Medical Data Compression and Sharing Technology Based on Blockchain

Yi Du[1] and Hua Yu[2(✉)]

[1] College of Engineering, Shanghai Polytechnic University, No. 2360 Jinghai Road, Pudong District, Shanghai 230020, People's Republic of China
[2] Information Center, Shanghai General Hospital, No. 100 Haining Road, Shanghai 200080, People's Republic of China
h.yu@shgh.cn

Abstract. Cloud service provides distributed storage spaces and avoids computing bottlenecks of centralized database storage, but it cannot provide secure data storage and sharing. How to store and share medical data efficiently and reliably is an important issue to ensure the safety of medical data in the process of medical data management of multi-regional hospitals. The public access of blockchains and the non-modifiable characteristics of the stored data make it an effective implementation scheme for medical data sharing. Meanwhile, with the continuous transactions in blockchains, the data of the blockchains are bound to be larger and larger, which leads to serious issues for the use and storage of data. Based on the LZW (Lemple-Ziv-Welch) algorithm, this paper presents a lossless compression technology for Chinese text compression with a compression storage and sharing scheme for medical data using blockchains to provide safer and more efficient access services for medical data.

Keywords: Blockchain · Medical data · Text compression · Data storage · Data sharing

1 Preface

Medical information is the valuable information from patients. However, in the current medical system of Chinese domestic hospitals, most information cannot be shared by all the hospitals, so the patients sometimes should apply for a new medical card to record the medical information of themselves, making the previous medical information of the patients useless or easily wrong. Although most hospitals use paper-based medical records, these records are very easy to be damaged or lost, which is a very unreliable way for medical information recording. On the other hand, with the development of cloud computing, distributed databases can realize medical information sharing, but may cause losses to patients because of the leakage of patients' information. Therefore, medical staff and patients need a system that can share medical information among hospitals and ensure the safety and reliability of stored medical data. As a distributed database system with multiple independent nodes, a blockchain is the ideal way to realize this system at present due to the advantages of decentralization, no trust, strong tamper resistance, etc.

Z. Zhang et al. (Eds.): AAIM 2020, LNCS 12290, pp. 581–592, 2020.
https://doi.org/10.1007/978-3-030-57602-8_52

A blockchain is expected to solve the common problems of low data security or poor sharing in the existing medical system. At the same time, with the continuous expansion of transactions in a blockchain, the amount of data it stores is also increasing. In order to improve the efficiency of a blockchain, this paper proposes the following solutions based on blockchains and text lossless compression technology:

1. Hospitals build a blockchain to jointly manage and maintain the stable operation of the blockchain, ensuring data security and reducing the cost of data protection;
2. The LZW (Lemple-Ziv-Welch) algorithm is used to compress the original medical data, compress and store the medical text data and realize the effective compression of medical information to relieve the storage pressure;
3. Based on the idea of text data compression given in this paper, the realization of medical data management and sharing under the blockchain is discussed, which provides a new idea for the effective management of medical data under the blockchain.

Section 2 of this paper introduces the related work of text compression algorithm and blockchain technology. Section 3 introduces the blockchain, text lossless compression and other related technologies involved in this method. Section 4 introduces the specific design of the method, and finally the work is summarized in Sect. 5.

2 Research Progress

2.1 Blockchain

In recent years, many researchers and institutions at home and abroad use blockchain technology to explore and practice in the fields of data protection and sharing [1]. In 2013, Araoz et al. realized the authenticity protection of electronic files by storing hash values in fields in blockchain transactions [2]. Based on the blockchain, Vaughan et al. proposed a general file protection framework, which computes the file hash values and builds the Merkel tree to reduce the cost of data protection [3, 4]. In 2016, Azaria and others constructed a decentralized medical data access and authority management system by using smart contracts to realize patients' ownership of their medical data and enable them to independently share and manage medical records [5]. In November of the same year, Weide Cai and others put forward the development method of application system based on blockchain, including the design model of account chains and transaction chains, as well as the application principle of parallel code execution model on the chain [6].

