Research on the Method of Eliminating Duplicated Encrypted Data in Cloud Storage Based on Generated Countermeasure Network

Lai-feng Tang[1](✉) and Qiang Wang[2]

[1] Xinjiang Institute of Technology, Aksu 843100, China
[2] School of Intelligence Technology, Geely University, Chengdu 641402, China

Abstract. In order to improve the efficiency of cloud storage and save network communication bandwidth, data De-duplication technology has been widely used. At the same time, data encryption brings new challenges to the De-duplication technology. Therefore, the method of data De-duplication based on cloud storage encryption of generative countermeasure network is proposed to get rid of the constraints of third-party servers. The popularity of user data is divided into two layers. The semantic security of non-popular data is ensured by double-layer encryption. The inner layer is convergence encryption and the outer layer is symmetry encryption. When data popularity changes, the cloud server just needs to remove the outer layer of encryption and store the convergent encryption results. The security analysis of the scheme is given, and the performance of the scheme is discussed by performance analysis and comparison. The simulation results show that the scheme is feasible and efficient.

Keywords: Generative countermeasure network · Cloud storage · Encrypted data · Data deduplication

1 Introduction

In order to protect data privacy, more and more users upload data to cloud server after encrypting. In the information age, the generation of massive data makes data storage a problem. Enterprises and individuals use cloud storage services to store data one after another. When users need to upload large amounts of data, it will cause huge delay, and there is a large amount of data redundancy on a global scale. In response to the current problem, researchers have proposed the data deduplication technology, which reduces data redundancy at the data block level or the file level to improve storage resources and network bandwidth utilization. Using this technology, for the same file, no matter how many users want to upload, if the file already exists in the server, all the file owners link to the file, the customer does not need to upload again.

A cloud storage based on emergent against networks encrypted data to the heavy method, its improved thought is: introducing a random number, ensure timely effectiveness of every file ownership certification process, even if the attacker intercepts ciphertext

W. Fu et al. (Eds.): ICMTEL 2021, LNICST 387, pp. 590–601, 2021.
https://doi.org/10.1007/978-3-030-82562-1_57

hash value, not a random number, also can't calculate the real-time and effective evidence, cannot pass file ownership certification, achieve the purpose of avoid replay attack [1]. The KP algorithm in ML scheme is used to extract the key from the original file instead of using the file itself as the encryption key. Therefore, the generative adversation network scheme not only improves the security, but also greatly reduces the amount of computation in the encryption and decryption process, enabling users to use cloud storage services more convenient and efficient [2]. Therefore, the purpose of this paper is to retain the original features of the scheme, repair the scheme, and make it more safe and efficient.

2 Cloud Storage Encrypted Data De-duplication Method

2.1 Cloud Storage Encryption Data Privacy Optimization

Considering the protection of user privacy information, this paper proposes a generative countermeasure protocol for De-duplication of anonymous encrypted data in a generative countermeasure network, which hides the communication between the network user and the cloud storage server by relying on the anonymous channel technology, and introduces digital certificates to ensure the normal access of data files in a generative countermeasure network [3]. The generative anti-network anonymous De-duplication model is shown in Fig. 1.

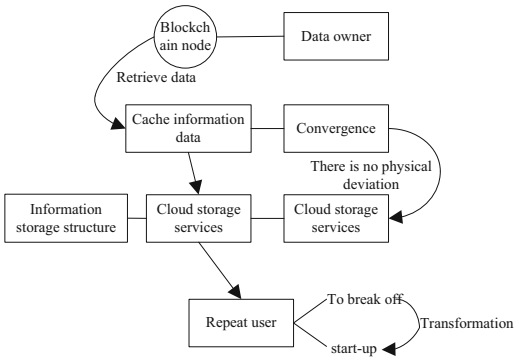

Fig. 1. Cloud storage encrypted data anonymous De-duplication model

Generative counterwork requires fast access to data files when using cloud storage servers against data owners and deduplication users in a network. The cloud storage server is responsible for securely storing user data files and is able to respond correctly to user requests. Trusted intermediaries are intended to enable anonymous channels, and all cloud storage server and user communications are forwarded through Trusted Intermediaries to hide user identity [4]. In order to ensure that the cloud storage server can correctly respond to the user's request for retrieving files, and at the same time can not get any information about the user, it is necessary for users to negotiate with each

other to obtain a digital certificate between the generated counterwork network and the cloud storage server, so as to realize the efficient anonymous data file storage.

