Research and application of privacy protection technology based on big data environment

Zhiping Ding
Institute of Information Technology and Creative Design, Qingyuan Polytechnics, Qingyuan, 511510, Guangdong, China
dingzhiping@qypt.edu.cn

Abstract. With the development of the Internet of Things, the popularization of the mobile Internet, and the rapid promotion of social networks, the growth of data has entered a big explosion, and the development of information technology has caused a torrent of big data. The continuous development of technology brings people convenience, speed and comfort, but also hides hidden worries. Combined with differential privacy and clustering, a clustering-based differential privacy universal dataset release method for mixed datasets is proposed: using k-prototype clustering algorithm to group records in the hybrid dataset to reduce differential privacy Query sensitivity and the amount of noise to be added to improve data utility while providing data privacy protection; perform attribute difference calculations for numerical attributes and categorical attributes combined with weights, and measure their information loss separately. Finally, Experimental results show that this algorithm can improve the usability of data publishing.

Keywords: Privacy protection; big data; k-prototype; Clustering; Data release.

1. Introduction

The issue of personal privacy and security is an additional product that the development of Internet technology must bring, and the development of big data technology only further deepens this problem. The fundamental problem faced by personal privacy infringement in the era of big data is: a lot of personal information did not cause damage to others in the past, but big data technology will bring together a large amount of personal information for analysis and combination, which may damage the autonomy of others. Decisions and autonomy. At the same time, in the era of big data, it is difficult for us to define the relationship between big data services and privacy violations. On the one hand, the development of big data has been changing the connotation and extension of personal privacy. Many information that we consider to be private is actually not the real privacy. The intrusion brought about by technological development is gradually escalating, and traditional privacy is gradually invaded and separated from reality. The goal of protecting privacy in the era of big data is more to avoid the abuse of personal privacy information. The focus of privacy protection has gradually shifted from managing the collection and use of personal privacy information to emphasizing personal control and control over their data. This article discusses the application research of privacy protection technology in combination with the current mainstream big data applications.
2. Definition and characteristics of big data

The definition of big data can be represented by 4Vs characteristics, and there are typically two categories: 1) Big data definition of International Data Corporation (IDC): use type, speed, volume and value (variety, velocity, volume, value) Define big data. Among them: variety includes various types of data such as structured, semi-structured, and unstructured; velocity means that the collection and processing of big data must be fast and timely in order to maximize the value of big data; Volume means a large amount of data; value means that big data has great social value. From the perspective of data and technology, the big data architecture can be divided into two layers: big data life cycle and big data platform.

1) Big data life cycle

The big data life cycle includes collection, storage, use, distribution, and deletion. Data enters the big data platform through collection and storage, discovers its potential value through use, transmits and shares data or analysis results through distribution, and finally deletes data that is no longer needed. Therefore, it can be said that the big data life cycle is the process of converting data into value.

2) Big data platform

The big data platform provides the infrastructure, storage and processing platform, and data analysis algorithms required by each link of the big data life cycle, and is the technical support for the entire big data architecture.

3. Grammatical privacy

Grammatical privacy is usually operated in statistical databases. Data is usually published in the form of tables. The tables contain different attributes. Grammatical data protection technology is usually based on the assumption that the release of data may make certain individuals Privacy caused leakage. The first step to protect the privacy of these individuals is to eliminate or use random numbers to replace the identification attributes in the form before releasing the data. However, this simple method of removing the identification attribute cannot provide sufficient privacy guarantee for the individual. This is because the combination of the semi-identification attribute and the publicly released information can still identify the true identity of the individual. DeMontjoye and others collected 15-month mobile data sets of 1.5 million individuals, and performed simple anonymization of these data sets, such as excluding names, home addresses, telephone numbers, and ID numbers. At the same time, the author also collected additional data sets, including hourly time and space records of certain individuals. As a result, the author can identify individuals with 95% accuracy using only 4 spatio-temporal points. Furthermore, they studied the credit card transaction records of 1.1 million individuals for 3 months, and again found that 90% of individuals can still be identified through 4 spatiotemporal points. In the delivery, we cite a very simple example. Table 1 shows the data that is anonymized to protect patients' private information (such as Disease). In detail, we exclude the Name and ICN attribute values.