In China, ant financial services carried out tracking management of donation and donation flow based on blockchain technology in 2017, improving the transparency, traceability and nonmodifiable characteristics of its system and data [7]. In 2018, Baidu applied the blockchain technology to the data protection of Baidu Encyclopedia, recording the historical version of each update of the encyclopedia entry, the author, editing time and other information on the blockchain, achieving the purpose of data protection and storage [8]. In October 2017, Zhang Ning et al. realized a solution framework of personal privacy data protection by using blockchain, database, asymmetric encryption and

other technologies, initially realizing the protection of personal privacy in the Internet car rental scene [9].

In the field of medical treatment, American scholar Kevin Peterson et al. used blockchain technology to realize medical information sharing, and proposed a novel consensus mechanism, which uses the accuracy of semantics as proof to mine and generate blocks [10]. Researchers such as Ekblaw from MIT Media Laboratory in the United States, used Ethereum as the platform to write smart contracts to realize a distributed information management system "MedRec", realizing the information security of sensitive medical information through identity authentication, encryption, sharing and distributed storage of data [11]. At the same time, a proof of concept mechanism was proposed to ensure the normal operation and maintenance of the system. Researchers at the University of California, San Diego, proposed a blockchain-based privacy protection framework for decentralized medical information "ModelChain", which analyzes the data in the context of the private chain with artificial intelligence, and can design the information certificate without explicitly displaying the patient's medical information The algorithm is used to determine the processing order, which provides a good solution to protect the sensitive data of patients by combining the related technology of artificial intelligence [12].

In China, Xia Qi and others from the University of Electronic Science and technology proposed MeDShare scheme, which uses blockchain to realize medical data sharing, and completes the verification, audit and information sharing control of medical data [13]. In the scheme, smart contract is also introduced to track data information and realize data traceability. If there is an illegal transaction, smart contract can automatically revoke the access rights of illegal users to improve the security of the system. In 2018, Daiying Dong and Xueming Wang proposed to let the nodes in the hospital alliance service group register the public key on the network, and then add the public key to the header of each item of transaction data. The client uses the API interface provided by Web3 to interact with the node, and encrypts the user's private key once through the user's password, so as to ensure that patients can query cases through identity documents [14]. Yanhui Ren from Xi'an University of Electronic Science and technology proposed a scheme of medical information privacy protection and sharing based on blockchain. On the basis of ensuring the dense storage of archive data, the scheme enables users to achieve fine-grained access control for each record, and verifies the feasibility and performance overhead of the scheme [15].

2.2 Data Compression Algorithm

In the early 19th century, researchers replaced the characters commonly used in text using the code named MC (Morse Codex's), that is, using shorter characters to encode commonly used characters to achieve the compression effect. Then, the S-F (Shannon fan) coding algorithm appeared. According to the probability of symbol occurrence, the algorithm uses shorter coding symbols to replace the original symbols. In 1952, David A. Huffman proposed Huffman coding, which was monopolized from 1960s to 1980s [16]. Although Huffman coding has some advantages in data compression compared with the compression algorithm before 1952, it also has a fatal weakness, that is, when the algorithm compresses the data, it needs to scan the original data twice [17]. In

1977, Abraham Lempel and Jacob Ziv used dynamic dictionaries to replace the repeated strings in the text with shorter symbols, which greatly improved the compression ratio. In 1978, they published the improved LZ78 algorithm, whose idea is to generate a static dictionary set based on the input data. The algorithm is more stable and effective.

According to the requirements of application background, there are more and more improved algorithms based on LZ77 and LZ78 compression algorithms [18], among which LZW algorithm is the most popular. LZW algorithm is an optimized version of LZ78 algorithm proposed by Terry Welch in 1984 [19]. In addition to inheriting the fast compression and decompression characteristics of LZ78 algorithm, the compression ratio of this algorithm is higher. LZW can dynamically generate a dictionary set to save the processed historical data when compressing and decompressing the data. When "prefix, character" cannot correspond to each other in the algorithm dictionary, the prefix is encoded and output to realize data compression. For medical data, this paper presents a dictionary-based LZW Chinese compression method in the process of blockchain management and sharing to reduce the increasing pressure of transaction data storage and improve the transaction efficiency in blockchains.

