obtain private information through guessing.

Differential privacy, as the name implies, is to prevent differential attacks, that is, if a class announces the results:

there are 50 students in our class, among which 10 of them fail, then if the attacker can get the results of 49 students, you can predict who the last student fails. If a class says that we have 50 students with a failure rate of 20%, and the attacker can only get 49 students with a failure rate of 18.36%, then he can't predict who the last student will fail. This is the core idea of differential privacy: for two data sets with only one record in difference, the yes query has a very close probability of obtaining the same value.

The data after differential privacy disturbance can satisfy the nature of two aspects: first, the attacker cannot reconstruct the real original data through the relevant background data after obtaining the disturbed data; second, after the data has added noise, the attributes of the data or data can be kept unchanged, such as statistical analysis of the data after adding noise. The size of the noise added to the differential privacy is independent of the data set size, so according to this advantage, the differential privacy has a good application experience in large datasets. At present, differentia privacy mainly applies relational database, recommendation system, trajectory information privacy protection and other scenarios.

2.2 Data anonymity and privacy protection technology

Data anonymous privacy protection technology is a privacy protection technology based on restricted release. Data anonymous privacy protection is primarily achieved by selectively releasing the risk of raw data, or not publishing or publishing sensitive data with low accuracy [3]. Currently, data anonymous privacy protection technology mainly includes the k-anonymous model, l-diversity model, t-close model, etc.

The k-anonymity model mainly aims to solve the problem of privacy leakage caused by user data link attacks. The k-anonymity is primarily manifest in generalization and concealment technology [4].

Generalization (Summary): Also known as generalization, this refers to the data being described more generally and abstractly. The original quasi-identifier attribute value is replaced with a general range value, making it impossible to distinguish the original specific value. For example, in the case of age, it may be summarized as an age range (e.g., age > 30 in Table 2 below).

Concealment (Suppression): Also known as inhibition, compression, certain information is not released. For example, the class may be replaced with an asterisk (*) in Table 2 below. By reducing the accuracy of published data, each record has at least the same quasi-identifier attribute value as the other K-1 records in the data table, thereby reducing the risk of privacy leakage caused by link attacks. (Table 2 below)

Table 1 Original Table

Student ID	classes and grades in school	age	surname and personal name
1	Class 1, Grade 1	18	J ack
2	Class 1, Grade 1	20	T om
3	Class 2, Grade 1	22	H elen
4	Class 2, Grade 1	26	J ackie
5	Class 3, Grade 1	36	J ohn
6	Class 3, Grade 1	39	C andy

Table 2 K – an anonymous schematic diagram

Student ID	classes and grades in school	age	surname and personal name
1	Class 1 *	age<=20	J ack
2	Class 1 *	age<=20	T om
3	Class 1 *	20<age <=30	H elen
4	Class 1 *	20<age <=30	J ackie
5	Class 1 *	a ge >30	J ohn
6	Class 1 *	a ge >30	C andy

Let's assume that the original table's name is hidden, but an attacker can still locate a record through the class and age. However, after k-anonymity is applied, the class and age are summarized and obscured. Even if the attacker knows the specific class and age of a student, they cannot retrieve the student's name. The same quasi-identifier {class from the table above, age} must have at least 2 records, making it a 2-anonymous model. Implementing the k-anonymous model ensures that the attacker cannot identify a user with a confidence level greater than 1/k using a quasi-identifier.

2.3 Data encryption, privacy protection technology

The use of encryption has always been the core solution in terms of protecting data privacy. Even in today's big data era, the application of data encryption to safeguard data privacy remains a common and essential technology.

In the protection of data information privacy, commonly used data encryption technologies include homomorphic encryption, attribute encryption, etc.

Homomorphic encryption (HE) allows encrypted data to be calculated directly without decrypting it [5]. The results of the operation are the same as that of the unencrypted data, so there is no need to know the original text during the operation. It is an encryption method that enables third parties to perform certain computable functions on encrypted data while retaining the functions and formatting features of the encrypted data.

Homomorphic encryption is often used in distributed environments, employing various encryption techniques to hide sensitive information in methods such as secure multi-party computing and other privacy computing scenarios.

3 Outlook of big data information privacy protection

The privacy protection of big data information is a complex and long-term system engineering, in today's increasingly developed technology and increasingly prominent data value. While meeting the business needs of customers, it is also necessary to ensure the security and privacy of data.

3.1 Strengthen the supervision of big data information platforms

With the promulgation and implementation of the Data Safety Law of the People's Republic of China and the Personal Information Protection Law of the People's Republic of China, the user information has been clearly stipulated and detailed, and excessive user information shall not be collected, and the collected data information shall not be shared and used at will. Therefore, it is necessary to strengthen the supervision of big data platforms, clarify the full responsibility of data, and clarify the responsibilities and obligations of users.

3.2 Increase the technology innovation of data and information privacy protection

At present, big data privacy protection technology has made good progress, but at present, there is a lack of relatively powerful and universal privacy protection solutions. Therefore, it is necessary to comprehensively strengthen technological innovation, increase technology research and development and upgrading, and effectively do a good job in data privacy protection.

3.3 Increase education and training on data and information privacy protection

It is essential to enhance the publicity and education regarding data and information privacy protection. Establishing a multi-dimensional publicity and education system is necessary to gradually improve individuals' and enterprises' awareness of data and information privacy protection. Additionally, strengthening the training of data and information privacy protection measures is necessary to improve users' ability to safeguard their data and information privacy. By practicing these measures in daily work, it can effectively protect the privacy of data and information.

4 Conclusion

To sum up, with the rapid development of advanced technologies such as the mobile Internet, cloud computing and the Internet of Things, data and information privacy protection technology has become a hot topic. This paper focuses on the basic definition, related technical concepts, and application scenarios of the current common privacy protection technologies such as differential privacy, data anonymity, and data encryption. Big data information privacy protection is a complex and long-term system engineering. In real life, we also need to study the in-depth use strategies of the above technologies, such as noise disturbance strategies such as differential privacy, for different practical scenarios. At present, the research of data information privacy protection in China is still in its initial stage. We still need to strengthen the supervision of big data information platform, increase the innovation and education training of data information privacy protection technology, effectively improve the national awareness of information privacy protection, and strive to better promote the healthy development of big data technology.

Reference

[1] Wang Rui. On data privacy protection [D]. Zhengzhou University, 2022.DOI:10.27466/d.cn ki.gzzdu.2022.001985

[2] Sun Daozhu. Classification and recommendation algorithms based on differential privacy protection [D]. The Strategic Support Force Information Engineering University, 2021. DOI:10.27188/d.cnki.gzjxu.2021.000120

[3] Peng-Ningbo. Review of domestic data privacy protection studies [J]. Library, 2021, (11): 69-75.

[4] Yang Fengjiao. K-anonymity algorithm based on sensitive privacy protection [D]. Tianjin University of Finance and Economics, 2015.

[5] Xiong Shiqiang, He Daojing, Wang Zhendong, etc. Review of Federal Learning and its Security and Privacy Protection Studies [J / OL]. Computer Engineering, 1-17 [2023-12-19] https: / / doi.org/10.19678/j.issn.1000-3428.0067782.