Range Query in Blockchain-based Data Sharing Model for Electronic Medical Records

Jingwen Li and Wenlong Dun

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, 211106, China

Corresponding author’s e-mail: lijingwen118@nuaa.edu.cn

Abstract. The data sharing of electronic medical records (EMRs) has great positive significance for research on disease and epidemic prevention. Recently, blockchain-based EMRs storing and sharing schemes have brought many benefits. Nevertheless, the existing medical data sharing solutions cannot fully meet the query requirements in multi-user scenarios where data users with different identities have different needs for data usage. In this paper, we propose a block construction method for storing EMRs. We combine the characteristics of the Merkle tree and prefix dictionary tree in the block to reduce the time consumption of queries in the block. Besides, the block header is identified by four attributes, namely hospital ID, Department ID, disease type, and time period. According to the query conditions, the block where the requested data is located is quickly found. In the blockchain-based data sharing model of EMRs, data requesters with different identities can make combined queries with multiple conditions according to their own needs, and then the block and transactions in the block are located accurately. In conclusion, we are the first to build the block based on disease types and our model supports more flexible combined queries with multiple conditions.

1. Introduction

In modern society, more and more infectious diseases have drawn our attention. To a certain extent, the sharing of medical data can prevent the large-scale spread of diseases. By analyzing the medical data of a large number of patients, we can find the symptom of disease transmission in advance and control the large-scale spread of diseases in time. Therefore, the Health and Epidemic Prevention Departments urgently need to get real and immediate numerous medical records for constant attention [1]. For the biomedical research institutions, they need large numbers of reliable samples to study the influencing factors to the incidence rate of diseases, and then take some measures to reduce it [2]. But beyond that, one patient may choose to go to more than one hospital. Even worse, each hospital has its database, which cannot effectively communicate with each other. Hence, patients may be asked to do some repeated tests, which leads to some unnecessary waste of money and time. What’s more worrying is that some patients may deliberately conceal their medical history. This will have negative effects on the safety of doctors and the precise treatment of patients [3]. Therefore, communication between hospitals is also extremely important.

With the arrival of the big data era, the medical management system continues to upgrade. At the same time, EMR has become the most original record of patients in the whole process of hospital treatment [4]. But the current EMR data management system is not perfect, every hospital has its own medical database. There is almost no effective data sharing between hospitals. For biomedical research institutions and health and epidemic prevention departments, they are also unable to efficiently access
a tremendous amount of EMRs resources. Massive resources of EMRs not only face the dilemma of uneasily being shared on a large scale, but also the risk of privacy disclosure in the existing data dissemination process [5].

At present, there are some researches [6-10] on the secure sharing of EMRs, but most of them adopt the access control strategy based on blockchain technology. The implementation of access control in the system is usually divided into three steps, i.e., identification, authentication, and authorization. In addition to only allowing legitimate users to enter, access control also ensures reliability. It keeps track of what operations every user performs in the system. To a certain extent, the traditional medical data management system based on access control prevents malicious users from accessing medical data and provides a data sharing access model. On the basis of secure data access control, we need to consider how to make the data requester efficiently query data according to their specific needs.

However, although the technology of the data sharing model based on the blockchain provides a new method for the sharing of EMRs, the unique blockchain architecture brings challenges to the data query. For the data stored in the general database, we can search directly. But in the blockchain, we need to search the block first, and then search the needed transactions contained in the block. In order to search data effectively in blockchain structure, some query schemes [11-13] have been proposed to solve these problems caused by the blockchain structure, but they are still insufficient. Due to the different identities between users, the usage of data for them is also various, which leading to different request conditions. The query method in the block-chain structure must meet the query requirements of users with different identities. For example, when a hospital requests access to data, it is generally to retrieve some patients’ previous medical records. The query request may be in the form of {"Hospital A", "Patient ID", "2020.02.10-2020.02.15", "Asthma"}. Data requests from biomedical research institutions may be in the form of {"Ophthalmology", "Cataract", "Age: 60-80", "Sex: male"}. The query request from the health and epidemic prevention department may be in the form of {"Respiratory department", "2020.02.01-2020.05.01", "Pneumonia"}.