3 Related Technologies

3.1 Main Idea of Blockchain

Blockchain technology can safely store bitcoin transactions or other data, and ensure the security of these data or information to prevent tampering and forgery. Different from the common relational database and the non-relational database, the core of blockchain is decentralization. By using encryption algorithms such as digital signatures, hash algorithms and distributed consensus algorithms, the stored data are very difficult to be tampered with, destroyed or erased from the database operation log. Under the premise that peer-to-peer nodes do not need trust, the decentralized point-to-point transaction, coordination and cooperation are realized to solve the problems of high cost, low efficiency and insecure data storage in the centralized database application system.

Blockchain system is divided into six layers: the data layer, the network layer, the consensus layer, the incentive layer, the contract layer and the application layer. Among them, the data layer encapsulates the underlying block of a blockchain, hash function, data encryption and time stamp; the network layer includes point-to-point technology, propagation mechanism and verification mechanism; the consensus layer encapsulates various consensus algorithms of network nodes; the incentive layer mainly includes the distribution mechanism; the contract layer mainly encapsulates various scripts, algorithms and smart contracts, which is the basis of the programmable characteristics of the blockchain; the application layer is the application of the blockchain in various scenarios. Smart contract is the core element of blockchain, which is triggered by events and runs on the blockchain data ledger. It is used to realize the interaction between a blockchain and contract application.

Each node in the blockchain can encapsulate the transaction data into a data block with time stamp, and link to the main block to form the latest block, and then synchronize the block information into the whole blockchain network. The data block includes block head and block body. The block header encapsulates the address of the previous block

and a series of hash values, which are used to link the front block and the back block; the block body mainly contains the main information of the block. The composition of a blockchain is shown in Fig. 1.

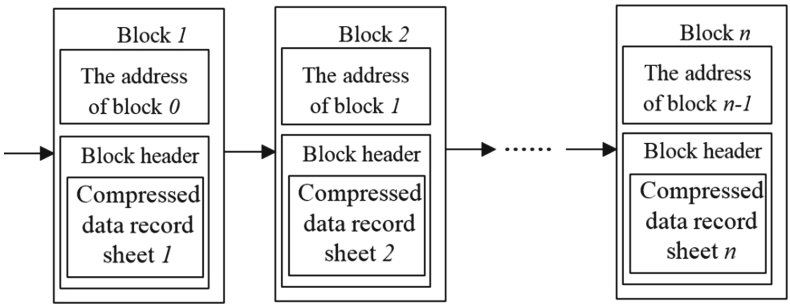

Fig. 1. The composition of a blockchain

3.2 LZW Algorithm

LZW algorithm is improved by Terry A. Welch on the bases of LZ78 and LZ78. It is a compression algorithm based on dictionary. The idea of dictionary compression algorithm is to store the strings appearing in the data as entries in the dictionary, and code these entries to replace the relatively long strings in the data with the encoding of entries. Its characteristic is that the dictionary does not need to be stored in the compressed file with the compressed data.

LZW algorithm initializes the dictionary at the beginning of compression. After initialization, the dictionary contains 256 single characters. After the compression starts, read the strings and match the entries in the dictionary one by one in order. If the match is successful, continue to read in the next character to form a new string and continue to match until the match fails. Number the string and add it to the dictionary as a new entry. Output the last matching to the number corresponding to the string. In this way, the number of entries in the dictionary will automatically increase as the data are compressed. Then the matching probability of strings in the data will be increased to achieve the purpose of data compression.

LZW algorithm is independent of the statistical characteristics of probability with compressed data. Therefore, this algorithm can be used in real-time data compression, which is very important because sometimes it is impossible to know the probability and statistical characteristics of each character in the compressed data in advance. For the data source with a high repetition rate, LZW algorithm can get a satisfactory compression rate.