Asymmetric encryption of a generative counterwork network requires a combination of public and private keys, one for encryption and the other for decryption. Private keys are highly secure, so to be kept, public keys are generally public [5]. Asymmetric encryption is characterized by long time and slow speed of encryption and decryption, so it is only suitable for a very small amount of data encryption. Otherwise, the entire transmission process will be prolonged due to encryption speed. Expression of asymmetric encryption.

Ciphertext:

$$c = Enc_{publicKey} \times m \tag{1}$$

In the formula, $Enc_{publicKey}$ represents the asymmetric encryption key.

Written:

$$m = Dec_{\text{private key}} \times c \tag{2}$$

In the formula, $Dec_{\text{private key}}$ means to write the encryption key.

Symmetric key:

$$m_n = Dec_k \times c \tag{3}$$

In the formula, Dec_k represents symmetric encryption key.

In a generative countermeasure network, two types of attackers are considered:

1) Internal attacker: refers to an opponent within the system, mainly refers to the cloud server. The cloud server is honest and inquisitive, allowing arbitrary access to the user data it stores.
2) "External attacker" refers to an opponent outside the system, mainly refers to an unauthorized user. Access by external adversaries to information about partially uploaded data through eavesdropping on public channels, the main purpose of which is to illegally obtain clear text information about user data stored on cloud servers

Generative countermeasure of encrypted data under the network to re-secure targets as follows:

1) Data privacy: The De-duplication scheme shall ensure the privacy of user data stored on the cloud server, including non-popular data and popular data. The cloud service should not get any clear text information about the user data it stores. Unauthorized users cannot get clear text information about the user data stored on the cloud service.
2) Data integrity: The De-duplication scheme shall ensure the integrity of user data stored on the cloud server. The scheme allows authorized users to verify the data integrity when downloading data.

Anonymous De-duplication schemes include the following steps: proof of full user ownership; encrypted data De-duplication; and user digital certificate negotiation [6]. In

order to reduce the redundancy of the data file, the process of judging and proving the ownership of the data file must be carried out in clear text, and the trusted middleman must be introduced to hide the identity information of both sides [7]. At the same time, users can use digital certificates to retrieve files, ensure that only legitimate users can retrieve data, and hide the relationship between users and data. The De-duplication model of anonymous encrypted data is shown in Fig. 2.

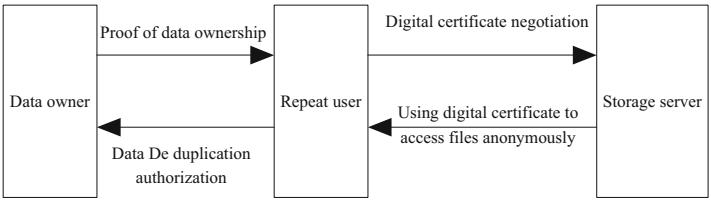

Fig. 2. Anonymous encrypted data De-duplication process

Suppose that the user in the generative countermeasure network encodes and preprocesses the data file and encodes the original file with erasure code. The bilinear mapping is used in the scheme. Assuming that the order of the curve filled in the massive data space is M, the S space range of the massive data set can be divided into $2^M \times 2^M$ grids, and each grid has four-dimensional spatial coding