In this paper, on the basis of the blockchain-based medical data sharing model, we propose a query method for the unique block-chain structure of the blockchain, which is applicable to three kinds of users, i.e., hospitals, research institutions, and health departments. We take all the EMRs of the same disease within a time period as a data package, and upload it to the blockchain network. After verification, it is mined into a block and added to the chain. All EMRs in each block are stored in the form of a Merkle tree, which combines the characteristics of the prefix dictionary tree. When the blockchain network in the storage layer receives the query request, the block where the requested data is stored is first positioned. If the requested data is fine-grained, the intra-query will be performed in the block. According to the query condition, start the query from the root node until the qualified data or leaf node. To summarize, our contributions made in this paper are as follows:

- We propose a new method to construct the Merkle tree in the block. All the EMRs of the same disease within a time period are packed in a block. We store these medical records in the Merkle tree with the characteristics of the prefix dictionary tree. This method of building a Merkle tree is suitable for the storage and sharing of EMRs.
- The Merkle tree root and each EMR in the block are defined with multi-attribute, which allows users of different identities to freely combine the query conditions, and accurately locate the blocks and medical records that meet the conditions.
- We design the system as a three-layer data sharing model, which provides a secure access environment for users.

The rest of the paper is organized as follows. Section 2 constructs the system architecture and builds the Merkle tree in a new method. The specific data access process is described in Section 3. Time complexity and security analysis are presented in Section 4. Finally, we conclude our paper in Section 5.
2. System architecture

The data sharing system we set is used for data sharing based on the blockchain between untrusted parties, to enable them to query efficiently and prevent privacy leakage during data transmission. The main structure of this system is mainly divided into three layers, i.e., user layer, agent layer and storage layer.

2.1. User layer

The user layer is mainly composed of different types of users. The main purpose of these participating users is to use medical data for research and analysis, disease prevention and control, and public health and epidemic prevention. The participating users can be hospitals, research institutions, or government health departments. Users can perform multi-condition range queries and combine multiple query conditions such as attributes \{"Hospital ID", "Department ID", "Disease Type", "Time Period", "Patient ID", "Sex"\} according to their needs. Each user has two pairs of keys. The private key \(x_1\) of the first key pairs \((x_1, y_1)\) is used to sign the data request to be issued, and the private key \(x_2\) of the second key pairs \((x_2, y_2)\) is used to decrypt received data package.

2.2. Agent layer

There are two main purposes for setting the Agent layer. The first purpose is to authenticate the user who sent the query message, ensure the validity of the requester’s identity and access rights to the querying data, and authorize access to the valid requester. In this model, the agent layer can check the access rights of each block. If the requester has the access rights of the query’s blocks, the agent layer will authorize this access and return the encrypted data package to the user. The second purpose is to record users who perform illegal operations and revoke the access rights of users who have entered the blacklist.

2.3. Storage layer

The storage layer stores all medical records and provides query results to the agent layer. The components of the storage layer include the blockchain network and smart contract center.

2.3.1. Blockchain network.

The blockchain network mainly realizes the uploading, storage, and query aggregation of medical data. The blockchain network is composed of separate broadcast blocks and linked together in chronological order. The main function of the blockchain network is to maintain an immutable database distributed in chronological order and includes the operations about transmitting and requesting the system data. The medical data upload process is mainly carried out in the blockchain network. The broadcast nodes (usually the hospitals that upload EMRs) in the network sign the EMR data, and then forward it to the adjacent nodes. After the adjacent nodes receive and verify it, they forward it to the next neighboring node until all nodes pass the verification. The mining nodes in the network package the verified data into a block and the new block become a part of the blockchain.

