Research on Data Security Audit in Cloud Computing for Big Data Environment

Li Shuanbao*
School of Artificial Intelligence, Henan Finance University,Zhengzhou, 450046, Henan, China
**Corresponding Author*

Abstract

The modernization of industrial industry cannot be separated from the development of big data.In order to meet this challenge, cloud data integrity audit has been proposed in recent years and received extensive attention. Based on the in-depth study of the impact of different cloud storage data types on the audit scheme, this paper proposes an audit scheme based on Dynamic Hash table.Based on this, this paper explores a variety of cloud storage audit algorithms for different data types to deal with different security challenges.Facing a series of data security problems brought by cloud computing, this paper analyzes the concept, working principle and characteristics of cloud computing, and discusses the data security risks brought by cloud computing from four aspects. At the same time, this paper elaborates the data security strategy from five aspects: data transmission, data privacy, data isolation, data residue and data audit. In this paper, we propose to adopt end-to-end data encryption technology, build private cloud or hybrid cloud, share table architecture, destroy encrypted data related media, and introduce third-party certification authority for data audit.

Keywords*: Cloud data, industry, data audit, data security, dynamic hash table.*

I. Introduction

In recent years, the first mock exam model is based on the demand supply of computing resources, which makes cloud computing applications more and more widely. It has led the trend of intensive, large-scale and specialized information technology, and has become a major revolution in the field of information technology [1]. Among them, cloud storage is a new branch that extends and develops from the concept of cloud computing. Its goal is to use cloud computing technology to coordinate a large number of different types of storage devices to provide data storage services [2]. Cloud storage provides storage resources as services to users through the Internet, which is an important form of infrastructure as a service (IAAs) in cloud computing [3]. At present, the popular cloud storage services include Dropbox, Google box, baidu disk, etc. these services provide a solution for enterprises and individuals to keep and efficiently access massive data. Due to the large storage capacity and low cost of cloud, more and more organizations and individuals tend to host large-scale data to cloud service provider (CSP).

However, the characteristics of data outsourcing in cloud storage environment bring great challenges to data storage security and data management [4]. In recent years, with the rise of cloud storage, cloud security incidents emerge in an endless stream [5-6]. For example, the Amazon cloud accident in 2011 not only caused many companies' servers to stop running, but also caused 0.07% of users' data loss [7]. It can be said that the core issues of data storage security and privacy protection have seriously affected users' confidence in cloud storage services. On the one hand, users can't control the access and use of data in a conventional way because of the data stored in the cloud. On the other hand, CSP may have dishonest behaviors driven by interests, such as concealing data loss information from users, deleting customer data to save space resources, and disclosing user privacy. According to the latest research report on the development of China's IDC industry, the main objective factor restricting the domestic cloud storage market is "data security issues".

The security issues typically represented by data storage security and privacy protection are hindering the

ISSN: 0010-8189

promotion and development of cloud storage. Therefore, the need to audit whether the data stored in the cloud is accurate and complete arises at the historic moment, and data integrity verification for untrusted cloud service providers has also become a focus in the field of cloud storage research [8-10]. As a small branch of information security issues, data integrity audit is closely related to cryptography, which is the cornerstone of solving information security issues. This is the reason why the security audit technology, which is supported by the relevant cryptography theory and aims at the verification of cloud data integrity, has made great progress in recent years.

However, there are obvious shortcomings in simply placing the audit process between the user and CSP: first, as both sides of the transaction, there is an interest relationship between the user and CSP, and any interested party as the audit role will affect the fairness and authority of the results; Second, the audit work needs to be carried out frequently and regularly, which will bring great extra cost to users, especially for mobile intelligent terminals (such as mobile phones, tablet computers, etc.). Therefore, with the help of a trusted third party, we need to conduct an objective and fair integrity audit on the data stored in the cloud, reconcile the trust contradiction between users and CSP, and reduce the burden of users, so as to finally create a healthy and trusted cloud storage environment.