$$M_0 = \left[\log_2 \frac{D_0}{H_1}\right] \tag{4}$$

In the formula: D_0 represents the total amount of data; H_1 represents the storage size of data block. Statistics coding data element information set K', assuming that the total number of data coding blocks is I, if the data block storage size H is greater than the maximum threshold percentage of massive data blocks, then the coding block should be divided into sample set k. Based on the linear filling curve aggregation characteristics of spatial four-dimensional Hilbert code, the massive data coding blocks are decomposed, and the corresponding storage sequence of each coding block H_2 is marked to form the corresponding spatial data partition matrix as follows:

$$F = \begin{vmatrix} H_{2\text{de}0} & H_{1a0} & s_0 \\ \cdots & \cdots & \cdots \\ H_{2\text{de}n} & H_{1an} & s_n \end{vmatrix} \tag{5}$$

In the formula: $H_{2\text{de}n}, H_{1an}, s_n$ represents the spatial elements of massive data in the generative countermeasure network. According to this element, the corresponding spatial code and the corresponding mass data block storage label after matrix matching are obtained, so as to complete the division of massive data security elements, and optimize the privacy of cloud storage encrypted data according to different types of security elements.

2.2 Improved Solution Security Improvements

In the process of uploading and downloading files, attackers break the data privacy in two ways. Specifically, in file uploads, attackers use hash proof attacks to trick cloud storage servers into downloading private data [8]. In file access, an attacker accesses unauthorized de-reduplicate data. Finally, the loss of shared data blocks in the event of a device failure reduces system availability, meaning that a large number of referenced files are lost and data is unrecoverable. Table 1 shows the main security problems of data De-duplication in cloud storage system.

Table 1. Key security issues for data de-emphasis in cloud storage systems.

Safety objectives	Political attack or problem	Specific implementation
Confidentiality	Brute force attack	Comparison between generating ciphertext well by traversing plaintext set and stealing ciphertext
Privacy	Hash proof attack	Attackers use fingerprints to cheat the server and download private data
	Unauthorized access	An attacker attempts to access private data that does not belong to his/her own rights
Usability	Shared data loss	Device failure results in the loss of shared blocks, aggravating the unavailability of files

Based on Table 1, the security threats faced by internal and external attackers of data De-duplication cloud storage system in Generative countermeasure network are analyzed. In order to protect data confidentiality, users will use their own key to encrypt data and generate different ciphertexts, so duplicate data cannot be found [9]. Convergence encryption *CE* is used to support the De-duplication of ciphertext, where the key is the data hash value. However, there are brute force attacks on convergent encryption, especially on low entropy files. That is, if the attacker knows that the ciphertext C of the target file D is in a set of known size n, the attacker can recover D from the set S by offline encryption. For each element, D attacker uses convergence encryption to get ciphertext C, and obtains plaintext D data De-duplication compared with $C = C_1$. Cloud storage system is faced with device failure (for example, disk error), which will lead to shared data loss and data unavailability.

1) In the data deduplication cloud storage system, only one duplicate data block is kept, and in case of equipment failure and loss of shared data, multiple files will be unreadable.
2) Devices used in cloud storage systems (e.g. disks) are also at risk of mechanical failure, resulting in loss of data and interruption of system services.

The data de-reencryption method generates a key for each data block, so the key security and clutter degree are typical data de-reencryption key management algorithms.

The following is an introduction and analysis of typical key management methods and their problems, including single-key server, master key management and secret sharing, users losing control over the privacy data in the data de-cloud storage system and data sharing among users, so the problems of unauthorized access in the storage system are more serious. This paper summarizes the research status of three typical methods of data de-re-access control and analyzes their advantages and disadvantages, including proxy re-encryption, attribute encryption and key tree encryption.

