International Conference on Advancements in Engineering, Technology and Sciences
Dhaanish Ahmed College of Engineering, Chennai (ICAETS-2018)
(Approved by AICTE, New Delhi and Affiliated to Anna University, Chennai)
16th-17th March 2018 www.conferenceworld.in ISBN : 978-93-87793-11-8

A SECURE MULTI-KEYWORD RANKED SEARCH SCHEME OVER AUDIT-FREE CLOUD STORAGE

Preethi.R[1], Dr.B.L.Velammal[2]

[1]*Department of Computer Science and Engineering,*
College of Engineering Guindy, Anna university(India)

[2]*Department of Computer Science and Engineering,*
College of Engineering Guindy, Anna university(India)

ABSTRACT

*The major aim in this paper is to solve the problem of multi-keyword ranked search over encrypted cloud data (MRSE) to protect the privacy in the cloud platform. Data holders usually outsource their private data to the cloud for large lexibility, easy access from anywhere and huge number of data storage in cloud. Inorder to preserve data privacy, the sensitive data should be encrypted before outsourcing it in the cloud, which performs traditional method based on plaintext keyword search. Hence an encrypted data search scheme is the most significant here. Incase of the large number of users and documents in the cloud, it can be searched with several keywords and the related documents are retrieved based on the keywords. This same searchable encryption mechanism is used in single keyword search or Boolean keyword search, and rarely sort the search results. For multi-keyword search scheme, has drawback that it results with as many matches as possible from the documents based on user query. It is necessary to measure the accurate matches ofthe documents and user query.Inorder to overcome this issues, index tree is constructed. Based on the Term Frequency (TF) and Inverse Document Frequency (IDF), the index tree is constructed, so the matched one could be exactly measured accurately by the inner product of the TF * IDF. On the other hand, directly outsourcing the TF and IDF in the cloud will break the index privacy or the search privacy. To prevent this issues, DES encryption technique is used to encrypt the TF and IDF value and then the encrypted values will be outsourced in the cloud. As a result, the data leakage can be eradicated and data security is guaranteed. And also a new cloud storage encryption scheme is established, that enables cloud storageproviders to createconvincing fake user secrets to protect user privacy. Since coercers cannot tell if obtained secrets are true or not, the cloud storage providers ensure that user privacy is still securely protected.*

Keywords :* Index tree construction, Searchable encryption, multi-keyword ranked search, cloud computing, attribute based encryption.

I.INTRODUCTION

Cloud computing is a conversational phrase used to express a variety of dissimilar types of computing ideas that occupy large number of computers that are connected through a real-time communication network i.e Internet. In science, cloud computing is the capability to run a program on many linked computers at the same time. The fame of the term can be recognized to its use in advertising to sell hosted services in the sense of application service provisioning that run client server software on a remote location. Cloud computing relies on sharing of resources to attain consistency and financial system alike to a utility (like the electricity grid) over a network. The cloud also centres on maximize the effectiveness of the shared resources. Cloud resources are typically not only shared by multiple users but as well as dynamically re-allocated as per demand. This can perform for assigning resources to users in dissimilar time zones. For example, a cloud computing service which serves American usersduring American business timings with a specific application (e.g. email) while the same resources are getting reallocated and serve Indian users during Indian business timings with another application (e.g. web server).

This mechanism must take full advantage of the use of computing powers thus decreasing environmental damage as well, since less power, air conditioning and so on, is necessary for the same functions. The expression "moving to cloud" also explains to an organization moving away from a traditional CAPEX model i.e buy the devoted hardware and decrease in value it over a period of time to the OPEX model i.e use a shared cloud infrastructure and pay as you use it. Proponents maintain that cloud computing Permit Corporation to avoid direct infrastructure costs, and focus on projects that distinguish their businesses as an alternative of infrastructure. Proponents also maintains that cloud computing permit schemes to get their applications should run faster, with better manageability and less maintenance, and enable IT to more quickly adjust resources to meet random and changeable business demand. Now a day's cloud computing has become essential for many utilities, where cloud customers can slightly store their data into the cloud so as to benefit from on-demand high-quality request and services from a shared pool of configurable computing resources. Its huge suppleness and financial savings are attracting both persons and enterprise to outsource their local complex data management system into the cloud. To safe guard data privacy and struggle unwanted accesses in the cloud and away from, sensitive data, for example, emails, personal health records, photo albums, videos, land documents, financial transactions, and so on, may have to be encrypted by data holder before outsourcing to the business public cloud; on the other hand, obsoletes the traditional data use service based on plaintext keyword search. The insignificant solution of downloading all the information and decrypting nearby is clearly impossible, due to the enormous amount of bandwidth cost in cloud scale systems. Furthermore, apart from eradicating the local storage management, storing data into the cloud supplies no purpose except they can be simply searched and operated. Thus, discovering privacy preserving and effective search service over encrypted cloud data is one of the supreme importance. In view of the potentially large number of on-demand data users and vast amount of outsourced data documents in the cloud, this difficulty is mostly demanding as it is really difficult to gather the requirements of performance, system usability, and scalability.