LZW algorithm is a compression algorithm based on dictionary. The dictionary is built dynamically in the process of data compression. In the process of compressed data transmission and storage, the dictionary does not need to be transferred and stored together with the compressed data. When the data is decompressed, the dictionary can be

generated dynamically in the process of decompressing. The dictionary constructed during compression is exactly the same with the dictionary constructed during decompressing. The capacity of the dictionary is not large, and the compression and decompression speed is high.

4 Method Design

4.1 Lossless Compression of Chinese Based on LZW

(1) **Chinese coding method based on LZW**

The core of LZW compression algorithm is a conversion table maintained in the process of compression, which is a dictionary. Because the basic processing unit of the general LZW algorithm is bytes for English characters, if it is directly applied to Chinese characters, the hidden semantic information in Chinese data coding will be lost artificially. Therefore, in this paper, the original LZW algorithm is improved to make it more suitable for the actual application scenarios.

In the improved algorithm given in this paper, GB2312 standard is used to obtain the code value of Chinese characters. In order to avoid too large initial dictionary, only common Chinese characters are added to the basic code set of the dictionary in advance, and space is reserved for undefined code words that may be encountered in the compression process. Through this dictionary, the longer Chinese string in the input data can be converted into shorter encoding as entry to achieve the compression goal instead of the relatively longer strings in the input data. The process of algorithm compression is shown in Table 1. In the process of compression, according to certain rules, the algorithm adds the first Chinese string encountered in the encoder to the dictionary, and assigns a unique flag value called code value to the added string.

Table 1. Chinese coding process based on LZW

Step 1	Initialize dictionary
Step 2	Input a Chinese character c, prefix string P=c
Step 3	Encoding conversion:
(1)	While c is not an end character do
(2)	If P+c is in the dictionary
(3)	Then P=P+c
(4)	Else
(5)	Find the code of P in the dictionary;
(6)	Add P and c to the dictionary;
(7)	P=c;
(8)	Output the encoding of P

(2) **Chinese decoding method based on LZW**
Compared with the compression process, the restoration process of LZW algorithm is shown in Table 2. The key to the restoration process is that the initialized dictionary must be consistent with the compression program, and the dictionary maintained during the restoration process is almost synchronized with the compression process. In a blockchain, this process is mainly used for data users to decode the acquired data information.

Table 2. Chinese reduction and decoding process based on LZW

Step 1	Initialize dictionary
Step 2	Input the first encoding num and assign it to the reserved string *O*
Step 3	Output *O*
Step 3	Decoding conversion:
(1)	While *num* is not an end character do
(2)	Find the string *N* corresponding to the encoding *num* in the dictionary;
(3)	If *N* is Null
(4)	Then *N*=*O*+*N*;
(5)	Output *N*
(6)	Add *O*+*N* to the dictionary;
(7)	*O*=*N*

4.2 Sharing and Acquisition of Medical Compressed Data

(1) **Blockchain model**
The compressed blockchain model proposed in this paper mainly includes three entities: a data owner, a data manager and a data demander. The details are as follows:

1. Data owner: the owner of data, responsible for the collection and provision of data. Here it mainly refers to patients or scientific research institutions or hospitals authorized by patients. The data owner needs to ensure the authenticity and reliability of data.
2. Data manager: responsible for compressed data storage and publishing. The data manager encrypts the compressed medical data and saves them on the cloud server. Only authorized users can download them.
3. Data demanders: entities that need to retrieve and use medical data. In the paper, they mainly refer to scientific research institutions or hospitals. Data demanders can query the medical data they are interested in from the blockchain network, record the hash values and download the complete data from the cloud server.