Data storage security, i.e., content security, is similar to transmission security, which guarantees the security of all data passing through the network, while content security guarantees the security of the user's data after persistence. If the user's data is stored on the cloud storage side, and how to ensure that the cloud storage side is agnostic to the user's content, users are required to encrypt their data for storage [10]. Combining the high security of asymmetric encryption with the high efficiency of symmetric encryption, users generate symmetric keys to encrypt their own data files symmetrically, and then take symmetric keys as additional properties of data files to be protected by asymmetric encryption with their own public keys [11]. Finally, the symmetric keys of users and the encrypted data files are handed over to the cloud for management. Because the private keys of users are unknown, the cloud cannot decrypt the symmetric keys of users and thus cannot obtain the clear text information of user data files [12, 13]. Based on this, the network De-duplication data backup instructions are standardized, as shown in Table 2.

Table 2. Network redo data backup instruction.

Send De-duplication instruction	Network De-duplication data backup execution
AT*DSO	Send the backup instruction of network De-duplication data to modem and wait for receiving
ATE0	Network De-duplication data backup arrangement, received instruction sequence number
AT XS01	Set the network De-duplication data carrier signal, and modify the signal change parameters
AT*W*DO	Configure the implementation content of De-duplication network data backup, and store all data in the database

On the one hand, the security of data storage in the cloud depends on the encryption of the client to prevent the leakage of data files, on the other hand, it depends on the distributed storage system in the cloud, such as distributed file system and distributed protocol [14, 15]. In order to ensure the integrity of data files, the cloud also needs to do integrity check and necessary redundant backup. In the cloud storage environment, the data transmission efficiency can be improved by increasing the effectiveness control of data link communication. However, in the actual transmission process, the time limit of data link communication is limited, so it is necessary to constantly adjust the data link communication, change the data transmission integrity, and make statistics The communication time limit of fixed information data link t', the real-time transmission time t'_s of fixed information data link and the shortest transmission time of a certain data

link with fixed information volume $t'_{\min}$. The relationship among as, t', t'_s and $t'_{\min}$ is analyzed, as shown in Table 3.

Table 3. Cloud storage encrypted data reload scheme.

Relationship	De-duplication scheme		
$t'_s < t'$	No control measures to eliminate duplication		
$t'_{\min} < t' < t'_s$	Changing the frequency of data transmission	Adjust transmission rate	Take anti-interference measures
$t' < t'_{\min}$	Replace data link network		

According to the quantitative content of the above-mentioned security evaluation indexes, the security of different network information transmission is tested to obtain the maximum read, write and service permissions of the host. The vector of network security confidentiality is $A_z = (A_z(c))_n$, the vector of network security integrity is $A_v = (A_v(c))_n$, and the vector of network security availability is $A_t = (A_t(c))_n$. The positive ideal criterion of the evaluation index is: $A_0^+ = \left\{A_0^+(1), A_0^+(2), \cdots, A_0^+(n)\right\}$, which is used as the evaluation data to obtain the optimal evaluation results of each index; The negative ideal standard of evaluation index is: $A_0^- = \left\{A_0^-(1), A_0^-(2), \cdots, A_0^-(n)\right\}$, take this standard as the evaluation data, and then get the worst evaluation result of each index. Through the establishment of the standard, we can get the best and worst base point, and compare the distance between different evaluation values. Furthermore, the evaluation of data De-duplication security theory is described. The data De-duplication security theory is shown in Table 4.

Table 4. Data de-emphasis security specification.

Level	Explain
Extremely unsafe	The security guarantee ability of De-duplication data transmission is poor, and the security situation is severe
Unsafe	The security guarantee ability of De-duplication data transmission is limited, and there are security risks
Safe	The De-duplication data transmission has certain security guarantee ability, and the environment is basically safe
Very safe	De-duplication data transmission has a strong security capability, and the environment is very safe

Users encrypt their own data files to prevent the cloud storage server from misusing their sensitive data, and must ensure that the cloud storage server cannot get the contents of user files by any means. Users store files in the cloud storage server through anonymous channels and retrieve them through digital certificates. These operations can hide the

user's identity information from the cloud storage server. Therefore, even if the cloud storage server and legitimate file owners attack, the encrypted files provided by the cloud storage server cannot provide any additional information. The trusted middleman is the proxy of the proxy reencryption protocol, so the communication information encrypted by the trusted middleman can not get the user file and key information. In this scheme, in order to reduce user information leakage, users need trustworthy intermediaries to hide identity information, and users can't authenticate directly. So trustworthy intermediaries in the system are a more powerful attacker.