On the one hand, to congregate the efficient data retrieval requirement, the huge amount of documents orders

the cloud server to achieve result relevance ranking, as an alternative of returning undifferentiated results. Such ranked search system allows data users to discover the most appropriate information quickly, rather than burdensomely sorting during every match in the content group. Ranked search can also gracefully remove redundant network traffic by transferring the most relevant data, which is highly attractive in the "pay-as-you-use" cloud concept. For privacy protection, such ranking operation on the other hand, should not reveal any keyword to related information. To get better the search result exactness as well as to improve the user searching experience, it is also essential for such ranking system to support multiple keywords search, as single keyword search often give up far too common results. As a regular practice specifies by today's web search engines i,e Google search, data users may lean to offer a set of keywords as an alternative of only one as the indicator of their search interest to retrieve the most relevant data. And each keyword in the search demand is able to help narrow down the search result further. "Coordinate matching", as many matches as possible, is an efficient resemblance measure among such multi-keyword semantics to refine the result significance, and has been widely used in the plaintext information retrieval (IR) community. Though, the nature of applying encrypted cloud data search system remains a very demanding task in providing security and maintaining privacy, like the data privacy, the index privacy, the keyword privacy, and many others. Encryption is a helpful method that treats encrypted data as documents and allows a user to securely search through a single keyword and get back documents of interest. On the other hand, direct application of these approaches to the secure large scale cloud data utilization system would not be necessarily suitable, as they are developed as crypto primitives and cannot put up such high service-level needs like system usability, user searching experience, and easy information discovery. Even though some modern plans have been proposed to carry Boolean keyword search as an effort to improve the search flexibility, they are still not sufficient to provide users with satisfactory result ranking functionality. The solution for this problem is to secure ranked search over encrypted data but only for queries consisting of a single keyword. The challenging issue here is how to propose an efficient encrypted data search method that supports multi-keyword semantics without privacy violation. In this paper, we describe and solve the problem of multi-keyword ranked search over encrypted cloud data (MRSE) while preserving exact system wise privacy in the cloud computing concept. Along with various multi-keyword semantics, select the efficient resemblance measure of "coordinate matching," it means that as various matches as possible, to confine the significance of data documents to the search query. A new cloud storage encryption scheme that enables cloud storage providers to create convincing fake user secrets to protect user privacy. Since coercers cannot tell if obtained secrets are true or not, the cloud storage providers ensure that user privacy is still securely protected.

II. PROPOSED FRAMEWORK

In the proposed system, a secure tree-based search scheme is proposed over the encrypted cloud data, which supports multi-keyword ranked search and dynamic operation on the document collection. Specifically, the widely-used "term frequency and inverse document frequency model are combined in the index construction and query generation to provide multi-keyword ranked search and GDFS scheme is used to obtain high search

efficiency. To Overcome different attacks in different threat models, the index tree is encrypted. The cloud storage providers create convincing fake user secret for unauthorized user to ensure that user privacy is securely protected.

