An Efficient Dictionary and Lingual Keyword based Secure Search Scheme in Cloud Storage

Sushil Kumar Verma
Indian Institute of Information Technology
Allahabad (India)

Sijo Mathew
Indian Institute of Information Technology
Allahabad (India)

Shashank Srivastava
Indian Institute of Information Technology
Allahabad (India)

S. Venkatesan
Indian Institute of Information Technology
Allahabad (India)

ABSTRACT

Cloud computing has become the more accepted answer to the current information needs. With today's enterprises operating across multilingual and multicultural spectrum, cloud computing search algorithms should be upgraded to perform searches based on multilingual languages. Data stored in cloud is in encrypted form this is in order to maintain the data integrity, the data search techniques used in the cloud needs to maintain this encryption as the cloud service provider should not be able to ascertain what is entering and leaving the cloud. This paper proposes a solution ,which will provides a method for finding exact encrypted data by using multilingual search queries, phonetic based and also keeping the data integrity of the stored data in the cloud intact. The paper also provides a comparative study between the various public key encryptions and the proposed scheme.

Keywords

Dictionary Search, Lingual Search, Phonetics Search, Semantic Search, Untrusted Cloud Storage

1. INTRODUCTION

Cloud computing has become the more accepted answer to the current information needs .With the advent of terminology like Big Data being commonly applied to data generating systems, and data availability being synchronous to the mobility of an individual cloud computing has carved itself as a omnipotent solution. Cloud computing provides seemingly unlimited storage solution to its user. It provides access to the stored data on demand and act as a single point of contact for the global operations of the user for data transaction. Data storage on the cloud has been skeptical due to the concerns related to data integrity and data proliferation. As the cloud services are being handled by third party handler and cloud service provider playing as a host to a wide spectrum of businesses gives rise to question of data security. Encrypted data storage provides redressal to these difficulties but poses new challenges. Encrypted storage of data in the cloud facility has provided a solution to the data integrity problem but has paved a way for problems relating to encrypted data search, and the limitations embodied with the traditional data search technique. A major limitation associated with encrypted data search is the inability of these algorithms to filter the match across multilingual languages. The latter limitation cited in evident in trans-lingual work environment being practiced in global organization. Challenges related how to search exact data in the stored encrypted data and how to filter search for encrypted data in multilingual work culture. The later part of the challenge refers to the problem in which different words to address a single object. The possible way for conducting a search in the encrypted data is to download the whole data on end user system and conduct the search, which proves to be an unviable option as this will consume a lot of computational and storage resource at the client end. Encrypted word search is the viable approach for this problem. The multilingual part of the problem can be addressed by the introduction of multilingual dictionary in the encrypted cloud environment which will help in the optimization of the whole system to cater a multilingual end user profile. Multilingual phonetic in this paper we have proposed a novel procedure for optimizing the encrypted search in cloud storage that will keep the integrity of stored data intact from the cloud service provider. The paper also presents a comparison between already existing two approaches and our proposed new lingual search algorithm.

2. PROBLEM STATEMENT

In a multicultural operating environment a word can be identified by various other words which may have the same or nearly same or synonymous meanings. For example the name "Indian Institute of Information Technology" can have various combinations to its attributes like India can be referred to as Bhartiya or India or Indian National having multiplicity according to the variousness in accordance to the end-user set. Similarly Institute can be referred to as University or College or "Sansthan" and Information Technology can be referred as "Suchna Prodhyogiki" the hindi translation of the afore said string. The search scheme should be intelligent enough to direct these various search combination to the appropriate content in the cloud environment.

Further elaborating, at an International platform multiple languages comes into scenario. Considering the afore said example University is English, Universitiet is the Dutch translation of university and Ollscoil is the Irish translation for the same. The search scheme robustness to search data in the encrypted form is needed to cut across the multilingual strata of the end users.

When, considering the various aspects of data search different phonetic alphabets also arise in the scenario. These phonetics and alphabets arise due to the language linguistics. This problem is discussed in more dilate form in the following example : "Umea University Sweden" is how it is typed in normal English whereas the original name is "Umeå University" is the proper noun of the university, similarly "University of Jane" is how the name is communicated in normal English whereas the original name is "University of Jaén". Here å and é are the phonetics associated with their native names but are largely represented as "a" and "e" only. This variables are also needed to be addressed to for and effective and efficient search.