2.3 Implementation of Cloud Storage Encrypted Data Deduplication

As one of the participants of the deduplication process, the cloud is mainly responsible for the deduplication of the information communication between the user and the verification service group, and the user's privacy between the user and the data holder can be amicably hidden by such third-party forwarding of the user's information. However, since the public key of the user and the user's information will be accessed by the cloud during the communication process, it may be assumed that the cloud server performs its own tasks according to the agreement, but it may not be reasonable for the cloud to obtain the user's data information through legitimate means based on the information already obtained, which is called "honest but curious". The security concerns require that the communication between users is invisible in the cloud, and the identity authentication of users is carried out by the cloud, and this forwarding mode has hidden man-in-the-middle attack. Convergent encryption is used to ensure the security of the information and the identity of the user based on the data owned by the user and the authenticator.

Data De-duplication can be divided into file level, block level and byte level; generally speaking, the smaller the granularity, the more duplicate data can be found, but the calculation cost is also higher. According to the range, data De-duplication can be divided into local and global De-duplication: local De-duplication can only occur in a single node and part of users, and its calculation and memory costs are less, but the less redundant data is found; global De-duplication can be achieved between multiple nodes and multiple users, which can obtain better compression ratio, but the computational and memory overhead is also higher. According to the time, data De-duplication can be divided into online and offline De-duplication; both online and offline De-duplication can reduce the storage overhead, and online De-duplication occurs before the data is written to the storage device, which may affect the performance of the system. According to the location, data De-duplication can be divided into target end and source side De-duplication. Target side De-duplication is called server side De-duplication, which can only reduce storage overhead. Source side De-duplication is also called client side De-duplication, which can not only reduce the storage cost of client, but also save transmission bandwidth. The data De reordering optimization is shown in Fig. 3.

Generally speaking, block level De-duplication can identify and eliminate more fine-grained redundancy, and the system throughput is high, so block level De-duplication is more widely used in storage systems. Data block is divided into fixed length block and content-based block; secondly, the calculation is to obtain the hash summary of the data block as the identifier; then index query uses the index based on locality and similarity to improve the retrieval speed. Finally, data management mainly includes

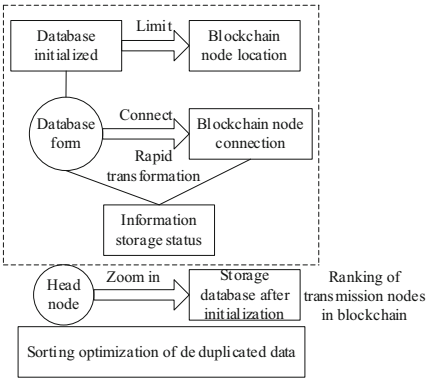

Fig. 3. Data de-sequencing optimization processing steps

container read-write and recovery performance. The storage layer is the bottom layer of cloud storage, which is composed of different storage devices and complex network devices. In addition, there is also a set of storage management system for centralized management, status monitoring, maintenance and upgrading of hardware equipment. In the environment of big data cloud storage, if the output rate of node is equal to the input rate, the output of node i data is determined by $i + 1$. According to the relationship between the input and output of nodes, the transmission control mode of single node is obtained, and the data De-duplication steps are optimized based on the above principles, as shown in Fig. 4.

