iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** V **Month of publication:** May 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Multi-Keyword Ranked Search over Encrypted Cloud Data with Privacy-Preservation

Abhila. G. K[1], Chithra. A. S[2]
[1]MTech Student, [2]Associate Professor, Dept of CSE
Lourdes MathaCollege of Science and Technology, Trivandrum, India

Abstract*: The use of cloud computing became popular due to it's flexibility and economic savings. The cloud data can access anywhere from the world with the help of internet. The cloud servers consists of lots of records which is outsourced by different owners. Here security is main factor. So records must be encrypted before outsourcing them. Different encryption algorithms can be used. The cloud consists of many records, so proper searching techniques wants to be used for searching a record based on user request. The searching of record is mainly based on keywords present in the search query. The single keyword search or Boolean keyword search are two commonly used searching techniques. This paper proposed a novel searching technique for encrypted record with multi-keyword ranked search over encrypted cloud data (MRSE) with privacy preservation. This technique use ranking schemes to list most prior record first according to how many times the searching keywords present in the record. This technique use two methods first one is "coordinate matching" which means as many matches as possible and second one is "inner product similarity" to quantitatively evaluate such similarity measure in records. The experimental results shows the proposed method has low overhead on computation and communication.
Keywords: cloud computing, privacy-preservation, ranked search, multi-keyword, encryption, semantic search

I. INTRODUCTION

The cloud computing is also called on demand computing because users can access cloud based on their needs whenever they wanted. With the help of internet the cloud can access anywhere from the world by a computer or laptop or mobile phones etc. Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS) are different cloud computing services. Sometimes the services offered by the cloud is at free of cost sometimes need to pay some money for the cloud services. The security is main issue concern to cloud. Cloud services offered by cloud servers placed in different places of world. Sometimes a third party may involved in dealing with cloud. So the records which are outsourced must be in secure form. For security reasons the files must be encrypted before outsourcing them to cloud. The cloud is a shared pool of configurable computing resources. Individuals and enterprises outsource their records to cloud mainly due to its great flexibility and economics savings[1].The sensitive datas such as tax documents, financial transactions, personal health records, etc must be encrypted before outsourcing them into cloud servers. Whenever the owner need a file it is practically impossible to download all files and decrypt them locally it will cause huge amount of bandwidth cost. Here search services provided and privacy preservation during searching is also somewhat important. The cloud servers consists of huge number of datas so while performing searching it is extremely difficult to maintain the requirements of performance, system usability and scalability.

To meet the effective data retrieval according to the data users need the relevance ranking scheme must be used instead of retrieving undifferentiated results. Such ranking based search systems enabled data users to retrieve most relevant data's quickly as possible [2].Considering the case of "pay-as-use" cloud paradigms, the ranked search can eliminate unnecessary network traffic. While performing the searching this method should not leak any keyword related information for privacy preservation. It is necessary for ranking system to support multiple keyword search to improve search result accuracy and also to enhance the user searching experience. The single keyword search may lead to far too coarse results. Today's web search engines such as google, yahoo etc provided multi keyword searching facilities. This will help to retrieve the most relevant data's. The search result can narrow down with each keyword present in the search query. The result relevance can be find out by "co-ordinate matching" scheme which means many matches as possible. It is an efficient similarity measure method which help to retrieve the result quickly. Information retrieval (IR) community is highly used by multi-keyword ranked search scheme for refine the result relevance. The single keyword based searching for encrypted cloud data provided less relevance in result. The Boolean searching scheme can be used for enrich the search flexibility compared to single keyword search method [3].A secure *k*-nearest neighbor (*kNN*) technique is adapted in MRSE for secure inner product computation .In this method use index construction with the data vector.

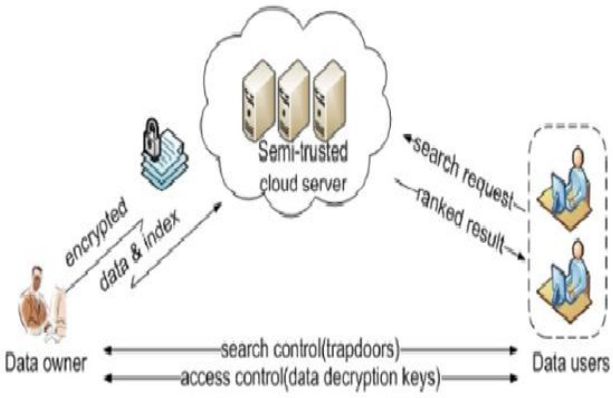
Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

The reminder of this paper consists of Section II, consists of the system model, the threat model, design goals. Section III consists of the MRSE framework and privacy requirements. section IV, which describes the proposed schemes. Section V presents simulation results. Section VI, conclude the paper.