In this scheme, we define 11 main attributes for each EMR. That’s to say, the medical records of each patient are listed as \{"Hospital ID", "Department ID", "Disease Type", "Time Period", "Attending Doctor ID", "Treatment Plan", "Patient ID", "Sex", "Age", "Address", "Origin", "..."\} including 11 main fields to identify. When the number of EMRs of a certain disease type reaches a certain value, these medical records are packaged. The Merkle tree structure of these medical records is shown in figure 1, which can be regarded as an improved prefix dictionary tree. The construction process of the Merkle tree in our scheme is as follows. First, all EMRs in a certain period time are divided into four parts, i.e., \{Male, age\leq 60\}, \{Male, age> 60\}, \{Female, age\leq 60\}, \{ Female, age> 60\}. The EMRs of each part are hashed in pairs until they form a Merkle tree containing the same prefix. The same prefix is used as an attribute attached to the non leaf node. The root of the Merkle tree is identified by four attributes, i.e., \{"Hospital ID", "Department ID", "Disease ID" and "Time Period"\}. That is, we can locate all the blocks containing one or more of the four attributes according to query criteria.
When the storage layer’s blockchain network receives a query request, it first locates the block where the requested data is stored according to the query conditions. If the granularity of the query data is at the block level, then all the blocks found are the query results. If the query conditions are more detailed, for example, including attributes as \{"Age", "Sex", ...\}, after locating the block, continue to query within the block. Assume that the four identification attributes of the Merkle tree root in figure 1 are \{"Z Hospital", "Ophthalmology Department", "January 1, 2020 – January 31, 2020", "Cataract"\}. That is, we want to inquire about all patients’ cataract medical records from January 1, 2020 to January 31, 2020 in Z Hospital’s Ophthalmology Department. According to the query conditions, the block that meets the conditions is first located, and then make a search in this block. The blue area in figure 1 is the query results that meet the conditions.

![Figure 1. The structure of Merkle tree.](image)

2.3.2. Smart contract center.
The smart contract center runs on the Ethereum platform. A smart contract is a script (code) written in advance. Once the predetermined conditions are activated, the script (code) on the contract can be executed without the help of external trusted authorization, and the executed transaction records are audited on the public platform. The main function of the smart contract is to monitor the user’s operation on the queried data and report the data operation to the agent layer. Once the user performs a violation operation, the smart contract can revoke the user’s right to access and operate the data, and the agent layer pulls the user into the access blacklist.

3. Access model
This section will detail the specific process that the user access to medical data, including data request, data query, data transmission and data monitor.

3.1. Data request
According to the attributes on the EMRs and Merkle tree root, the user can freely combine them forming the query conditions. Using the ElGamal digital signature scheme [14], the query request is signed by the user using the private key \(x_1\) of the first key pair \((x_1, y_1)\). And then sends the data request to the agent layer. The agent layer receives the query request from the user layer. And then verifies the validity of the signature by using the requester’s public key \(y_1\) generated and shared before
sending the query request. If the signature is valid, the query request will be accepted, otherwise the invalid request will be deleted. The agent layer in the model can check the access rights of each block in the blockchain. If the requester has all access rights to the queried block, the agent layer will forward this data request to the storage layer.

Figure 2. Search path for query.

3.2. Data query
When the storage layer receives this data query request, then searches in the blockchain network according to range query conditions. We can note that the Merkle tree is composed of some transactions and identified by several attributes. Therefore, when performing range conditions search for data on the blockchain, we first locate all the blocks that meet the conditions. If the query conditions include attributes in the EMRs, then continue the search in these blocks. Search down from the Merkle tree root until the query conditions are met or leaf nodes. For example, we will query the medical records in the form of \{"Affiliated Hospital of Z University", "Ophthalmology", "Cataract", "2020.02.10-2020.02.27", "Hua Li"\}. Figure 2 shows the search path.