3. RELATED WORK

Searching of encrypted data using key word efficiently was first introduced in Song et al. (2000).Thereafter there has been considerable work in this direction by many individual. Key milestones were achieved by Boneh et al. (2004) by proposing a public key encryption with keyword searching (PEKS) scheme, which resulted in tracking keywords in emails without reading the whole content. This also paved way for reducing the computational overhead required for decrypting and encrypted data. Our paper draws its inspiration from the work of Qin Liu et al. (2009) who introduced an efficient privacy preserving keyword search scheme, the main outcome of this work was it allowed a service provider to participate in partial decryption to reduce a client's computational overhead, but this scheme was in effective for a cloud environment. The work was further refined by the work of et al. R. ChinnaSamy (2012) where an efficient semantic secure keyword based search (ESSKS) scheme suitable for cloud storage is proposed.

Our paper proposes a approach of introducing dictionary and lingual based secure search scheme which enables end user to get a result which will filter keywords with the flexibly to accommodating lingual, phonetic and dictionary combinations of words, with keeping the integrity of the encrypted stored data in the cloud intact.

4. DICTIONARY AND LINGUAL BASED SECURE SEARCH SCHEME

User (A): In cloud system, encryption, decryption and updating of the file operation and secure file object can be performed by the user (author). Key management is mandatory for this user.

User (R): In cloud system searching and decryption can be performed by the user but there is no right for the user(reader) to access to manipulate the secure file object.

Untrusted Cloud Server: Server will contain the data into two parts first one is encrypted file and second one secure file object.

Distributed key server: It is distributed in nature. It stores the file number of the files, key and hash values of random words from the file. The key server system is also connected to the internet. It is placed inside the relevant organizations.

According to Fig.1 the DLBSS scheme works as follows:

Step 1: The user (A) sends encrypted format of filename, encrypted keywords, file ID number , user access list containing hash value of private key ,time stamp a hash value of random words from the file and encrypted modified and linguistic dictionary with word ranking methodology to the cloud storage system.

Step 2: After sending the information to Cloud Storage Server, the user sends the file ID number, key, same timestamp and hash value of the random word to the key server which is locally established in particular organization and connected to the internet.

Step 3: User(R) will search the specific file by sending encrypted keywords and hash value his/her private key.

Step 4: Then UCS sends the ciphertext of the file, ciphertext of the file ID number, hash values of random words from the file and timestamp when the keyword matches to user querying keywords. In case the user keyword sends by the user(R) matches with more than one file from the database, it will show the all the files matched with the keywords according to rank basis.

Step 5: The Distributed Key Server returns the particular key of the file and hash values of random words from the file and timestamp to the user(R).

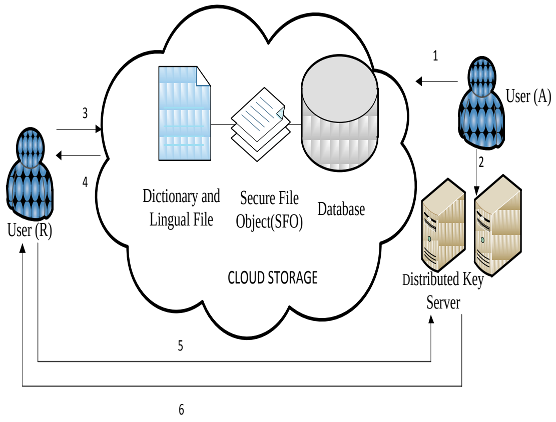

Fig 1 : Process of DLBSS Scheme

The detailed process undertaken in the DLBSS Scheme is described as below

Uploading of encrypted files to UCS

$M(User(A) \rightarrow UCS(db)) = [C_F, C_{Fid}, C_{FN}, H_{RK}, t\ ]$

C➔ Ciphertext of the file

C_{Fid}➔File identification number

C_{NF}➔Ciphertext of the file name

H_{RK}➔Hash function of random keywords

T➔ Time stamp

1. File's Text Encryption Methodology FENC:The user U (A) encrypts a file F with symmetric encryption algorithm E, which generates a ciphertext C = EK(F), here the encryption key K is random and it will be unique for specific file. The random key will be transmitted to the other user securely so that the authorized user can decrypt the file.

2. File's Component Encryption Methodology FC_{ENC} : The user U (A) encrypts all file components (File ID Number ,File Name, Hash function of random keywords, Time Stamp , Keywords for the file , User Access List with symmetric encryption algorithm E, which generates a ciphertext C = EK(FCi), here the encryption key K is random but it will not unique it will be same for all the component the file as well as it will be known to all the least authorized user.

