Big Data Security and Privacy Protection

Wentao Zuo

Guangzhou College of Technology and Business, Guangzhou, Guangdong, 528138, China

Keywords: Big Data, Privacy Protection, Internet Security Technology

Abstract: Big data is a research hotspot in the current academic and industrial circles, which is affecting people's daily life style, work habits and thinking patterns. However, at present, big data faces many security risks in the process of collection, storage and use. The privacy leakage caused by big data is seriously plagued by users, and false data will lead to incorrect or invalid big data analysis results. This paper analyzes the technical challenges of implementing big data security and privacy protection, and sorts out some key technologies and their latest developments. The analysis points out that big data is an effective means to solve information security problems while introducing security issues. It brings new opportunities for the development of information security.

1. Introduction

With the extensive use of data by various industries, big data has become a major symbol in the field of information technology after the mobile Internet, cloud computing, and the Internet of Things. Due to the formation of such a large and complex data system, people's analysis and in-depth study of data information has become less convenient. To handle and manage such complex data systems requires more comprehensive security and privacy protection technologies, but now people are facing the growing information security and privacy issues of big data. This is a major challenge that requires the entire information technology industry to focus on and actively seek solutions.

2. Sources and characteristics of big data

Big data refers to a large and complex collection of data that is difficult to process using current database management tools or data processing methods. Data sources can be divided into: (1) various data information such as pictures, texts, and audio that people voluntarily issue on the Internet; (2) various types of logs and files generated by machines and stored in computers. , database, media materials, etc.; (3) item attribute class, device record data, such as various product information recorded in the warehouse, data calculated in astronomical glasses, and so on. Characteristics of big data: (1) Scale – As mentioned above, big data is large and complex. According to statistics, the total amount of information in the world in 2012 has been 2.7ZB, and it is expected to increase to 8ZB in 2015; 2) Diversity-In the past, in order to facilitate storage and viewing, data is mostly structured data based on text. Now, due to the diversification of information carriers, non-structures containing information such as pictures and audio are made. More and more data; (3) value - through the analysis and statistics of the overall data, extracting valuable parts for users to use is also one of the basic characteristics of big data; (4) high speed - -- In the era of information explosion, there is a growing need for efficient processing of information and the provision of real-time information.

Big data analysis is mostly used in different fields such as science, medicine, and commerce, and its uses are very different, but their goal of analyzing data is no more than three. (1) In order to obtain valuable information, the original data containing a large amount of information is analyzed and integrated at different angles, and finally the more excellent information is summarized to help people understand the essence of things, grasp the development and operation of things, and then The next step in the development of things and things to make a predictive response. For example, in the field of fashion sales, the marketing department can understand the consumer's consumption

DOI: 10.25236/csam.2019.012

trend and demand by analyzing the consumer's consumption data, and can produce more market-oriented products in advance to satisfy consumers. (2) By analyzing the accumulated data in multiple dimensions, not only can people grasp the general group characteristics, but also can specifically describe the differences between different individuals. Enterprises can use this data to introduce more humanized services to customers. For example, Amazon analyzes behaviors (ie, search, browse, join shopping carts, and purchases) before users purchase items, and can understand the user's purchase goals and psychological activities at the time of purchase, and implement effective recommendations. (3) Under the condition that information can be quickly transmitted through the network, it is more necessary to analyze the authenticity of the data information. The information provided by the erroneous data may cause the user to make an incorrect decision, and sometimes cause irreparable errors. Therefore, it is necessary to conduct a detailed and in-depth analysis of the data, to take its brilliance and to ruin it. For example, filtering spam in a mailbox can also use big data analysis technology to protect users from interference.

3. The security test facing big data

Judging from the leakage of user information in recent years, the leakage of user privacy has caused great trouble to users. According to the privacy content that needs to be protected, privacy protection can be divided into: protection of unknown privacy, anonymous protection of identifiers, and anonymous protection of connection relationships. But in fact, apart from the disclosure of user privacy, there are still some problems. Some enterprises predict the behavior and life status of users through big data analysis, and then grasp the user's living habits, hobbies, consumption records, etc., and recommend advertisements to users. Wait. Many companies today simply perform anonymous anonymity on user privacy, thinking that as long as the public information does not contain a user identifier, the user's privacy can be well protected, but it is not. At present, the privacy protection of users in the process of collecting, storing, managing and applying user information mainly depends on the self-discipline of enterprises, and there is a lack of corresponding supervision standards and regulations. Users have the right to know how their data information is used and used in business activities.