(2) **Blockchain sharing process of medical data**
The sharing of medical data refers to the sharing of safe and reliable medical data within the owner of the medical blockchain or outside the blockchain by using the

demanders through intelligent contracts and hybrid encryption mechanisms, and can ensure the safe and efficient storage of data in the blockchain. It includes:

1. Building a data sharing model: in this stage, build a medical data sharing model on each sharing node, which includes data processing module;
2. Data processing stage: in this stage, the data owners participating in the sharing use the data processing module to collect, compress and store the data under their jurisdiction, make classification marks on the data as the sharing labels, and sign the data information using the private key;
3. Data communication stage: in this stage, the data users participating in data sharing perform node initialization configuration, and generate index data blocks containing the unique identity of the node in each node of the blockchain;
4. Data acquisition: the data demander applies for obtaining encrypted data information, and obtains the decryption key by sending his/her own identity; the data sharing module checks the authenticity of the shared record through the smart contract to determine whether the shared information matches successfully;
5. Data operation: each node in the area chain determines whether to allow this operation. If the number of nodes allowed for this sharing request is less than half of the total number of nodes, the data sharing request will be rejected. Otherwise, the operation will be allowed, and the operation will be time-stamped to record, generating a data operation block.

5 Performance Analyses

The mixed compression method including Chinese and English characters is implemented using Java. The experimental environment is given as follows:

- CPU: Intel (R) Cor e(TM) i7-5600U CPU @ 2.60 GHz;
- Memory: 12 GB;
- Operating system: Windows 7 professional (64 bit);

In order to test the effectiveness of the improved LZW algorithm, five groups of medical data files were randomly selected as experimental data to test the compression effect. The code length of the Chinese dictionary is 16 bit and the number of entries is 64K. Due to the fact that both Chinese and English characters are included in the conventional medical data information, the compression effects of Chinese medical data, English medical data and mixed medical data were tested and compared in the experiment. In the compression process of Chinese and English medical data, the experiment is completed by separating Chinese characters and English characters, and then compressing them respectively. The experimental results record the data compression ratio (data size after compression/data size before compression) and the compression time. The experimental results of compression performance are shown in Figs. 2, 3 and 4, respectively. Through the comparison of the three figures, it is not difficult to find that although the amount of mixed data changes, the compression rate of the algorithm remains constant, which continues the better compression performance of LZW compression algorithm.

Tables 3, 4 and 5 show the time taken to compress different data files. It can be seen that the compression time increases with the increase of data volume, but the compression time is also within the acceptable range for users.