3. Random Function Hash Methodology RwH : In this methodology the random function picks up defined random words from the file content having word length greater than a threshold value, then apply the hash function to the those keywords only

H_{RK}=H(K_{W1}, K_{W2}, K_{W3}, K_{W4}, K_{W5}, ...).

Creating of Secure File Object (SFO) as on fig2

M(User(A)➔ USC(SFO))=[$C_{Fid,}$ $C_{NF,}$ $K_{wi,}L_{UA,}$,t]

C_{Fid}➔File identification number

C_{NF}➔Ciphertext of the file name

K_{Wi}➔ Encrypted Keywords for the particular file

L_{UA}➔ User Assess list for authorized user

T➔ Time stamp

Components of Secure File Object (SFO)
File ID Number
Ciphertext of the file name
Encrypted Keywords for the particular file
User Assess list for authorized user
Time stamp

Fig 2: Components of Secure File Object (SFO)

First one will be phonetics parts second part will be consist of lingual words and the last but not the least third part is semantic based. All the parts will be ranked accordingly.
Keyword :[{ Same phonetics words},{Similar Lingual Words},{Same Semantic Words}]
K_{Wi}:[{$K^{*}_{w23,}$ $K^{*}_{w85,}$ K^{*}_{w45}},{ $K^{*}_{w22,}$ $K^{*}_{w27,}$ $K^{*}_{w2374,}$ K^{*}_{w12}}, {$K^{*}_{w23,}$ $K^{*}_{w289,}$ K^{*}_{w99}}]
Here the ranking of keyword will be:
K^{*}_{w23} =0
K^{*}_{w85} =1
K^{*}_{w45} =2
K^{*}_{w22} =3
K^{*}_{w27} =4
K^{*}_{w74} =5
K^{*}_{w12}=6
K^{*}_{w23} =7
K^{*}_{w89} =8
K^{*}_{w99} =9

Searching of the Data

First of all user(R) sends the encrypted keywords and hash function of his/her private key and the whole message will go to the Dictionary & Lingual file after that all the keywords will be matched and modified keyword will emerged for this process.

Original keyword by user:
M(User(R))➔ USC=[K_{wi}{ K_{w1} , K_{w2} , K_{w3} , K_{w4} etc},{H_{PK}}]

Modified keywords send form USC to SFO:

USC➔SFO=[K_{w1}($K^{*}_{w23,}$$K^{*}_{w85}$),$K_{w2}$($K^{*}_{w99,}$$K^{*}_{w74,}$$K^{*}_{w12}$), K_{w3}($K^{*}_{w45,}$$K^{*}_{w27,}$ K^{*}_{w89}) , K_{w4} , K_{w5} etc},{H_{PK}}]

After matching with these keywords with SFO file's keywords cloud system will generate the list file ids with ranking order.

Distributed Key Server

The User (R) processes the following
M(UCS➔ User(R) =[C_{NF}, C_{Fid}, C, H_{RK},t]
M(User(R)➔ DKS= [C_{NF}, C_{Fid}, H_{RK},t]

From the above process, after receiving a C_{NF}, C_{Fid}, C,H_{RK},t, user (R) sends C_{NF}, C_{Fid}, H_{RK},t to the distributed key server. The DKS compares C_{NF}, C_{Fid}, H_{RK},t to the information that is stored and if they are equal and matches , returns the key K and H_{RK} to user (R). Then the user (R) uses K key to decrypt C to get the decrypted version of file F and check the hash value of the random words from the file content H_{RK} for data integrity.

From the whole process, it can be seen that the cloud storage server does not take part in neither the encryption nor decryption and nor having any type of key with it, in this scheme everything is encrypted with symmetric key encryption in such a way that cloud server never know who is the users of that organization and to maintain the data integrity and non-repudiation with an add-on security like secure file object (SFO).

5. PERFORMANCE ANALYSIS

Key management, data redundancy, data integrity, security, time and dictionary and lingual based output are some major factors that can be analyzed for measuring performance.

None of the available schemes for secure search in cloud environment presents the dictionary based approach for reducing time and efficient search.

Where PEKS and EPPKS uses single key pair for encryption and decryption processes. Both search schemes stores the key pair at cloud end which is a soft target for attacker.

