The Key Technology Research on Privacy Protection Based on Big Data

Xueguo Li [a], Yinglan Shen [b]

Chong Qing Creation Vocational College, ChongQing, 402160, China;

[a]Lixueguo@126.com, [b]1121294519@qq.com

Keywords: Big data, Big data security, Privacy protection

Abstract. Big data is a research focus in the academic and industrial circles currently, which is affecting people's daily life-style, work habits and thinking modes. But at present, the big data is confronted with many security risks in the collection, storage and use processes, privacy disclosure caused by data bring serious problems to the users, false data will lead to wrong or invalid analysis results of big data. Technical challenges that the realization of security and privacy protection of big data faces are analyzed in this paper, and several key technologies and its latest progress are compiled. This article points out when the big data introducing safety problems, it is the effective means to solve the problem of information security as well.

1. Introduction

Nowadays, the development of social informatization and network bring about explosive growth of data. According to the statistics, average 2 million users use Google search per second, what Facebook users share are more than 4 billion every day, the number of twitter that Twitter process is more than 340 million every day. At the same time, a large number of data are generated in scientific computation, medical health, finance, retail industry and so on, Total amounts of global information has reached 7.5 ZB in 2015, and this figure is expected to reach 9.2 ZB by 2016.

Turing Award winner JimGray put forwards the fourth paradigm of scientific research in academia, namely, data intensive scientific research on the basis of big data; "Nature" launched a special release on big data edition to discuss in 2008; the IT industries act more positively, data reuses are continued to be focused on, and the potential values of big data are dug. At present, the big data has become another growth point of information industry in the field of information technology after the cloud computing. According to prediction of Gartner, big data will drive global IT spending $75 billion in 2015, the Gartner list "big data" technology as strategic one of the top ten technologies and trends in many companies and organizations. Governments also are the main drivers of big data technology promotion, 35 countries and regions have formally built their own data open portal website all over the world up to now. In our country, China Institute of Communications, China Computer Federation and other important academic organization successively set up big data expert committee in 2012, and provide academic consulting for application and development of big data in China.

The development of the big data is still confronted with many problems currently, the privacy problem is one of the key problems that people generally acknowledge [1-2]. Nowadays, people's every word and action are in the hands of merchants on the Internet, including shopping habits, friends' contact situation, reading habits, retrieval habits and so on. A number of actual cases show that even harmless data are largely collected; it will be exposed personal privacy [1]. In fact, the secure meaning of big data is broader, the threat that people are faced with is not limited to personal privacy disclosure. The big data is similar to other information; which is confronted with many security risks in storage, processing, transmission process; and big data have data security and privacy requirements. To achieve security and privacy protection of big data is more difficult, compared with the past other safety problems (such as the data security in cloud computing). This is because in cloud computing, although service providers control the data storage and operation environment, but the users still have some means to protect their data, for instance, cryptography technology are used to realize storage and safety computing of data, or safe operating environment can be realized by way of trusted computing way, etc., Facebook and other merchants are not only the producers of the data, they are also data storage, managers and users under the background of big data,

therefore, it is extremely difficult to realize the users' privacy protection by limiting merchants use users' information simply[1].

2. Research Summaries of Big Data

2.1 Sources and Prosperities of Big Data

It is universally thought that sources and properties of big data; big data are the large scale and complex; so that it is data set that is difficult to use the existing database management tools or data processing application to process. Common prosperities of big data include volume, velocity and variety.

According to different sources, the big data can be roughly divided into the following categories [3]:

(1) From the people. All kinds of data that people produce in the process of Internet activity and in using mobile Internet, including texts, images, videos and other information;

(2) From the machine. Data generated by all kinds of computer information system, which exist in the form of documents, databases, multimedia and so on, including auditing, log, and other automatically generated information;

(3) From the material. Data collected by all kinds of digital devices, such as digital signals generated by camera, people' various eigenvalues in medical Internet of things, a large amount of data generated by astronomical telescope, etc.

