Research on Medical Big Data Application Model Based on Data Separation

Hai-feng Li* and Hong-rui SUN
Dalian Jiaotong University, Liaoning, Dalian, China
*Corresponding author

Keywords: Medical big data, Big data security, Data separation.

Abstract. Data attributed data, privacy attributed privacy. At present, big data development now going all walks of life, in order to better carry out medical research and so on, the medical industry select development of healthcare big data, but the data are still some big problems in terms of security: user privacy information disclosure and other. This paper proposes a new way of thinking: data separation storage method, by the user's information data into information data block medical data and personal information data blocks, try a way to solve the existing problem of private data leakage, and study the practical application ability of medical big data based on this data separation storage method.

Introduction

Big Data is the information explosion in recent years in the development of the status quo from the next-generation technology, through these years of development, its advantages of being continued to be confirmed, in this context, all walks of life to seek their own for the better development, competing with big data technology. However, in studies of big data, the security issue is one of concern focus. Data leakage or damage will cause considerable losses to the producers and users of the data. Therefore, it is particularly important to focus on the development of data security.

At present, around Big Data security or medical applications of this issue, the relevant domestic and foreign scholars, enterprises subject and government level in the ongoing efforts. Overseas: For example, the US federal, state, and other organizations are required to comply with strict confidentiality, integrity, and availability rules when collecting and processing data [1]; in order to reduce the impact of data mining on privacy leakage, Agrawal proposed privacy in 2000 Protected data mining technology PPDM (Privacy Preserving Data Mining). The methods that PPDM have developed over more than a decade can be basically divided into based on homomorphic encryption, OT (Oblivious), and Secure Multi-party Computation (SMC) [2]. Three types of methods; and domestic: Back in December 2009, the Ministry of health formulated the "basic norms of electronic medical records (trial)", the current electronic medical records have been widely promoted, the use of more detailed standard format is gradually updated [3,4]. In June 2016, the "Guiding Opinions on Promoting and Regulating the Development of Health and Medical Big Data Applications", drafted by the former National Health and Family Planning Commission, proposed to promote the sharing and openness of health and medical big data resources, and promote governance, clinical and Scientific research and public health big data applications [5].

In summary, this article proposes a method related to the security of big data. This method may solve the security problems of user privacy leakage. In view of the fact that this method is implemented before data is collected, it can largely protect user information security, although it is powerless for corporate information theft, but it can at least provide some small help for security research. This solution aims to protect the privacy of user's data. Therefore, this article will specifically study the feasibility of this method in combination with medical big data, and analyze its specific methods, use effects, and implementation possibilities.
Related Theoretical Definitions

Healthcare Big Data
Medical big data is an important application branch of big data onto the medical field. It refers to data related to individual life and health generated in activities related to human health. According to its data sources, medical big data includes clinical data, medical knowledge data, experimental examination data, medical imaging data, genetic data and public health data [6].

Large Data Distributed Storage
The essence of big data storage is to store a large amount of data onto hundreds or thousands of machine nodes in a distributed manner [7]. With the continuous expansion of data scale, traditional relational databases cannot meet the needs of large data storage, but rely more on distributed databases and non-relational databases. Distribute database system is a plurality of computer nodes connection to the database system via a network. Although each of the distributed database node physical address is relatively scattered, but logically associated with the use of time can be regarded as a complete database. Abstract distributed database system 4 layer structure model: a global layer, the concept of global layer, local layer, and the concept of local inner layers of between the corresponding inter-layer mapping (refer with: Fig. 1) [8], data is stored separation methods presented herein are the inner layer established in the local database.

![Distributed database system 4 layers structure model.](image)

Analysis of Medical Big Data Security Issues
Through literature review, in Comprehensive data security issues and large medical data analysis application basis points on the pain, the medical large data security aspects of the existing problems are summarized in the following four aspects: data security, access security, data credibility or data deceive, user privacy disclosure.

Data Security Issues
Data security is more about the security of data itself, such as the timeliness of data backup, data storage strategies such as stability, whether data is stored in a local or a third-party cloud services platform, data security has always been a problem. At present, medical big data is mostly maintained by itself. For example, HIS data and clinical data are stored locally in medical institutions, which places higher requirements on the safety of data backup in medical institutions.