DLBSS involves both encryption and decryption to be done at user end, thus reduces the computational overhead at cloud.

A single key pair is used with an extra key available for every file objects separately it addresses the data integrity and security of the file object.

The scheme presented by the paper reduces the computational overhead generated by presently available search schemes to a much extent. Furthermore, the problems arising due to key management and data integrity can be resolved in a method more adaptable to cloud services environment.

Table 1. provides analysis of all available three schemes with the scheme that this paper proposes.

	PEKS	EPPKS	ESSKS	DLBSS
Memory Power & CPU capability	High[4]	Less[8]	Less[1]	Less
Security	Less	Less	Less	High
Key Management	Hard[1]	Hard[1]	Easy[1]	Easy
Data Integrity	No[1]	No[1]	Yes[1]	Yes
Dictionary and Lingual Based Result	No	No	No	Yes

6. SECURITY MECHANISM

The search scheme suggested in this paper deals with security by using private key of the user. The private key of the user is hashed using known hash algorithm. The result of this hash is then encrypted using symmetric key algorithm to obtain a unique key value. This unique key value serves as an authenticator. The secure file object contains this authenticator value for authentication. The hash value send by the user is matched with the set of hash values already maintained in the cloud environment by organization. When both values matched, the access is granted to the user.

7. CONCLUSION AND FUTURE WORK

In this paper, we have presented a new dictionary and lingual based secure search scheme for cloud storage environment which reduces the time consumed in searching the data from the cloud storage system. It provides a platform which enables of user of multicultural backgrounds to filter there search irrespective of their lingual and phonetic disparities. This scheme maintains the integrity of the data during data transfer from cloud to the user and vice versa.

Our future work mainly focuses on distribution of dynamic key to provide enhanced security and computational requirements can be reduced at user ends.

8. REFERENCES

[1] R. ChinnaSamy, Dr. S. Sujatha. An Efficient Semantic Secure Keyword Based Search Scheme in Cloud Storage Services. In International Conference on Recent Trends in Information Technology ICRTIT-2012

[2] Ning Cao, Cong Wang, MingLi, Kui Ren, and Wenjing Lou. Privacy-Preserving Multi-keyword Ranked Search over Encrypted Cloud Data. In IEEE INFOCOM 2011

[3] Chang Liu, Liehuang Zhu, Longyijia Li, Yu'an Tan. Fuzzy Keyword Search on Encrypted Cloud Storage Data With Small Index. In Proceedings of IEEE CCIS2011

[4] Boneh D, Crescenzo G, Ostrovsky R, Persiano G. Public key encryption with keyword search. In: Proceedings of Eurocrypt 2004, Lecture notes in computer science, vol. 3027; 2004. p. 506–22.

[5] Thulasimani Lakshmanan and Madheswaran Muthusamy. A Novel Secure Hash Algorithm for Public Key Digital Signature Schemes .In The International Arab Journal of Information Technology, Vol. 9, No. 3, May 2012

[6] Anup Mathew. Survey Paper on Security & Privacy Issues in Cloud Storage Systems. In EECE 571B, TERM SURVEY PAPER, APRIL 2012

[7] Boneh D, Waters B. Conjunctive, subset, and range queries on encrypted data. In: Proceedings of TCC 2007, Lecture notes in computer science, vol. 4392; 2007. p. 535–54.

[8] Liu Q, Wang G, Wu J. An efficient privacy preserving keyword search scheme in cloud computing. In: Proceedings of IEEE TrustCom-09 in conjunction with IEEE CSE-09; 2009. p. 715–20.

[9] Ari Pirkola, Turid Hedlund, Heikki Keskutalo,and Kalerro Jarvelin, Dictionary-Based Cross Language Information Retrieval: Problems, Methods and Research Findings. In Information Retrieval 4(3/4),209-230

[10] Y.-C. Chang and M. Mitzenmacher. Privacy Preserving Keyword Searches on Remote Encrypted Data. Proceedings of ACSN 2005, Lecture Notes in Computer Science 3531, pp. 442-455.

[11] Liu Hong-xia, Dai Jia-zhu, Jiang Chao "Research on Privacy Preserving Keyword Search in Cloud Storage "Computer Science and Information Technology ICCSIT 2010 3rd IEEE International Conference

[12] D. X. Song, D. Wagner and A. Perrig. Practical Techniques for Searches on Encrypted Data. Proceedings of the 2000 IEEE Symposium on Security and Privacy, pp. 44-55