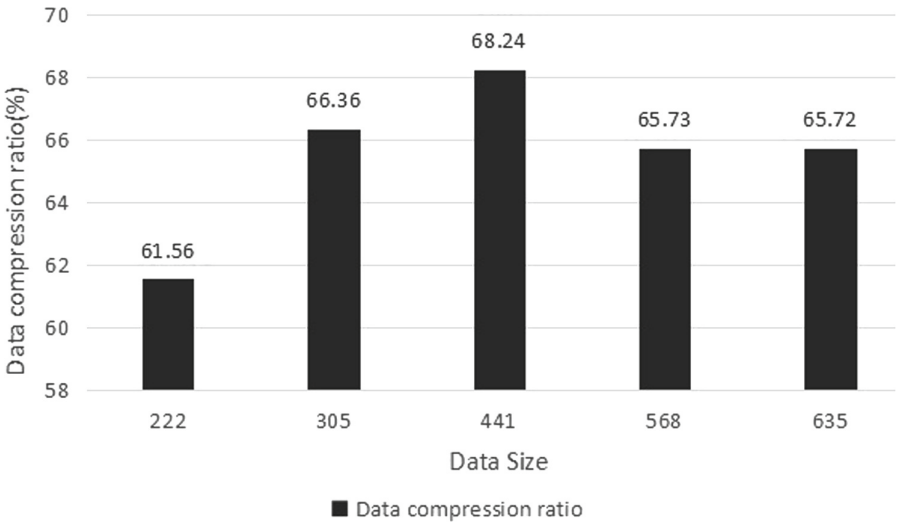

Fig. 2. The compression effect of mixed medical data

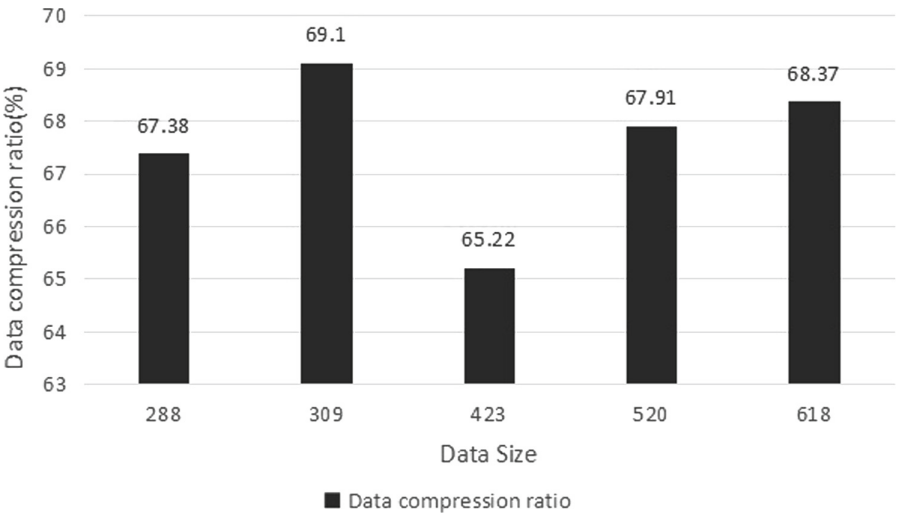

Fig. 3. The compression effect of Chinese medical data

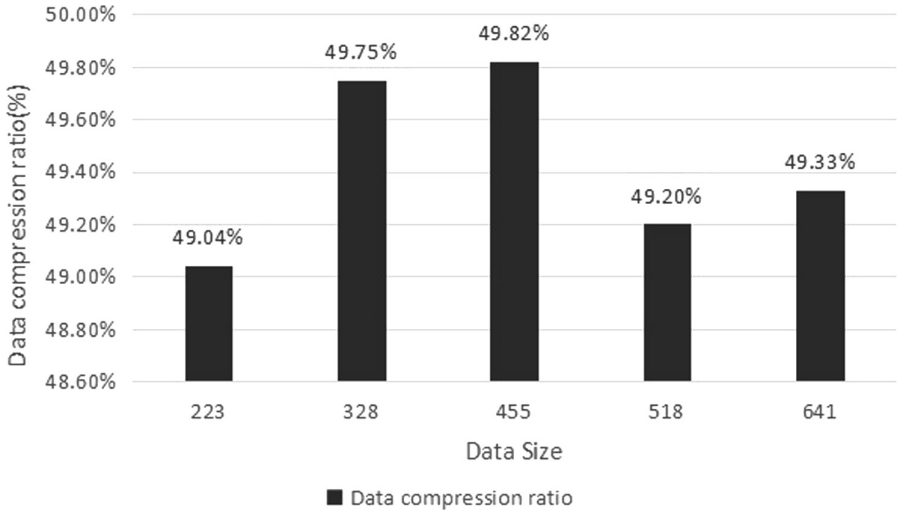

Fig. 4. The compression effect of English medical data

Table 3. The compression effect and time of mixed medical data

File size	Mixed medical data	
	Compression ratio (%)	Compression time (ms)
222	61.56	136
305	66.36	181
441	68.24	253
568	65.73	320
635	65.72	337

Table 4. The compression effect and time of Chinese medical data

File size	Chinese medical data	
	Compression ratio (%)	Compression time (ms)
288	67.38	123
309	69.10	162
423	65.22	242
520	67.91	297
618	68.37	318

Table 5. The compression effect and time of English medical data

File size	English medical data	
	Compression ratio (%)	Compression time (ms)
223	49.04	117
328	49.75	156
455	49.82	235
518	49.2	308
641	49.33	347

6 Conclusions

At present, the sharing of medical data among medical related institutions is always a hot research issue, so it is of great significance to ensure the privacy of medical data and realize the sharing of electronic medical compressed records based on blockchains. In this paper, based on the characteristics of blockchains, such as decentralization and non-modifiable characteristics, and in view of the increasing data volume with blockchain transactions, a medical data compression and sharing scheme based on blockchain is proposed. In this paper, a mixed LZW algorithm is proposed. In this experiment, some medical data information is randomly extracted, and the effectiveness of the algorithm is analyzed from the aspects of compression effect and compression time. The corresponding blockchain compressed data model and sharing process are given. It provides a technical idea for realizing the safe and efficient sharing of medical data among authorized users.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Nos. 41672114, 41702148).