In big data, there is a lot of data that is confusing or false. If you don't judge it carefully, it will be deceived by data. There are two reasons for this kind of data: First, the data itself is fake, or someone makes up or hears for a certain purpose, and the wind comes from the wind; Second, the data is distorted, due to the operator's work mistakes in the process of data collection. The difference between the collected information and the real information affects the final result of the data analysis. It may also be that during the process of communication, the information changes cause the data to not reflect the real situation. For example, a restaurant's ordering phone has been changed for some reason, but the original number has already been included in the database, so the number that the user sees after searching is inconsistent with the actual number. Therefore, in order to improve the credibility of the data, the data user should understand the source of the data, the route of transmission, and the process of data processing, etc., to prevent invalid conclusions.

4. Big Data Security and Privacy Protection Technology

In terms of structured data, to effectively implement user data security and privacy protection, data publishing anonymous protection technology is a key point, but this technology needs to be continuously explored and improved. Most of the existing data publishes the basic theory of anonymous protection technology. The setting environment is mostly that users publish data once and statically. If the identifiers are grouped by tuple generalization and suppression processing, the collection of common attributes is anonymized using the k anonymous mode, but it is easy to miss a particular attribute. But in general, the reality is changeable, and data is generally published continuously and repeatedly. In the complex environment of big data, it is more difficult to implement data distribution anonymous protection technology. An attacker can obtain various types of information from different publishing points and different channels to help them determine a user's information. This also requires researchers in the information field to invest more energy and research.

Unstructured data containing a large amount of user privacy is mostly generated in social networks. The most prominent feature of such data is the graph structure, so data distribution protection technology cannot meet the security privacy protection requirements of such data. Generally, attackers will use the relevant attributes of points and edges to re-identify the user's identity information through analysis and integration. Therefore, in the social network to implement data security and privacy protection technology, it is necessary to combine the characteristics of its graph structure, user identity anonymity and attribute anonymity (point anonymity), that is, hide the user identification and attribute information during data publishing; Anonymity (anonymity) of relationships between users, that is, data publishing is hidden from the relationship between users. This is the main point of social network data security and privacy protection. It can prevent an attacker from cracking anonymous protection by inferring data published by users in different channels, or by tying between users to infer users who were originally protected by anonymity. Or in the complete graph structure, the super node is used to perform partial segmentation and re-aggregation of the graph structure, so that the side anonymity can be realized, but this method can reduce the availability of data information.

Watermarking technology refers to embedding identifiable information into a data carrier in some unobtrusive manner without affecting the content of the data and the use of the data. Generally used in media copyright protection, there are also some database and text files that apply watermarking technology. However, the application of watermarking technology on the multimedia carrier and the database or text document is very different. The characteristics of the data based on the disorder and dynamics of the two are not consistent. Data watermarking technology can be divided into strong watermarks, which are used to prove the origin of data and protect the original author's creative rights. Fragile watermarks can be used to prove the authenticity of data. But watermarking technology is not suiTable for big data that is now mass-produced quickly. This is something that needs improvement.

Research on data traceability technology began in the database field and is now being introduced into big data security and privacy protection. Marking the source data can shorten the time when the user judges the authenticity of the information, or help the user to verify whether the analysis result is correct or not. The marking method is the most basic method in the data traceability technology, mainly the calculation method (Why) and the data source (Where) of the recorded data. Data traceability technology also plays a significant role in the traceability and recovery of files.

5. Conclusion

When the era of big data came, it brought opportunities for technological development, but also brought new problems and challenges. Big data security and privacy protection are just some of the problems that need to be solved. Through the specific research and technology mining of the status quo of big data security and privacy protection, this paper discusses the key technologies that can solve the existing information security and privacy protection problems, such as anonymous technology, watermark technology and traceability technology. Of course, these are not just these. Realize big data security and privacy protection. At the same time, we must master some national policies to provide a good environment for the development and application of related technologies, so that big data can better promote the development of human society information technology.

References

[1] Feng Dengguo, Zhang Min, Li Wei. Big Data Security and Privacy Protection [J]. Computing Machine, 2014(1).

[2] Ge Yueying. Information Security in the Age of Big Data and Protection of Citizens' Personal Privacy [J]. China Information Industry, 2014(1).

[3] Xie Bangchang, Jiang Yefei. How to protect privacy in the era of big data [J]. China Statistics, 2013 (6)

[4] Hou Fuqiang. Personal Information Protection Issues and Legal Countermeasures in the Age of Big Data [J]. Journal of Southwest University for Nationalities, 2015(6):106-110.

[5] Li Baihan, Wang Yuzhu. Data Security in Big Data Cloud Computing Environment [J]. Electronic Technology and Software Engineering, 2017(15): 213.

[6] Lu Yifeng, Wang Zugang. Characteristics, Problems and Management Strategies of Big Data Application--Based on the Perspective of Internet Financial Stability [J]. Journal of Changchun Finance College, 2017(4): 5-10, 27.