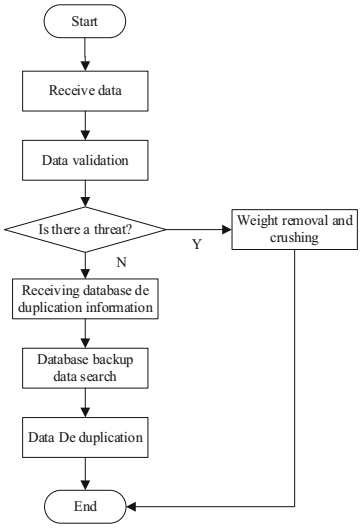

Fig. 4. Data removal restep optimization

In order to save the network bandwidth, this scheme only requires the initial uploader to upload the file data, and the subsequent uploader only needs to verify the file without uploading data. Moreover, this scheme has higher flexibility, if the user has higher security requirements, this scheme can be improved to server-side data deduplication with little cost, and can effectively resist side channel attack.

3 Analysis of Experimental Results

The experimental environment is set up as follows: fix the nodes on the 120 m × 120 m plane, the number of nodes is 80, and the running PC is configured as Pentium (R)4CPU2.40 GHz. The computing operation on the elliptic curve (prime order of the elliptic curve |p|=256bit) is implemented by using the open cryptography library Miracl, and the BLMLE scheme is simulated by using C++ language on the Windows10 operating system, VisualStudio2008, as well as the scheme in this paper. The experimental device's processor is InterCoreTMi5-7200U, the CPU is 2.50 GHz, and the memory is 8 GB. The HASH algorithm uses SHA-256 and the encryption key uses 256bit AES encryption. The experimental parameters are set as shown in Table 5.

Table 5. Experimental parameter settings.

Field	Type	Explain
ID	Short	Host identification
Name	Character string	Host name
Ip	Character string	IP address
YellowAlarm	Short	Warning value

The tests used synthetic and real data sets, respectively. The synthetic data set contains files populated with randomly generated content, and each file is divided into fixed-length data blocks. Fslhomes contains a mirror of the user's home folder, including source code, binaries, documents, and virtual machine images.

Because the label calculation of bl-mle needs to segment the ciphertext block and then perform S-TIMES exponential operation, the efficiency of bl-mle is very low. In the experimental environment, if $p = 256$, it takes $s = (2 \times 1024)/(256/8) = 64$ exponential operations to calculate the tags corresponding to 2 KB data blocks, 128 times for 4 KB blocks, and so on for other sizes.

As can be seen from the Table 6, the label computation overhead of the proposed scheme is much less than that of the BL-MLE scheme. This is because the hash operation is much faster than the exp calculation on the elliptic curve, although this scheme needs two more hash operations. So as the block size increases, the number of exp calculations increases, and the time gap between the two schemes becomes larger. For this project, the number of calculations is fixed, the increase in the size of data blocks will only bring about an increase in AES and hash overhead, and a 16 KB of hash or AES only needs about 1 ms, an exp calculation takes about 20 ms, so the calculation time of this project

Table 6. Computational overhead for generating block labels.

Method	Block size			
	2 KB	4 KB	8 KB	16 KB
BL-MLE method	1406	2496	5609	11154
The method of this paper	46	48	49	52

increases slightly with the increase of data blocks, which proves that this method has a relatively good effect in the actual application process, further compare the accuracy of the application of the two methods, and record, specifically, as shown in Fig. 5.