2.2 Technological Architecture of Big Data

Big data processing involves collection, management, analysis and display of data, etc. Fig. 1 is a related technical sketch

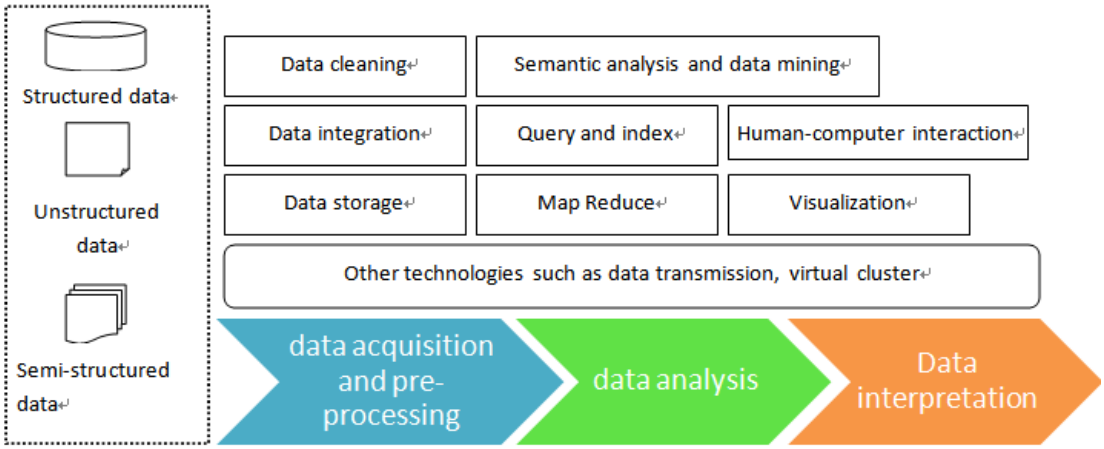

Fig.1 architecture of big data

(1) Acquisition and preprocessing of data

The data sources of big data are diversified, including database, text, image, video, webpage, and various structured, unstructured, semi-structured data. Therefore, the first step of big data processing is acquired from data source and makes preprocessing operations, and provides unified and high-quality data sets for the subsequent processing.

Because the sources of big data differ, there may be descriptions with different modes, even contradictions exist. Therefore, the data are made cleaning in the process of data integration; it is indispensable to eliminate similar, repetitive or inconsistent data. According to the properties of the big data, cleaning of unstructured or semi-structured data and data integration with very large scale can be used.

(2) Data analysis

Data analysis is the core process of the big data application. According to different levels, data analysis can be roughly divided into three categories: computing architecture, query and index, data analysis and processing.

MapReduce is widely used big data set computing model and framework in computing architecture aspect. In order to adapt some analysis requirements that the task completion time

requires higher demand, the performance can be optimized, a data flow analysis solution plan based on MapReduce architecture has been proposed.

In term of query and index, because the big data contains a large amount of unstructured or semi-structured data, the query and index technologies of traditional relational database are restricted, and NOSQL type database technologies get more attention. A hybrid data access architecture HyDB can be adopted and a concurrent data query and optimization method is made operation.

The mainly relevant technologies include semantic analysis and data mining and so on in data analysis and processing aspects. Because data present diversified properties in the big data environment, so when the data is in the semantic analysis, it is harder to unify terminology and dig information. In view of the big data environment, a kind of efficient terminology standardization method that can solve the terminological variations problem has been put forward [2]. The traditional data mining technologies are mainly aimed at structured data, so it is imperative that unstructured or semi-structured data mining technologies need studying. A mining technology for image documents has been proposed [3].

3. Security Challenges Presented by Big Data

Science and technology is a double-edged sword. Security problems caused by the big data and its values are equally striking. But the recently eruptible "prism" event more added people's concern about big data security. Compared with the traditional information security problems, the challenging problems big data security faces mainly reflected in the following aspects.

3.1 Users' Privacy Protection in Big Data

A large number of facts show that big data is not properly handle, which will cause great violations to users' privacy. According to difference of contents need protecting, privacy protection can be further subdivided into location privacy protection, identifier anonymity protection, connection anonymity protection, etc.