Access Security Aspects
Access security more emphasis on security in data transmission processes in the network, although the data encryption algorithm continues to the good evolution, however, access security is a growing
problem. Whether the data is stored locally or hosted in a third-party cloud, whether it's data from the medical industry or data from other industries, good access security are equally important.

**Data Credibility or Data Deceive**

The beginning of large data definition itself means data non-precision, based on this analysis of the data may make misleading result of the judgment, it will produce data to deceive. The threat from medical big data credibility mainly comes from two aspects: at the subjective level, individual patients are sometimes unwilling to over-disclose some of their true information. This will cause the data to have errors when collected at the source; objectively, such as the scribbled handwriting of doctors' medical records and inconsistent recording standards, which will cause errors in the stored data such as electronic medical records, resulting in data credibility decrease or data spoof occurs.

**User Privacy Leak**

User privacy is the most closely related to the interests in data owners and users, and it is the security problem that needs to be addressed most. This will also be the focus on this article based on the data separation method. Medical large data pointing to the relevant data onto personal medical services, a record of a single individual's medical record, medical insurance, and other relevant information related to personal and more private data area, so the event of a leak or illegal use of criminals will cause property, information and other losses to the owner or user.

**Data Separation Storage Analysis and Application Strategy**

**Data Separate Storage Related Definitions**

The point of this chapter is based on the big data distributed storage strategy, which roughly divides the generated data into two major parts: personal blocks (personal privacy data) and data blocks (non-personal block data), the two are only associated locally. At the level of big data, will be observed only to each and every one individual data block. In the method, the interaction between the two block is connected by the terminal's identification. This identification can be a status code set by the terminal itself, or some complex password locks. The access to the personal block is only this identification which is used to ensure the security of user privacy data onto the big data level.

With large medical data, for example, the user's personal information such as home address, name and other privacy strong message as a personal block database; in a state of a user signs of disease history and other information for the data block databases, the two databases, there is an association relationship within each of them, and the unique key connecting the two databases may be the user terminal identity. So that in the data mining process, it plays a role in protecting users' privacy data when obtaining user information data, as well as to form information partition between the large information data analysis and user privacy.

**Definition of Information Partition and Its Effect Analysis**

By separating the data storing method, when companies or organizations to collect data acquisition only operations required data, thereby isolating each privacy data and information data, so that each other to produce Information partition. So, can this kind of information partitioning really to make big data analysis judge the user's privacy data?

Big data analysis mostly analyzes the relationship between data by machine learning data mining algorithms for massive data. Through to establish statistics model, Deviation will be generated and gradually accumulated, affecting the statistical results. However, this deviation accumulated largely to affect the amount of data being affected by the size and the parameter information, so that the large data analysis will be easier to identify these deviations. According to effective statistics, when the data sample size is greater than the amount of 4000, the deviation accumulation will show an exponential rise (refer with: Fig. 2).
Meanwhile, when the data size is gradually increased, The underlying laws between the data are becoming clear, in order to identify certain corresponding relationship and sublimation of non-relation into relation, at this time, coupled with the data positioned with the user data calculation, you can understand that the data belongs, and the user's behavior pattern will be predicted by the specific attribution. Therefore, if the information is cut off, big data can only be a kind of feedback only approximate scale trends, and cannot be traced back to the personal information that is not included - personal information will appear too small in the wave of big data.

![Deviation changes with the increase of sample.](image)

**Figure 2. Deviation changes with the increase of sample.**

**Related Advantages and Disadvantages of Separate Data Storage Applications Analysis**

The main role of separate data storage method is to protect the user's private data its effect is designed to let the user's private data on the network "does not exist", in which when private information data become the necessary information on the server side should be provided to a limited extent with the permission of the agreement, so that the information isolation between the private data and the operational data can be achieved. However, after isolating a large amount of private data, the data separation storage method may produce the following problems:

1. Is there a loss of efficiency for data storage?
   For big data, massive data cannot and should not be stored on a single terminal. For data separation and storage, user generated information data and privacy data are already stored on the terminal, so this method will not have much impact on the storage of big data. Assume that the status of a certain physical sign is monitored. The user's local terminal device records the user's data one by one, and requests the server for corresponding suggestions according to internal data instructions when the user queries. When data is copied, the data is directly copied and uploaded according to the big data mapping algorithm. In this process, user data onto the personal block will not be accessed.