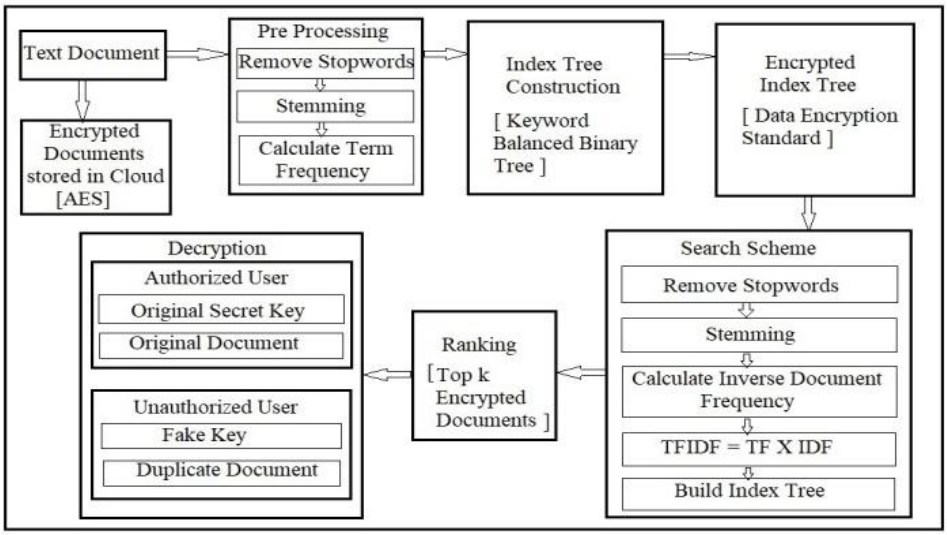

Fig 1. Block diagram of the proposed system

In Figure 1, is the overall system architecture of the Multi-Keyword Ranked Search Scheme over Audit free cloud storage. Here the text documents is given as input and it is encrypted to produce cipher text. All the documents are encrypted and stored in Cloud Server. To Perform high efficient search operation, an secure tree based search scheme is established. Extract the text document, read the content of the document and perform stopword removal and stemming to produce keywords. Calculate the term frequency for each keyword and then construct the index tree based on tf value and perform encryption. While performing search operation, the keyword and k value is given by the user, from which again extract the keyword, remove stopwords , perform stemming operation and then calculate Inverse Document Frequency for each document. In accordance to the search result, top k encrypted documents are retrieved from cloud. If the user is authorized person , then with the help of the key can decrypt the document else if the unauthorized user tried with any fake key, then inorder to fool the user , an fake convincing user secret is provided by the cloud providers.

From the above architecture diagram, it is divided into 6 different modules inorder to perform the search scheme.The first step is Pre-Processing which perform 3 operations such as remove stopwords, stemming, and calculating Term Frequency. The Text Document is given as input to the pre-processing module, where the content of the text document is extracted and remove stopwords and punctuation which produce keyword. After producing keywords, stemming operation is performed in all the keywords and produce a keywords such as it should be a perfect noun. Then based on the keyword calculate the number of occurance of the keyword in all the document and by using the formula calculate term frequency(TF) value. The next module is Index Tree Construction, here the calculated term frequency value is used to construct index tree with the help of keyword

balanced binary tree algorithm. Then the Index tree is encrypted based on Data Encryption Standard(DES) algorithm. The Term Frequency values in the tree is encrypted inorder to preserve keyword privacy. Then in Search Scheme, it performs 5 operations, when the user search for the keyword in the cloud, it removes stopwords from the user query, then perform stemming and then calculate Inverse Document Frequency(IDF) . Then based on the Term Frequency and Inverse Document Frequency values, multiply both the values to produce Term Frequency Inverse Document Frequency(TFIDF) values. The TFIDF values is used to construct index tree again to perform Greedy Depth First Search and produce top K values related to Documents. From the top K Docments, User can select the required document , to view the original document, the user have to send request to admin for key. If it is authorized user, then only the admin sends key to user mail-id or else no key will be revealed by the admin. If in case, the unauthorized user try to access the document or try with duplicate key, then the cloud provider will provide them with fake convincing document inorder to fool the unauthorized user.

III. PERFORMANCE ANALYSIS

1) Precision:

Precision is known as positive predictive value. It is defined as the number of correct result divided by the number of the retrieved result. Precision returns the exactness (accurate) of the result. If the threshold value changes the precision may increase or decrease accordingly.

$$\text{Precision} = \sum TP / (\sum TP + \sum FP)$$

where,

TP = True positive, total number positive reviews which is positive

FP = False positive, total number negative reviews which is positive

2) *Recall:*

Recall is known as true positive value or sensitivity. It is defined as the number of correct result divided by the number of the relevant result. Recall returns the completeness of the result but doesn't depend on any threshold value. . The recall value is calculated after finding the new rating, to check whether the user had given the positive or negative ratings correctly.

$$\text{Recall} = \sum TP / \sum P$$

where,

TP = True positive, total number of true positive conditions ranked.

P = Total number of correctly ranked.

3) *Accuracy:*

The Accuracy or recognition rate measures the proportion of true results to the total number of the instances in

the dataset. It specifies how many instances correctly predicted. This affects the system performance. Misclassification measures the error rate (ie) the proportion of false results to the total number of the instances for the given dataset. It specifies how many instances falsely predicted. This affects the system performance.

$$\text{Accuracy} = (\sum TP + \sum TN)/(\sum P + \sum N)$$

Where,

TP=True positive, total number of true positive reviews is classified as positive.

TN= True negative, total number of true negative reviews is classified as negative.

P = Total number of correctly ranked.