In order to protect privacy from being leaked, the grammar protection technology usually directly modifies the value of the semi-identification attribute of the initial data to ensure the privacy of the data. Privacy disclosure can be divided into 1) identity disclosure (record linkage), that is, the disclosure of the true identity of the individual; 2) attribute linkage, that is, the disclosure of sensitive information of the individual; 3) table linkage, that is, whether the individual The information in the data set was leaked. The grammatical privacy protection technology can be used in the big data collection stage and the big data release application stage. In the big data collection stage, the data producer can anonymize his own data and upload it to the data collector. In the stage of big data release and application, data collectors can use grammatical privacy protection technology to anonymize the data and then share it with other third parties.
4. Semantic Privacy
Semantic data protection technology is to protect the privacy of the data set whether or not the individual is publicly released. First, this article lists a simple example to illustrate. Now there is a data set that provides a query function, that is, you can query the annual tax and fee paid by people in a certain industry in a certain area. It is assumed that the annual tax payment of an individual is sensitive information and cannot be known by others. In this case, an attacker learned that the tax paid by Tom was 800 yuan lower than the average tax paid by civil servants in Shanghai. Although the attacker does not know any information about Tom’s taxes, but through the query function provided by the data set, the attacker can infer the taxes and fees Tom paid each year. What needs to be remembered is that Tom’s private information leakage does not depend on whether Tom is in the disclosed data set.

Semantic privacy protection technology is generally applicable to the following two environments:
Non-interactive environment: This environment means that there is no direct interaction between the data collector and the data application. The data collector can directly share and publish the data set, and the data collector does not limit the data application of the data application.
Interactive environment: This environment means that data collectors provide data-based query services and will not publish data publicly. This protection technology is usually used to ensure that the query results are not used by attackers to obtain information that needs to be kept confidential.
Sentence protection technology usually modifies the original data before publishing the data, while semantic protection technology usually adds noise to the actual published query results. Because these two technologies both anonymize and obscure the original data set, they have obtained privacy protection to a certain extent.

5. Key Technologies of Privacy Protection Research in Big Data Environment
Anonymization technology is a grammatical privacy protection technology that began to appear around 2000. Among them, K-Anonymity is the first method proposed, and subsequent anonymous protection methods such as l-Diversity and t-Closeness are derived privacy protection methods based on it. These methods have been applied in many fields and achieved good results.

1) K-Anonymity algorithm
Around 2000, Samarati and Sweeney introduced a method called K-Anonymity to protect data privacy. K-Anonymity requires that the records in the data set are indistinguishable from at least a certain number of records. Since it is usually assumed that the attacker uses semi-identifying attribute groups to identify someone’s identity, the basic idea of K-Anonymity is that the published data set must ensure that each equivalence class (a group of the same semi-identifying attribute value) has at least k groups the records are the same, which reduces the probability of being re-identified. In order to realize K-Anonymity, we made the following changes in sequence: 1) Change the "Age" attribute in the table from a value to an interval, such as changing age 22 and 25 to an interval [20,30]; 2) Change the "Job" attribute in the table to represent a wider range of occupations, such as changing Engineer and Lawyer to Professional; 3) Eliminate all the "Nationality" attributes in the table; 4) Change the "ZIP" attribute in the table to represent a wider range of occupations, such as changing Engineer and Lawyer to Professional; 5) Use K-Anonymity to anonymize data. Each equivalence class (Age, Gender, Job, ZIP) has at least two sets of records. In this case, the maximum probability that an attacker can re-identify an individual becomes 1/k. As the value of k is larger, the effect of privacy protection is better. However, the value of the data is also reduced after anonymization.