According to the above query method, the requested data is obtained. Aggregated the data and then forward it to the smart contract center in the same layer. In the smart contract center, set the operation permissions on the request data according to the identity of this requester, and generate a smart contract. A Smart contract is a section automatically executed code, which is encapsulated as a package together with the request data. Therefore, the data package (DP) contains request data (QR) and smart contracts (SC). A Smart contract can effectively monitor the lifeline of this data package. The package is identified by a package ID. Using ElGamal public key encryption scheme [14], encrypt the package with the public key \(y_2\) of the requester’s second key pair \((x_2, y_2)\). The encrypted data package is in the form of \(DP = E_{PKU}[Data-ID|QR|SC]\). Only the requester having a correct private key and the valid identity can decrypt this data package.

3.3. Data transmission
The storage layer forwards the package to the agent layer. Then the agent layer uses the ElGamal digital signature scheme to sign it. The signed encrypted package is in the form of \((DP, (r, s))\). The signature is to ensure that this package is forwarded by the agent layer. Besides, the signature can ensure that the package has not been maliciously tampered. Then, the agent layer forwards the signed package to the user layer.

3.4. Data monitor
The user layer receives the query result in the form of \(Q = (DP, (r, s))\). The user firstly verifies the signature of the agent layer, and decrypts the encrypted data package after verification. The operation
of decrypting will activate the smart contract attached to the request data. Once the smart contract is activated, all the user’s operations on the requested data will be monitored. When the preset permissions such as malicious modification is exceeded, the smart contract will run a program to automatically destroy data. At the same time, it will send the user’s illegal operation behavior to the agent layer. The agent layer will add this user into the blacklist after receiving report, and also will reject to receive this user’s query request.

4. Discussion

In this section, we will discuss our scheme from two aspects including time complexity and security analysis.

4.1. Time complexity

When our query condition is accurate to a patient’s medical record in specific time, the time consumption is mainly in two stages. Firstly, a specific block will be located through query condition in the form of \("Hospital ID", "Department ID", "Disease category", "Time range"\). Secondly, based on the EMR level attributes contained in the query condition, we continue searching in the block. The Merkle tree constructed in our scheme combines the characteristics of prefix dictionary tree. According to the query conditions, find the matching branches, and continue searching down until the EMRs matching all the query conditions or no result.

The time consumption in our query scheme is divided into two part. The first part is to locate a specific block. According to the blockchain sequence, in the best case, the query block is in the first, so the time complexity is $O(1)$. In the worst case, the query block is the last one, so the time complexity is $O(n)$. The average time complexity is $O(n)$. The second part of time consumption is to locate specific medical records or subtrees by searching in Merkle tree which stores numerous EMRs. If sex and age attributes are given in the query condition, make the binary search on Merkle tree. The average time complexity is $O(\log n)$. The average time complexity contained inter-block query and intra-block query is $O(n)$.

4.2. Security analysis

The user in the user layer makes a data request and signs it. The agent layer receives the request from the user layer. And then the agent layer will verify the identity of the user and confirm that it is a legitimate user. If the user has access to the requested data, this data request will be forwarded to the storage layer. Even if a malicious adversary forges a data request, the agent layer can identify the malicious user and delete the malicious access request by verifying the signature on the request. According to the query conditions, the storage layer will make a search in the blockchain network of this layer. The query result will be attached a smart contract and encrypted with the user’s public key.

The encrypted package is forwarded to the agent layer. The agent layer signs and forwards the package to the user layer. If a malicious adversary obtains the encrypted package, he still can’t decrypt the package and forge the signature. After receiving the encrypted package, the user will verify the signature. If the signature is verified, it proves that the package has not been forged or maliciously modified. Then, the user decrypts this package. At this time, the decryption action triggers the smart contract. The smart contract monitors the user’s operation behaviour in the whole process. Once there is any illegal operation such as modifying the attending physician’s signature on the EMR, this behaviour will be sent to the agent layer. The agent layer will pull the user into the blacklist. The whole process of data sharing is secure and can’t be tampered with by malicious adversaries.