II. Cloud storage data integrity audit model and objectives

2.1Audit model

From the existing literature, cloud storage data integrity audit model can be divided into private audit model and public audit model. The former prototype is remote data integrity verification, and the private audit model (as shown in Figure 1) contains two entities: CSP and user. CSP manages and cooperates with large-scale cloud servers to provide stable and efficient data outsourcing services. Users are users of cloud storage services. They want to host data to CSP to minimize the local storage and computing costs. However, due to the outsourcing of data to the cloud, users want to be able to verify and confirm whether the cloud holds their data completely at any time. Therefore, in this model, users act as auditors, CSP is required to provide evidence of data integrity and audit results. The audit model is relatively simple, but there are obvious shortcomings: first, users as users and CSPs as service providers are both sides of the transaction, and there is an interest relationship between them, and the audit conducted by the interested party will affect the fairness of the results. Second, the additional cost of audit work is hard for users (especially for mobile end users, such as mobile phones and tablets). Therefore, in order to achieve the fairness and efficiency of cloud storage data integrity audit, the existing schemes generally adopt the public audit model based on the third party (as shown in Figure 2), that is, to entrust the audit work to a trusted third party, It not only ensures the credibility and authority of audit results, but also transfers the user's computing overhead. The public model in the figure consists of three entities: cloud service provider (CSP), user and Third Party Auditor (TPA). As mentioned earlier, CSP is the provider of cloud storage services. Users are users of cloud storage services (including data owners (DOS) and data users (DUS). Users can be organizations or individuals). However, the audit task originally completed by users is now handed over to TPA, that is, audit the integrity of cloud storage data regularly or at the request of users, and feed back the audit results to users.

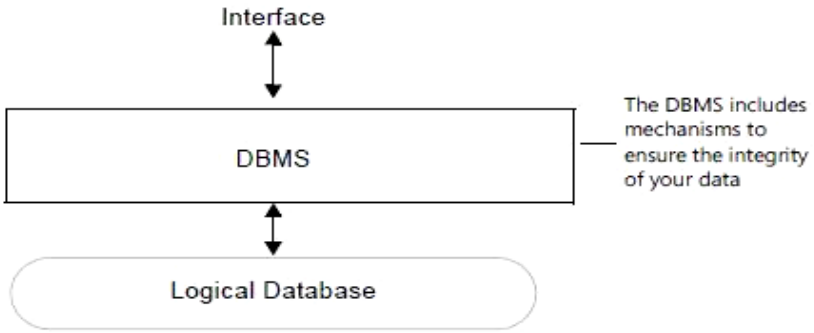

Figure 1 Private audit model of cloud data integrity

2.2Achieving goals

In the public audit of the integrity of cloud storage data, first of all, we assume that tPA can provide reliable audit results, but in order to protect user privacy, the verification scheme must ensure that tPA cannot obtain any information about the data content in the audit process. Then, CSP is usually regarded as untrustworthy in the design of audit scheme, especially when the data integrity is damaged due to various accidents, CSP may launch the following attacks in order to pass the audit

(1) Forgery attack: CSP forges the audit information needed by TPA in order to conceal the loss or destruction of data;

(2) Substitution attack: CSP uses the correct part of the data file to replace the wrong part to generate relevant audit information for verification;

(3)Replay attack: when the data is wrong, CSP attempts to feed back the audit information generated in the past to TPA to pass the audit;

(4) Collusion attack: in the case of multiple servers collusion in CSP, or in the case of multi-user sharing, CSP colludes with users whose access rights have been revoked to forge data or audit information. Obviously, in order to achieve secure audit, cloud storage data integrity audit scheme should be able to effectively resist the above attacks.

In addition, from the aspect of audit function and efficiency, the ideal audit plan should also achieve the following goals:

(1) Privacy protection: the audit scheme should ensure that tPA cannot obtain any data information involving privacy.

(2) Support batch audit: to improve audit efficiency, TPA should be able to respond to multiple audit requirements from different users and different clouds at the same time.

(3) Support multi copy audit: in the case of multi copy storage, the audit scheme should ensure that all copies stored in CSP are complete and correct.

(4) Support dynamic data audit: the audit scheme should support the dynamic update of data, and be able to effectively verify the integrity and freshness of data.

(5) Cost minimization: while ensuring the correctness of audit results, the audit scheme should reduce the computation and communication costs as much as possible.