The typical operation to implement K-Anonymity is to generalize and suppress semi-identified attributes, while non-sensitive and sensitive attributes remain unchanged. The generalization operation refers to replacing the original value with a more general value. For example, the values of the "Job" attribute Engineer and Lawyer in Fig.1 have been changed to Professional. The suppression operation refers to deleting all or part of the attribute value, such as deleting all the "Nationality" attribute values in Fig.1. The combined use of generalization operations and suppression operations can minimize the number of generalizations required to satisfy K-Anonymity, so that more detailed information can be published to increase the value of data.
While using generalization and suppression operations to achieve K-Anonymity, as the detailed information of the data is reduced, the utility of the data is inevitably reduced. In order to maximize the utility of the data (i.e., minimize the loss of information), it is necessary to calculate a K-Anonymity to minimize generalization and suppression. This calculation of the optimal K-Anonymity data set is an NP-hard problem.

| Name | ICN | Age | Gender | Job   | Nationality | ZIP   | Disease |
|------|-----|-----|--------|-------|-------------|-------|---------|
|      | *   | 22  | F      | Engineer | German   | 10045 | Hepatitis |
|      | *   | 25  | F      | Lawyer | Italy     | 10046 | Stroke  |
|      | *   | 31  | M      | Lawyer | America   | 11027 | Flu     |
|      | *   | 38  | M      | Engineer | China   | 11220 | Hepatitis |
|      | *   | 41  | F      | Dancer | America   | 11245 | Stroke  |
|      | *   | 44  | F      | Dancer | America   | 11238 | Flu     |
|      | *   | 56  | M      | Writer | America   | 10041 | Flu     |
|      | *   | 56  | M      | Writer | English   | 10028 | Flu     |

Fig. 1 Original data set

| Name | Address | City     | ZIP   | Age | Sex | Job |
|------|---------|----------|-------|-----|-----|-----|
|      |         | Queen Ave| New York | 10041 | 56  | M   |
| John |         |          |       |     |     | Lawyer |

Fig. 2 External data obtained from public

2) l-Diversity algorithm

The K-Anonymity method is an effective solution to protect the privacy of individuals in the published data set. Generally speaking, the operation to implement the K-Anonymity method is to only operate on the semi-identified attributes in the data set, and do not modify or remove any sensitive attributes. This makes the K-Anonymity method unable to defend against certain specific attacks, namely attribute linkage attacks. In detail, a data set processed by the K-Anonymity method may still be inferred (or to reduce uncertainty) the individual's sensitive attribute value. This is because K-Anonymity does not hook the sensitive attributes in the equivalence class, which may lead to the disclosure of sensitive information that may be attacked.

The l-Diversity method requires that each equivalence class of the anonymized data contains at least one well represented sensitive attribute value. There are multiple definitions of representative values, the simplest of which is that these sensitive attribute values are different, that is, each equivalence class contains at least 1 different sensitive attribute values.

3) t-Closeness algorithm

Although the l-Diversity method guarantees the diversification of sensitive attributes in some special circumstances, this method provides the attacker with more private information. For example, for a given virus test result, the probability of being negative is 0.99. If the attacker knows that the target user is in the data set, we also know that his positive probability is 0.01. However, after the data is operated by the l-Diversity method, the probability of finding a positive result through the equivalence class reaches 0.5. In other words, the l-Diversity method provides more private information to the attacker under certain specific circumstances. This is mainly because although the l-Diversity method guarantees the diversification of sensitive attributes, it ignores the global distribution of sensitive attributes, and the attacker may confirm the sensitive value with a high probability.

The t-Closeness method overcomes the shortcomings of the K-Anonymity and l-Diversity methods, and requires the distribution of sensitive attribute values in all equivalence classes to be consistent with
the global distribution of the attribute. By sending seeds, the probability of a skew attack on the sensitive attribute value of a target user will be the same as the probability distribution in the entire data set. In other words, the value of the sensitive attribute of the target user will not change, and the attacker cannot obtain more private information from it. Similarly, t-Closeness can also reduce the efficiency of similar attacks. In the t-Closeness method, the distance between the frequency distribution of sensitive attribute values in the published table and each equivalence class needs to be calculated. There are different standards for calculating these distances, such as the Earth Mover distance.