5. Conclusion

In this paper, we introduce some existing problems about data sharing of EMR in the multi-user scenario. According to the existing problems, we design a blockchain-based data sharing model of EMR. In this model, we construct a kind of Merkle tree combining the characteristics of the prefix dictionary tree, which is specially used to store EMRs. In the three-layer data access model composed
of the user layer, agent layer, and storage layer, each EMR is defined by multiple attributes and packaged according to disease category. By defining four attributes for each block in the blockchain network, we can quickly search the block where the requested data is located. Then, the binary tree is searched according to the attributes of EMR level contained in the query condition. This scheme solves the multi-condition combination query in a multi-user scenario.

This paper provides a new direction for the storage and sharing of EMRs. How to improve the query efficiency, how to ensure the security and privacy of query processing, and how to expand the storage space of blockchain are all problems worthy of in-depth study.

References

[1] Klompas, M., McVetta, J., Lazarus, R., Eggleston, E., Haney, G., Kruskal, B. A., ... & Lee, M. (2012). Integrating clinical practice and public health surveillance using electronic medical record systems. American journal of public health, 102(S3): S325-S332.

[2] Kho, A. N., Pacheco, J. A., Peissig, P. L., Rasmussen, L., Newton, K. M., Weston, N., ... & Kullo, I. J. (2011). Electronic medical records for genetic research: results of the eMERGE consortium. Science translational medicine, 3(79): 79re1-79re1.

[3] Wei, W. Q., & Denny, J. C. (2015). Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome medicine, 7(1): 41.

[4] Hazumi, M. (2001). U.S. Patent Application No. 09/864,261.

[5] Perera, G., Holbrook, A., Thabane, L., Foster, G., & Willison, D. J. (2011). Views on health information sharing and privacy from primary care practices using electronic medical records. International journal of medical informatics, 80(2): 94-101.

[6] Xia, Q. I., Sifah, E. B., Asamoah, K. O., Guo, J., Du, X., & Guizani, M. (2017). McDShare: Trust-less medical data sharing among cloud service providers via blockchain. IEEE Access, 5: 14757-14767.

[7] Xia, Q., Sifah, E. B., Smahi, A., Amofa, S., & Zhang, X. (2017). BBDS: Blockchain-based data sharing for electronic medical records in cloud environments. Information, 8(2): 44.

[8] Guo, R., Shi, H., Zhao, Q., & Zheng, D. (2018). Secure attribute-based signature scheme with multiple authorities for blockchain in electronic health records systems. IEEE access, 6: 11676-11686.

[9] Dubovitskaya, A., Xu, Z., Ryu, S., Schumacher, M., & Wang, F. (2017). Secure and trustable electronic medical records sharing using blockchain. In AMIA annual symposium proceedings. American Medical Informatics Association. p. 650.

[10] Liu, J., Li, X., Ye, L., Zhang, H., Du, X., & Guizani, M. (2018). BPDS: A blockchain based privacy-preserving data sharing for electronic medical records. In 2018 IEEE Global Communications Conference. Abu Dhabi. pp. 1-6.

[11] Xu, C., Zhang, C., & Xu, J. (2019). vchain: Enabling verifiable boolean range queries over blockchain databases. In Proceedings of the 2019 international conference on management of data. Amsterdam Netherlands. pp. 141-158.

[12] Zhang, X., & Poslad, S. (2018). Blockchain support for flexible queries with granular access control to electronic medical records (EMR). In 2018 IEEE International conference on communications. Kansas City. pp. 1-6.

[13] Chen, L., Lee, W. K., Chang, C. C., Choo, K. K. R., & Zhang, N. (2019). Blockchain based searchable encryption for electronic health record sharing. Future Generation Computer Systems, 95: 420-429.

[14] ElGamal, T. (1985). A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE transactions on information theory, 31(4): 469-472.