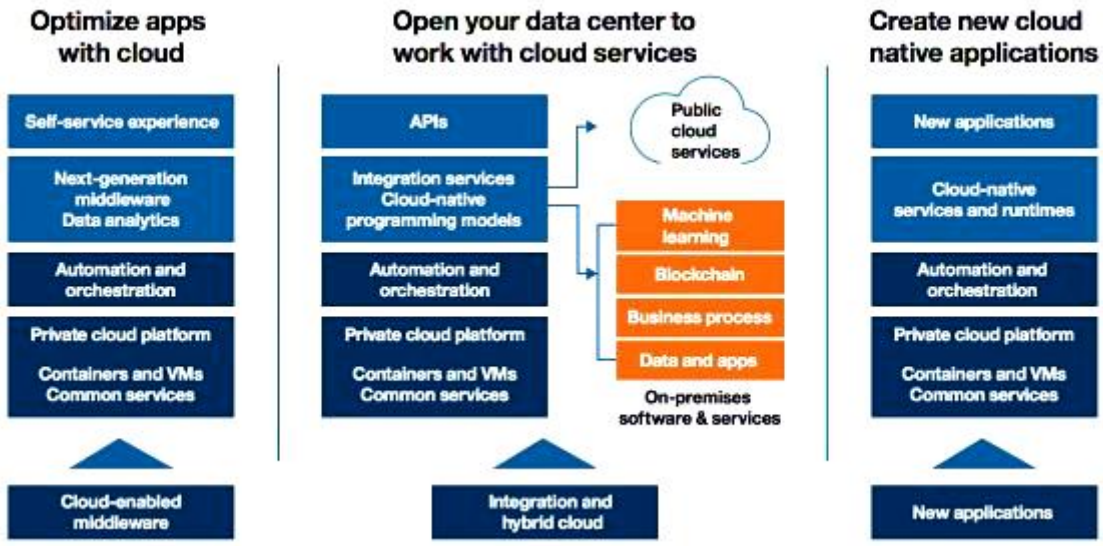

Figure 2 Public audit model of cloud storage data integrity

III. Integrity audit based on data classification

3.1 Common signature technology in cloud storage environment

With the continuous development of cloud storage technology, the amount of data stored in the cloud is increasing, and the audit demand is also increasing. In order to reduce the communication cost, storage cost and computing cost in the audit process, signature technology with different characteristics is constantly applied in the audit program.

(1) Message authentication code (MAC)

Cloud data integrity audit as a major trend, not only need to reduce the cost of the audit process, but also should ensure the privacy of the audited data. It is obviously extremely inefficient and unsafe to audit the data back to the auditor. Therefore, the implementation of audit without data retrieval is one of the basic requirements of cloud data audit scheme. In the initial non retrieval audit, message captcha is used to achieve the above goals. The main idea is that the user generates a series of message captcha including randomly selected message captcha secret key and data file for the auditor in advance. Each time the audit operation is performed, the auditor sends a key of the verification code to CSP, and uses the given key to verify the message verification code returned by CSP. The use of message verification code reduces the communication cost of each audit. However, in this type of scheme, the user key is needed for Mac verification, so the scheme can only be used as the auditor, which is difficult to be extended to public audit; Moreover, the number of MAC keys selected in advance is limited, and the key re generation process after use will bring huge computational overhead to users.

(2) RSA signature

In view of the shortcomings of message verification code, the audit scheme based on RSA signature technology is proposed by researchers, which mainly uses the homomorphism of RSA signature to save the computational and communication costs for users. Compared with the MAC based audit scheme, the RSA signature based audit scheme not only does not need to retrieve data, but also does not limit the number of audits, reducing the burden of users in communication and computing. However, the length of label based on RSA signature is related to the security factor of signature, and the communication overhead and storage overhead still need to be improved.

(3) BLS signature

In order to save the storage cost of data block label, researchers found that BLS signature instead of RSA signature can reduce the length of data block signature under the same security factor, so as to achieve the purpose of reducing the communication cost and storage cost in cloud data audit scheme. The cloud storage data audit scheme using BLS signature technology mainly uses the nature of bilinear mapping. Its specific content and implementation method have been mentioned in the second chapter, which will not be repeated here. Compared with the audit scheme using RSA signature technology, the length of tags generated by BLS signature is shorter, which is related to the length of parameters, thus reducing the storage of tags. However, the label generation process of BLS signature is more complex, which requires more computational overhead than the former two kinds of signatures.

(4) Algebraic signature

Algebraic signature is applied as a new signature technology in cloud data integrity audit. The main idea is: the auditor gives a tuple over a finite field, which is a vector composed of multiple elements. CSP can aggregate the data block and its algebraic signature, and then send them to the auditor. If the aggregate value of the data block matches the aggregate value of the algebraic signature, the audit will pass. Since the algebraic signature itself is a string, the generation of signatures and the operations between signatures are actually logical operations between binary systems. Therefore, the generation time and length of the algebraic signature are related to the size of the signed file. Compared with BLS signature technology, with a certain file size as the critical value, its generation time and length will be better than BLS signature technology. In the current cloud data integrity audit scheme, BLS signature and algebraic signature are more commonly used.