2. The question of whether data mining will produce efficiency loss
   For data mining efficiency loss issues, mainly how to use data and business organizations are closely linked, for absolutely rely on user data to personalize the recommended business enterprise, the data separation technology of data mining would be greatly reduced, does not apply to the use of data separation technology; however, for large medical data, the impact is not great, large medical institutions based on medical data for selected clinical pathways, disease judge, drug selection decisions, all the above model training do not rely on the user's personal information data block. During the entire data mining process, the user's privacy data is rarely analyzed interactively, only when the model is initialized and the result is output, it will interact.

**Relevant Strategies for Data Separation and Storage in Practical Applications**

The implementation possibility of the data separation storage method is discussed above, then we should think about its advantages and disadvantages in specific application.
(1) On the basis of separate storage-based data, we can establish a security zone in the storage division of the privacy of personal data block, if and only if the server with the correct use of identity when we can call these data in order to achieve user privacy information access operation. Specifically, the user's private data can be stored in the local security zone so that all applications can interoperate access to these data, during the network, server needs to has a formal verification "identity" in order to ensure the safe interaction of private data.

(2) The separation of private data can also effectively prevent data leakage caused by machine learning results. Machine learning and artificial intelligence technology can more and more learn to iterate by continuously collecting data. However, not every piece of data is true. When false information exceeds a certain ratio, is likely to lead to machine learning to make erroneous judgment, then the separated data may be provided for this case at least a chance of buffer layer.

(3) In terms of data blocks, with the support of various companies or organization’s agreements or the support of certain laws, data can be directly shared without considering the leakage of user data caused by information concatenation, which will facilitate big data. The expansion of the shared resource pool further strengthens the data foundation of big data analysis and strengthens its application advantages.

To sum up, in terms of large medical data, applications based on data separate storage technology can solve most of the privacy security problems, separate data storage method as an inexpensive application method, to ensure their strengths and weaknesses practical application.

Conclusion

Regarding privacy and security issues in medical big data, previous measures such as data encryption have been weak. Therefore, this paper proposes a method based on data separation and storage, emphasizing that the “factors” that may lead to leakage are stripped off at the beginning of calculation, so the non-existent information will no longer be stolen, although this method is only in the basic theory conceived, but there is still the possibility of realization, and the disadvantages are small. Especially in the application of medical big data, relatively better results will be obtained. At the same time, the user's private data information will be properly protected. In turn, such a better application of data privacy protection will also make it widely used by itself.

Ultimately, data belongs to data, and privacy belongs to privacy.

Acknowledgement

Liaoning Natural Science Foundation, 201800177
The subject of educational science planning in Liaoning Province, JG16DB054

References

[1] Feng Dengguo, Zhang Min, Li Hao. Big Data Security and Privacy Protection [J]. Chinese Journal of Computers, 2014, 37 (1): 246-258.
[2] Xiong Jing, Wei Meng, Ma Siming, etc. A review of the impact of the era of big data on the development of medical and health archives management [J]. Low Carbon World, 2018 (1): 347-348.
[3] Wang Zechuan. Medical big data and its opportunities and challenges [J]. Information Recording Materials, 2018 (4).
[4] Beckmann JS, Lew D. Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities [J]. Genome Medicine, 2016, 8 (1): 134.
[5] Li Changrong. Big Data Thoughts in Medical Informatization [J]. Science & Technology and Innovation, 2018 (4): 89-90.
[6] Shen Tao, Cui Yong. Medical Big Data: Expectation and Reality [J]. China Digital Medicine, 2015 (7): 2-4.

[7] Yu Shuo, Du Zhuolin. Research on Application Technology of Medical Big Data [J]. Information and Communications, 2018 (1): 94-96.

[8] Luo Xu, Liu Youjiang. Research status and clinical application of medical big data [J]. Journal of Medical Informatics, 2015, 36 (5): 10-14.