The threat people are confronted with is not limited to personal privacy disclosure and the prediction for the people' status and behavior based on big data. A typical example is that a retailer is earlier than the parents know the fact that his daughter is pregnant through historical analysis, and mail related advertising information. The social network analysis and research also shows that the user's attributes can be found by the properties of these groups. For instance, through the analysis of users' Twitter information, the user's political leanings, consumption habits and favorite team can be found.

3.2 How to Realize the Access Control of Big Data

Access control is the effective means to realize controlled sharing of data. Because the big data could be used in a variety of different scenarios, the access control demand is very prominent.

The properties and the difficulties of the access control of big data lie in:

(1) It is difficult to preset role and realize the role division. Because the application scope of big data is widespread, it is usually accessed by visitors from different organizations or departments, identification and purpose, the implementation of access control is the basic demand. However, under the scenarios of big data, there are a large number of users which need to implement rights management, and users' specific rights are unknown. In the face of unknown large amounts of data and users, it is very difficult to preset role.

(2) It is difficult to predict the actual permissions of each role. Because the big data scenes contain huge amounts of data, the security administrator may lack enough professional knowledge and cannot accurately specify accessible data scope for the users. And from the point of view of efficiency, it is also not the ideal way to define users' all authorization rules. Taking medical applications for example, doctors may need access a lot of information in order to do their job, but data access should be determined by the doctor, it should not require administrator to do special configuration for each doctor. But at the same time, the detection and control of doctors' access behavior should be able to provide, and limit doctor excessively access patients' data.

In addition, diversified access control requirements may exist in big data with different types. For example, in web 2.0 personal user data, access control based on historical records exist; i there are

access control requirements based on the scale and precision of data in the geographical map data; there are access control requirements of data time interval in stream data processing and so on. How to uniformly describe and express the access control requirements is a challenging problem as well.

4. The Security and Privacy Protection Key Technology of Big Data

The users' privacy protection, data content trusted authentication, access control and other security challenges that the big data face need launching security key technologies of big data currently.

4.1 Anonymous Protection Technology of Data Release

The anonymous protection of data release is key core technology and basic means to realize its privacy protection in terms of structured data (or called relational data) in the big data, which is still in the stage of continuous development and improvement currently. The typical K anonymous scheme is as an example. The early plan and its optimized plan through the tuple generalization, suppression and other data processing, the quasi-identifier are grouped. The quasi-identifier are similar in each group and contains K number tuple at least, thus each tuple is indistinguishable with other tuples of K-1 number at least. Because K anonymous model is aimed at all attribute collections, it is not defined certain specific attribute, it is prone for certain property to treat insufficiently. If sensitive attribute value in an equivalence class is consistent, then attacker can effectively determine the attribute values. Researchers have proposed 1 diversity anonymity for the problem. Its properties are that the diversity content of sensitive data in each anonymous attribute group are greater than or equal to l. The implementation methods include scheme based on cutting algorithm and plan based on the data exchange, etc. In addition, there are some plans between l anonymity and 1 diversification. Further, because l-diversity just makes occurrences frequency of sensitive data average. When the data in the same equivalence class scope is very small, the attacker can guess its value. T closeness plan require that distribution of sensitive data in the equivalence class is consistent with the distribution of data in the whole table. Other works includes (k, e) anonymous model, (X, Y) anonymous model, etc. The above research aims at static and one-time release condition. But in reality, data release often face scenes that data released continuously and several times. It is necessary to prevent the attacker to analyze multiple released data union, destroy original anonymity of data.

4.2 Anonymous Protection Technology of Social Network

The data generated by the social network is one of the important sources of big data, at the same time, these data contains a large number of users' privacy data. The user’s members of Facebook have reached 1.55 billion by December 2015. Because the social network have graph structure properties, the anonymous protection technologies are quite different from structured data. The typical anonymous protection requirements in social network are user identity and attributes anonymity (also called peer anonymity), the user's identity and attribute information are hided in the release of the data; and the relationships among users’ are anonymous(also called edge anonymity), when the data are released, the relationship among users are hided. The attacker tried to use various properties (degree, tag, some specific connection information, etc.) of node, identify the identity information of the node in the graph again. The edge anonymous schemes are mostly based on addition or deletion of the edge currently. Random addition or deletion exchange method can effectively realize edge anonymity.