3.2 Characteristics and differences of cloud storage data

The contents of the files stored in the cloud are various and have remarkable characteristics. According to the latest cloud storage industry and user behavior survey report, it can be found that the types of files stored in the cloud are different, but the main part is documents, movies, videos and photos. Therefore, after analyzing the most common data features of the above three categories, this section divides them into archive data (multimedia files, PDF files, etc.) and frequently updated data (working documents, etc.), and discusses the impact of their respective characteristics on audit methods.

(1) Archive data, also known as infrequent updating data (such as pictures, audio, video, etc.), is one of the common data types uploaded to the cloud.

(2) It occupies large-scale storage space, and files are generally above MB level, most of which exceed GB level.

(3) The amount of updates is very small, and the files are generally uploaded once, and there is no modification in the follow-up. Therefore, when considering the integrity audit scheme of archived data, it is necessary to take the data quantity as one of the key points, that is, the audit scheme of archived data should not be affected by the amount of data.

According to the characteristics of different signature technologies in the previous section, we can get:

(1) The calculation amount and length of BLS signature are related to its security coefficient, but not affected by the size of the signature block. If the size of the archive data block is increased and the number of blocks is reduced, the effect of the data size on the efficiency of BLS signature will be reduced.

(2) The calculation and length of algebraic signature are proportional to the size of signature file. Therefore, no

matter how the archive file is divided into blocks, the file size will have an impact on the efficiency of algebraic signature.

(3) Because the amount of update of archived data is small, the re generation of data block tags is not much, and the complexity of label generation calculation can slow down the requirements. And it is not necessary to segment data blocks to support high frequency data update. Based on the above three points, we can get that the integrity audit scheme of archived data is more suitable for BLS signature technology.

The frequently updated data, such as working documents and microblog data, has been increasing rapidly with the development of the Internet. Such data has the following characteristics:

(1) The update speed is fast. Take Weibo data as an example: according to the survey report issued by China Internet Center, the current user scale of microblogging is 242million and the average daily active user is about 75million. Assuming that the average active users update the microblog once a day, the daily update will reach 75million times.

(2) The granularity of the update is small. The granularity of each update will not exceed 1MB, whether it is e-commerce records, official documents, etc. Take microblog data as an example, with the limit of the number of words published, the granularity of each update is no more than 50kb.

Therefore, the above characteristics need to be considered when choosing the integrity audit scheme of frequently updating data. The existing dynamic data audit scheme takes the data block as the update granularity, that is, all the adding and deleting operations must take the data block as the minimum unit. Therefore, the frequently updated integrity audit scheme should support the segmentation of the data block and optimize the calculation cost of generating the label of the data block.

According to the description of BLS signature technology and algebraic signature in the previous section, it can be obtained as follows:

(1) The length of BLS signature and calculation cost are not sensitive to the size of data block. Therefore, the smaller the data block (i.e. the finer the update granularity) is, the more expensive the BLS signature is. Because the process of BLS signature generation is complex, the computing cost is less than other schemes.

(2) The computational cost and length of algebraic signature are sensitive to the size of the file. Therefore, the block of data block of the same capacity has little effect on the calculation cost and length of algebraic signature. The generation process of algebraic signature is related to the length of data block, so the smaller the data block is, the smaller the calculation amount of label generation is.

Based on the above two points, we can see that when the file size is equal, the finer the data block is, the greater the influence on BLS signature is, but it has no influence on the algebraic signature of the whole file. Therefore, the frequent updating data related to the actual update requirements is more suitable for the use of algebraic signature technology. In short, this scheme conducts classification audit on user uploaded files, and the archiving files adopt BLS based audit scheme, and frequently update files adopts audit scheme based on algebraic signature to reduce the storage amount of data block tags and reduce the calculation amount of label generation. Meanwhile, the dynamic hash table of data structure shall be used in the audit of frequently updating data, Achieve efficient data update (the system model is shown in Figure 3).