Analysis of Key Issues of Privacy Protection Technology in Big Data Environment, In the big data environment, traditional privacy protection technologies for small data still have great limitations and face new technical challenges. Big data has the characteristics of large amount of data, various data types, fast data generation speed and low value density. In addition, personal privacy changes dynamically with many factors, making it even more difficult to protect privacy in the era of big data. Specifically, the main problems faced by privacy protection research in the big data environment are reflected in the following three aspects:

(1) Privacy protection cost sharing technology in the big data collection stage
In the big data environment, the real-time changes brought about by the massive scale of data and the rapid generation speed make it difficult for traditional passive privacy protection technologies to adapt. At the same time, the subjects who implement these privacy protection methods are usually data collectors, without mobilizing the subjective initiative of data generators, so that they rarely actively participate in the privacy protection system. In the data collection stage, the data producer uploads data related to himself to the data collector. For traditional privacy protection technology, it is often inefficient to only consider allowing a single data generator to protect its privacy. In order to save the cost of privacy protection and the effect of protection, this article will discuss data generators working together to protect their privacy, that is, privacy protection cost sharing technology. It should be noted that no protection technology is limited to certain specific scenarios. For example, data producers such as location-based services can anonymize the data themselves without affecting or slightly affecting the services they enjoy. For other scenarios, such as when the patient is presenting his condition to the doctor, the patient must be unable to hide his condition to protect his privacy.

(2) Interrelated secret protection technology in the application stage of big data
Compared with the traditional privacy protection technology, the privacy protection technology in the big data environment is more complicated. The integration and fusion and correlation analysis of the multi-source data in the big data environment greatly increase the degree of data privacy risk. Traditional privacy technology only specifies the corresponding anonymization strategy for the attack hypothesis on a single data set. However, the large-scale and diversity of big data make traditional privacy protection technologies lose sight of one another. Under this challenge, when considering their own privacy mechanisms and privacy parameters, data collectors need to pay attention not only to the utility of the data, but also to the impact of interrelated data on him. This turns the traditional privacy and utility balance problem into a game problem, that is, each data collector maximizes data utility under certain privacy constraints. In this case, a new definition of privacy needs to be designed to evaluate the degree of privacy under interrelated data sets.

(3) Big data privacy risk control technology based on network insurance
Although data producers and collectors use various methods to protect privacy, privacy leaks still occur from time to time. Therefore, a method is needed to mitigate the loss caused by privacy leakage. This article introduces network insurance as a means of risk control to realize proper remedial measures for big data privacy leakage. Nevertheless, the application of cyber insurance to privacy risk control also faces some new problems: (1) The addition of cyber insurance increases the burden on data generators and collectors, which makes them not necessarily willing to use cyber insurance; (2) Network insurance reduces the loss of privacy leakage of data producers and collectors, so that data producers and collectors may reduce their degree of privacy protection to improve their effectiveness. (3) Network insurance may also enable data generators and clever collectors to perform some illegal actions to obtain additional
benefits. Therefore, the widespread application of network insurance needs to solve the above-mentioned problems.

6. Conclusion

In the era of big data, the explosive growth of data has brought unprecedented opportunities and challenges to human society. On the one hand, data is a company’s valuable assets, important economic investment and the cornerstone of new business models. Many companies and organizations collect, process, analyze, use, publish or share data based on the huge value of data. On the other hand, in the process that big data is widely used, the problem that big data technology cannot avoid is the risk of data privacy leakage. Since data often contains information about individuals or groups, although some data cannot directly represent individuals or groups on the surface, it can still be traced back to individuals or groups through attack techniques. Big data has the characteristics of large amount of data, various types of data, fast data generation speed and low value density. In addition, personal privacy changes dynamically with many factors, making it even more difficult to protect privacy in the era of big data. Therefore, this article has carried out relevant research work on the problems and technical requirements of privacy protection at different stages in the current big data environment, and proposed corresponding solutions, which have certain important theoretical and application values.

References

[1] X. Wu, X. Zhu, G. Wu, W. Ding. Data mining with big data, IEEE Transactions on Knowledge and Data Engineering, 2018.
[2] X. Xiao and Y. Tao. Personalized privacy preservation, in Proceedings of SIGMOD, 2016.
[3] M.E. Nergiz, C. Clifton, A.E. Nergiz. Multirelational K-Anonymity, in Proceedings of ICDE, 2017.