Another important way of thinking is division and integration of graph structure based on super node. For instance, anonymous scheme based on node integration, implementation scheme based on genetic algorithm, realization scheme based on simulated annealing algorithm and super node scheme that first filled, then divided. Although anonymous scheme based on super node can realize the anonymity of edge, but it is quite different from the original social structure figure, and at the cost of the sacrificing availability of data.

4.3 Watermark Technology of Data

Digital watermark is a method that the identity information is embedded within the data carrier in imperceptible way and does not affect its use; it is seen more in copyright protection of multimedia data. There are some watermark schemes for the database and text document.

The method that watermark is added in the database documents is quite different from multimedia carrier, which is determined disorder and dynamic properties of data. Its basic premise is that the

redundant information or can tolerate a certain precision error exist in the above data. For example, Agrawal et al., based on error tolerance range exist in numeric data in the database, a small amount of watermark information are embedded randomly selected least important place in these data. While Sion put forward a kind of plan based on the statistical property of the data set, one bit watermark information is embedded in a set of attribute data, prevent attacker from damaging watermark. In addition, the fingerprint information in database are embedding in watermark, which can identify the owner distributed object of the information and, it is conducive to tracking leakers under distributed environment, watermark public authentication without key can be achieved by using independent component analysis (ICA) [4].

There are many various generating method of text watermark, which can be roughly divided into fine-tuned watermark based on the document structure, dependent on the tiny differences between character spacing and row spacing formats and so on; watermark based on text content [5], dependent on the modified document content, for instance, the space are increased, punctuation are modified, etc.; and watermark based on natural language, the change can be achieved by understanding semantics, for example, the synonym for replacement or sentence changes, etc.

4.4 Traceability Technology of Data

As previously mentioned, data integration is one of the early stage processing steps in the big data. Due to the diversification of data source, it is necessary to record the source, transmission and computing process of data, and provide auxiliary support for mining and decision-making later.

As early as before the emergence of big data concept, data traceability technology has been widely studied in the field of database. The basic starting point is to help people to determine the all the data source in data warehouse, such as understanding what data items and which table they are operated by, therefore, the correctness of the result can easily check, or update the data with minimal cost. The basic method of data traceability is notation, for instance, in [6] the data query and communication histories in data warehouse are recorded by marking data. The subsequent concepts are further refined and Why and Where- two kinds, they are focused on computation method of the data and the data source, respectively. In addition to the database, it also includes the traceability technology of XML data, streaming data and uncertain data. Data traceability technology can also be used to trace and recover document.

5. Conclusion

Big data brings new security issues, but itself is also the important means to solve the problem. The relevant key technologies of security and privacy protection of big data are combed from the privacy protection, trust, access control of big data in this paper currently. But overall, the domestic and foreign related researches that aimed at the security and privacy protection of big data are not enough currently. The security and privacy protection problem of big data can be better solved only through technological means and related policies and regulations.

References

[1] Viktor Mayer-Schonberger, Kenneth Cukier. Big Data: A Revolution that Will Transform How We Live, Work and Think.Bost on: Houghton Mifflin Harcourt,2013

[2] Li Guo-Jie, Cheng Xue-Qi.Research status and scientific thinking of big data. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647-657(in Chinese)

[3] Dean Jeffrey, Ghemawat Sanjay. MapReduce: Simplified data processing on larg clusters//Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation. San Francisco, USA, 2004: 107-113.

[4] Kang U, Chau D H, Faloutsos C. Pegasus: Mining billion-scale graphs in the cloud//Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto, Japan, 2012: 5341-5344

[5] Jiang Chuan-Xian, Sun Xing-Ming, Yi Ye-Qing, Yang Heng-Fu. Study of database public watermarking based on JADE algorithm. Journal of System Simulation, 2006, 18(7): 1781-1785(in Chinse)

[6] Pease A, Niles I, Li J. The suggested upper merged ontology: A large ontology for the semantic web and its applications//Proceeding of the AAAI-2002 Workshop on Ontologies' and the Semantic web. Edmonton, Canada, 2002: 1-4