ISSN: 0010-8189

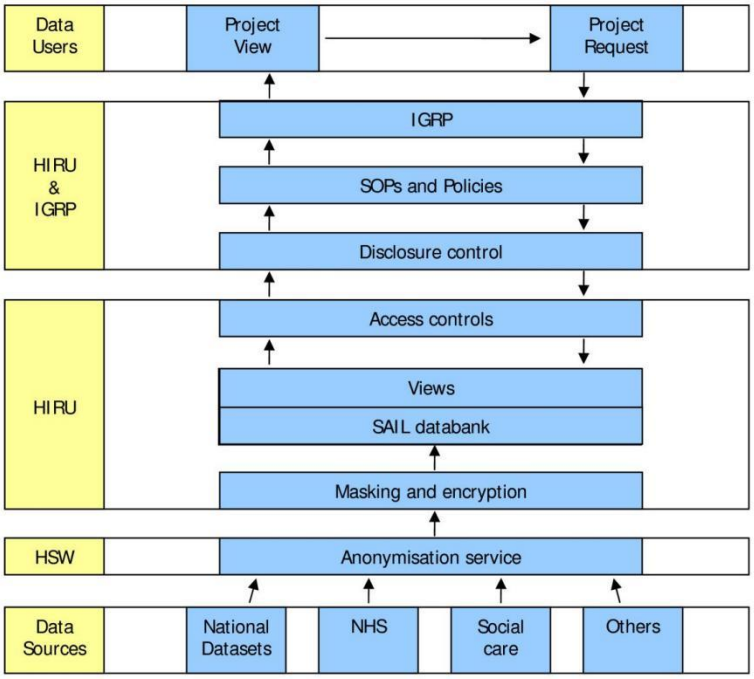

Figure 3 Data audit system model diagram

3.3 Performance analysis

According to the theoretical analysis of BLS signature and algebraic signature, we can get the following conclusions:

(1) The storage capacity of label generated by BLS signature technology is directly proportional to the number of data blocks, while the storage capacity of label generated by algebraic signature technology is directly proportional to the file size.

(2) There is no requirement for the partition of data blocks in archive data, and the requirement for dynamic update data is related to the actual update situation.

Therefore, in this section, the number of data blocks and the size of data blocks are taken as variables to compare the storage cost of the two signature techniques, and the parameters that can achieve the best efficiency are obtained for the audit scheme. In order to reflect the impact of file data block partition on the storage capacity of BLS signature, we compare the storage cost of BLS signature when 1GB files are divided into different data blocks. As shown in Figure 4, the storage capacity of BLS signature decreases significantly with the increase of data block size.

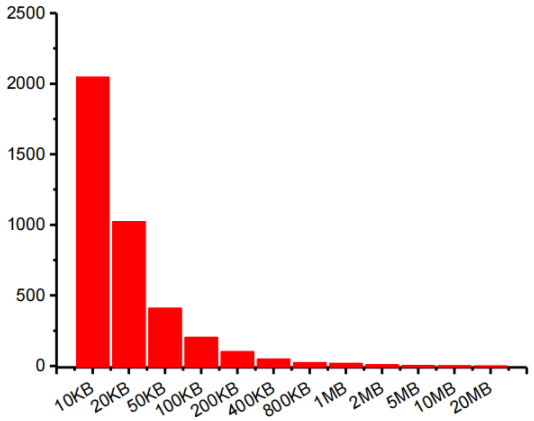

Figure 4 Label storage overhead changes with data block size

Then, in order to compare the influence of different block sizes on BLS signature and algebraic signature, we give the storage capacity of BLS signature and algebraic signature with the change of block size. As shown in Figure 5, the size of algebraic signature is directly proportional to the size of data block, while the size of BLS signature is not affected by the size of data block. Based on the above theoretical derivation, the simulation experiment is given. The basic environment of this simulation experiment has been given in the previous paper, and will not be repeated here.

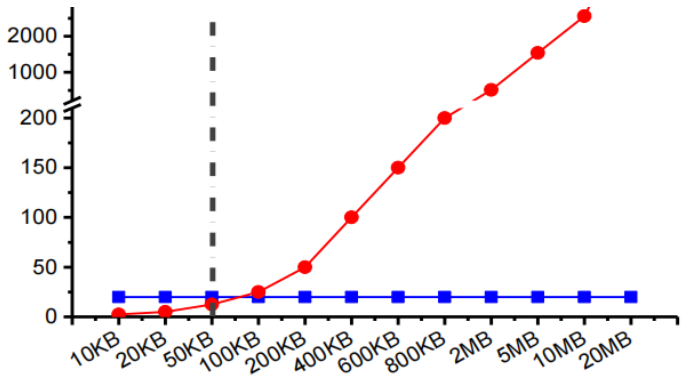

Figure 5 Experiment result

What needs to be explained here is: the prototype of algebraic signature is a string, assuming that its signature length is l, then its calculation is based on a finite field GF composed of 2l different characters. The multiplication and division operations in the field are the same as those in the real number field, while the addition and subtraction XOR operations are the same.