Table 1: Calculated Precision, Recall and Accuracy values

No.of Documents	**50**	**100**	**150**	**200**	**250**
Recall	0.9	0.98	0.95	0.8	0.85
Precision	0.8	0.75	0.9	0.6	0.8
Accuracy	0.97	0.96	0.95	0.89	0.88

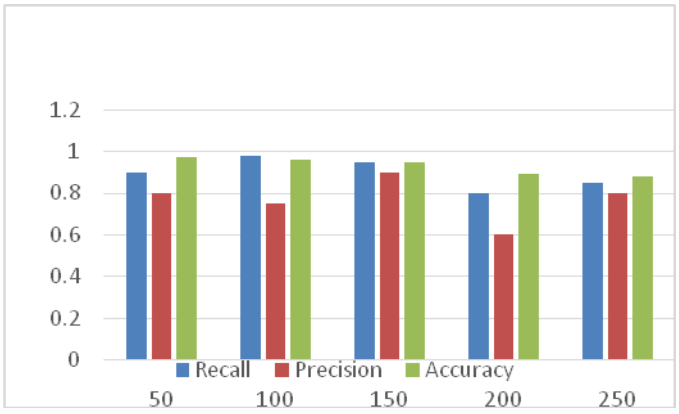

Fig. 2 Comparison of Precision , Recall and Accuracy

IV. CONCLUSION AND FUTURE WORK

In this paper, a secure, efficient and dynamic search scheme is proposed, which supports not only the accurate multi-keyword ranked search but also the dynamic deletion and insertion of documents. We construct a special keyword balanced binary tree as the index, and propose a "Greedy Depth-first Search" algorithm to obtain better efficiency than linear search. In addition, the parallel search process can be carried out to further reduce the time

cost. The security of the scheme is protected against two threat models by using the secure KNN algorithm. Experimental results demonstrate the efficiency of our proposed scheme. There are still many challenge problems in symmetric SE schemes. In the proposed scheme, the data owner is responsible for generating updating information and sending them to the cloud server. Thus, the data owner needs to store the unencrypted index tree and the information that are necessary to recalculate the IDF values. Such an active data owner may not be very suitable for the cloud computing model. It could be a meaningful but difficult future work to design a dynamic searchable encryption scheme whose updating operation can be completed by cloud server only, meanwhile reserving the ability to support multi-keyword ranked search. In addition, as the most of works about searchable encryption, our scheme mainly considers the challenge from the cloud server. Actually, there are many secure challenges in a multi-user scheme. Firstly, all the users usually keep the same secure key for trapdoor generation in a symmetric encryption (SE) scheme. In this case, the revocation of the user is big challenge. If it is needed to revoke a user in this scheme, we need to rebuild the index and distribute the new secure keys to all the authorized users. Secondly, symmetric SE schemes usually assume that all the data users are trustworthy. It is not practical and a dishonest data user will lead to many secure problems. For example, a dishonest data user may search the documents and distribute the decrypted documents to the unauthorized ones. Even more, a dishonest data user may distribute his/her secure keys to the unauthorized ones. In the future works, we will try to improve the SE scheme to handle these challenge problems.

REFERENCES

[1] C. Wang, N. Cao, K. Ren, and W. J. Lou (2012), "Enabling secure and efficient ranked keyword search over outsourced cloud data," IEEE Trans. Parallel Distrib. Syst., Vol. 23, no. 8, pp. 1467–1479.

[2] C. Chen, X. J. Zhu, P. S. Shen, and J. K. Hu (2014), "A hierarchical clustering method For big data oriented ciphertext search," in Proc. IEEE INFOCOM, Workshop on Security and Privacy in Big Data, Toronto, Canada, pp. 559–564.

[3] P. K. Tysowski and M. A. Hasan (2013), "Hybrid attribute and re-encryption based key management for secure and scalable mobile applications in clouds." IEEE Transaction Cloud Computing, pp. 172–186.

[4] W. Zhang, S. Xiao, Y. Lin, T. Zhou, and S. Zhou (2014), "Secure ranked multi-keyword search for multiple data owners in cloud computing," in Dependable Syst. Networks (DSN), IEEE Transaction, pp. 276–286.

[5] Zhihua Xia, Xinhui Wang, Xingming Sun, and Qian Wang (2015), "A Secure and Dynamic Multi-keyword Ranked Search Scheme over Encrypted Cloud Data," IEEE Transactions on Parallel and Distributed Systems, pp.1-8.

[6] X. D. Song, D. Wagner, and A. Perrig (2010), "Practical techniques for searches on encrypted data," IEEE Symposium Security Private, pp. 44–55.