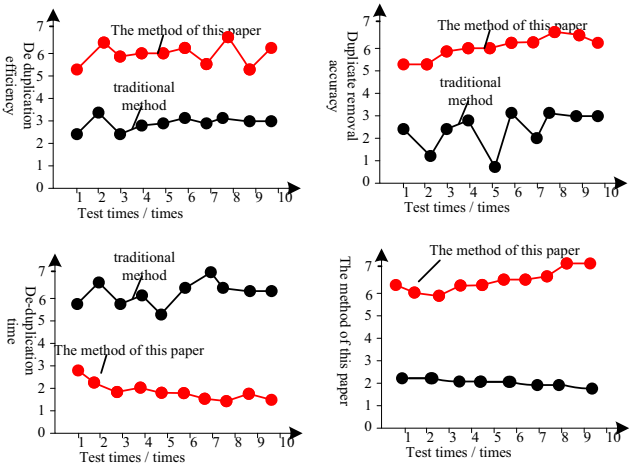

Fig. 5. Data reduplication efficiency comparison

According to the detection results in Fig. 5, the depth and accuracy of cloud storage encryption data De-duplication method based on generative adversary network proposed in this paper are obviously better than traditional methods in the actual application process, and the time consumption is relatively short. Therefore, it is confirmed that the cloud storage encryption data De-duplication method based on generative adversary network fully meets the research requirements.

4 Conclusions

In this paper, some shortcomings of existing cloud storage De-duplication schemes are improved, and a new De-duplication scheme is proposed. Compared with the original scheme, it not only saves the advantages of the original scheme, but also improves the computational efficiency of block label generation in client and block label comparison in cloud storage. At the same time, the file partition method of the new scheme is more

flexible. Compared with the traditional scheme, this scheme has obvious advantages in computing efficiency, and is more suitable for the existing cloud storage system.

References

1. Cui, H., Deng, R.H., Li, Y.: Attribute-based cloud storage with secure provenance over encrypted data . Futur. Gener. Comput. Syst. **79**(2), 461–472 (2018)
2. Hu, C.: Calculation of the behavior utility of a network system: conception and principle. Engineering **4**(001), 78–84 (2018)
3. Feng, X., Su, X., Shen, J., et al.: Single space object image denoising and super-resolution reconstructing using deep convolutional networks. Remote Sens. **11**(16), 1910–1915 (2019)
4. Zhu, F.: Dynamic channel allocation method for emergency communication network in vehicle networking. J. Xi'an Polytechnic Univ. **033**(003), 296–301 (2019)
5. Zhang, J., Ou, P.: Privacy-preserving multi-receiver certificateless broadcast encryption scheme with de-duplication. Sensors **19**(15), 3370–3378 (2019)
6. Hochberg, G.K.A., Shepherd, D.A., Marklund, E.G., et al.: Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions. Ence **359**(6378), 930–935 (2018)
7. Asgari, N., Ayoubi, S., Jafari, A., et al.: Incorporating environmental variables, remote and proximal sensing data for digital soil mapping of USDA soil great groups. Int. J. Remote Sens. **41**(19), 7624–7648 (2020)
8. Liu, S., Bai, W., Zeng, N., et al.: A fast fractal based compression for MRI images. IEEE Access **7**(99), 62412–62420 (2019)
9. Bandi, T., Justin, J., Rinesh, S., et al.: Efficient client-slide deduplication of encrypted data with public auditing cloud storage. Test Eng. Manage. **82**(1), 10425–10430 (2020)
10. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (2019)
11. Chengetanai, G., Osunmakinde, I.O.: QUACS: routing data packets in ad hoc networks on buffer-constrained load balancing conditions during emergency rescue crisis. Wireless Pers. Commun. **99**(10), 1–31 (2018)
12. Liu, S., Li, Z., Zhang, Y., et al.: Introduction of key problems in long-distance learning and training. Mob. Networks Appl. **24**(1), 1–4 (2019)
13. Inayat-Hussain, S.H., Fukumura, M., Muiz Aziz, A., et al.: Prioritization of reproductive toxicants in unconventional oil and gas operations using a multi-country regulatory data-driven hazard assessment. Environ. Int. **117**(4), 348–358 (2018)
14. Simpson, S.L., et al.: A mixed-modeling framework for analyzing multitask whole-brain network data. Network Neurosci. **3**(2), 307–324 (2019)
15. Ramadhani, E.H., Kabetta, H., Amiruddin, A.: Exploration of the security of free data encryption applications for cloud storage. IOP Conference Series: Materials Science and Engineering **1007**(1), 12–18 (2020)