References

1. Zhao, Z.: Research and Design of Digital Archive Management System Based on Blockchain. University of Science and Technology of China, Hefei (2018)
2. Proof of existence–An online service to prove the existence of documents (2018). https://docs.proofofexistence.com/
3. Merkle, R.C.: A digital signature based on a conventional encryption function. In: Pomerance, C. (ed.) CRYPTO 1987. LNCS, vol. 293, pp. 369–378. Springer, Heidelberg (1988). https://doi.org/10.1007/3-540-48184-2_32
4. Verify a chainpoint proof directly using Bitcoin (2017). https://runkit.com/tierion-/verify-a-chainpoint-proof-directly-using-bitcoin
5. Azaria, A., Ekblaw, A., Vieira, T., et al.: MedRec: using blockchain for medical data access and permission management. In: Proceedings of the International Conference on Open and Big Data, pp. 25 − 30. IEEE (2016)
6. Tsai, W.T., Yu, L., Wang, R., Liu, N., Deng, E.Y.: Blockchain application development techniques. J. Softw. **28**(6), 1474–1487 (2017)

7. Blockchain + public welfare, concept or trend (2017). http://www.xinhuanet.com/gongyi/2016-12/21/c_129414848.htm
8. Baidu's 'Wikipedia' now logs revisions on a blockchain (2018). https://www.coin-desk.com/-baidus-wikipedia-now-logs-revisionson-a-blockchain
9. Zhang, N., Zhong, S.: Mechanism of personal privacy protection based on blockchain. J. Comput. Appl. **37**(10), 2787–2793 (2017)
10. Peterson, K., Deeduvanu, R., Kanjamala, P., et al.: A blockchain-based approach to health information exchange networks. In: Proceedings of the NIST Workshop Blockchain Healthcare, vol. 1, pp. 1–10 (2016)
11. Ekblaw, A., Azaria, A., Halamka, J.D., et al.: A case study for blockchain in healthcare: "MedRec" prototype for electronic health records and medical research data. In: Proceedings of IEEE Open & Big Data Conference, vol. 13, pp. 1–13 (2016)
12. Kuo, T.T., Ohno-Machado, L.: ModelChain: Decentralized Privacy-Preserving Healthcare Predicting Modeling Framework on Private Blockchain Networks. arXiv preprint arXiv:1802.01746 (2018)
13. Xia, Q., Sifah, E.B., Asamoah, K.O., et al.: MeDShare: trust-less medical data sharing among cloud service providers via blockchain. IEEE Access **5**, 14757–14767 (2017)
14. Dong, D., Wang, X.: Research on electronic medical record sharing model based on blockchain. Comput. Technol. Dev. **05**, 1–4 (2019)
15. Ren, Y.: A privacy protection and sharing scheme of medical information based on blockchain. Xi'an University of Electronic Science and technology (2018)
16. Huffman, D.A.: Technique for the construction of minimum-redundancy codes. Proc. IRE **40**(9), 1098–1101 (1952)
17. Anas Almarri, A., Al Yami, B., et al.: Toward a better compression for DNA sequences using Huffman encoding. J. Comput. Biol. **24**(4), 280–288 (2017)
18. Ziv, J., Lempel, A.: Compression of individual sequences via variable-rate coding. IEEE Trans. Inf. Theory **24**(5), 530–536 (1978)
19. Welch, T.A.: A technique for high-performance data compression. Computer **17**(6), 8–19 (1984)