Since the update amount of frequently updated data is usually small, in order to verify the impact of the update amount (data block size) on the update efficiency of frequently updated data, we give the efficiency of different data block sizes when inserting data blocks into the file, as shown in Figure 6. It can be concluded that the time of data update operation (insert data block) increases with the increase of file size.

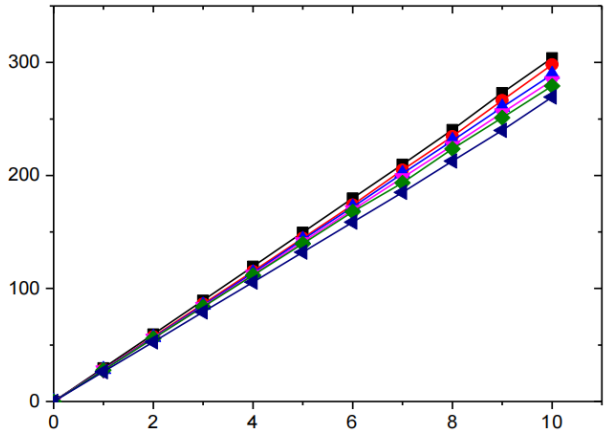

Figure 6 Change of update time (insert data block) with file size

V. Conclusion

When the data block size is small (varying from 1KB to 10KB), there is little difference in the time cost of data update operation. According to the above situation, based on the following reasons: first, because the archived data is not updated, its data block partition can be set directly according to the file size and other parameters. Secondly, in order to save the cost of updating data frequently, the data block partition is related to the actual update situation. Therefore, in the experimental environment given in this paper, for file f, the following conclusions are drawn:

(1) For archived data, data block partition is related to file size. According to the theoretical analysis, we can set the data block size to 1MB;

(2) For frequent updating data, because fine-grained updating is one of its characteristics, we want the data block size to be as small as possible. According to the above experiments, when the data block size changes from 1KB to 10KB, the update efficiency is not affected. So we can set the block size to 8KB.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. U1636107, 61972297), the Science and Technology Project of Henan Province (China) (Grant No. 182102210215, 192102210288), the Soft Science Project of Henan Province (China) (Grant No. 182400410482).

References

[1] Jiang Wei, Fang Binxing, Tian Zhihong. Network Security Evaluation and Optimal Active Defense Based on Attack Defense Game Model. Acta Computer Sinica, 2009, 32 (004): 817-827
[2] Miao Yongqing. Stochastic Model Method and Evaluation Technology of Network Security. China Science and Technology Investment, 2017, 4: 314
[3] Yi Hua Zhou, Wei Min Shi, Wei Ma. Research on Computer Network Security Teaching Mode for Postgraduates Under the Background of New Engineering. Innovation and Practice of Teaching Methods, 2020, 3 (14): 169
[4] Bao Xiuguo, Hu Mingzeng, Zhang Hongli. Two Quantitative Analysis Methods for Survivability of Network Security Management Systems. Acta Communication Sinica, 2004, 25 (9): 34-41
[5] Yang Yi, Bian Yuan, Zhang Tianqiao. Network Security Situation Awareness Based on Machine

Learning. Computer Science and Application, 2020, 10 (12): 8

[6] Li Zhiyong. Hierarchical Network Security Threat Situation Quantitative Assessment Method. Communication World, 2016, 23: 70-70

[7] Hu Wenji, Xu Mingwei. Analysis of Secure Routing Protocols for Wireless Sensor Networks. Journal of Beijing University of Posts and Telecommunications, 2006, 29 (s1): 107-111

[8] Wei Yonglian, Yi Feng, Feng Dengguo, Yong W, Yifeng L. Network Security Situation Assessment Model Based on Information Fusion. Computer Research and Development, 2009, 46 (3): 353-362

[9] Xu Guoguang, Li Tao, Wang Yifeng. A Network Security Real-time Risk Detection Method Based on Artificial Immune. Computer Engineering, 2005,31 (12): 945-949

[10] Li Weiming, Lei Jie, Dong Jing. an Optimized Real-time Network Security Risk Quantification Method. Acta Computa Sinica, 2009 (04): 793-804

ISSN: 0010-8189