II. PROBLEM FORMULATION

A. System Model

The system architecture of cloud data hosting consists of three different entities, 1)Data owner 2)Data user 3)Cloud server. The architecture is shown in fig(1) [1].



fig(1) system architecture of cloud data hosting and searching

A collection of data documents *F* which are in encrypted form is represented as *C* this is to be outsourced by data owners. To ensure the data privacy the owners of documents must be encrypted before outsourcing it. Here the encrypted index is marked I. In search control mechanism for searching a document collection owner can use t keywords ,from this the owner acquires a corresponding trapdoor *T* for searching purposes. This will help to improve searching speed and accuracy. Best example of this is broadcast encryption[4].The cloud servers get documents trapdoor *T* depending on the searching keywords then the servers find out index *I* of the corresponding encrypted file and retrieve it to the owners. The search results to be ranked by certain ranking criteria such as coordinate matching this is for improving the document retrieval accuracy. The Access control mechanisms is used to manage the decryption capabilities of users [5].

B. Threat Model

In this model the server is considered as " honest-but-curious" one. Because the sever work on designated protocol specification and act in "honest" way. This model is "curious" in case of data including index in its storage, analysing etc. The threat model is classified into two according to what information is known by the server .

1) *Known Ciphertext Model:* In this model the server is known two factors they are encrypted dataset *C* and searchable index *I*.
2) *Known Background Model:* This model is also known as stronger model. The server in this model known more than server in the known ciphertext model. The server known the background informations such as correlation relationship of given search requests through trapdoor and dataset related statistical information etc.

C. Design Goals

The system design goals is mainly to simultaneously achieve security and performance guarantees by the following ways.

1) *Multi-Keyword Ranked Search*: This type of search is used to avoiding returning undifferentiated results during searching.
2) *Privacy-Preserving*: It is used to preserve privacy by preventing server from knowing other information such as index of keyword etc.
3) *Efficiency:* This model should provide goals in functionality and performance with low communication and computational overhead.

D. Notations

1) *F* : the collection of m set plaintext document is denoted by $F = (F1, F2, \ldots, Fm)$.

2) *C:* the collection of m set document encrypted is denoted by $C = (C1, C2, \ldots, Cm)$.
3) *W:* the dictionary which consists of the keyword set consisting of *n* keyword is denoted as $W = (W1, W2, \ldots, Wn)$.
4) *I* : the collection of m set encrypted documents searchable index associated is denoted as $(I1, I2, \ldots, Im)$ where each subindex *Ii* is built for collection of plaintext document *Fi*.
5) f*W:* the subset of keywords in a search request is denoted as $fW = (Wj1, Wj2, \ldots, Wjt)$.
6) *T*f*W:* the search request fW is for corresponding trapdoor.
7) *F*f*W:* all documents with ranked id list with their relevance to f.

III. FRAMEWORK AND PRIVACY REQUIREMENTS FOR MRSE

In this section describe about framework required for MRSE and various privacy requirements for the framework.

A. MRSE Framework

The MRSE framework consists of four algorithms as follows.

1) *Setup(1ℓ):* the data owner take a security parameter ℓ as input and outputs *SK* as symmetric key .
2) *BuildIndex(F, SK):* Based on the dataset *F,* the data owner encrypted the dataset F with the symmetric key *SK* as a result produce a searchable index *I* and then outsourced the encrypted document to the cloud server. The document collection can be independently encrypted and outsourced after the index production.
3) *Trapdoor(fW):* With t keywords of interest in By using f*W* as input with keywords of search query , a corresponding trapdoor *T*f*W* is generated by this algorithm.
4) •*Query(T*f*W, k, I):* A query request as *(T*f*W, k)*is received by the cloud server, with the help of index *I* and trapdoor *T*f*W* it performs the ranked search, and finally returns the with *F*f*W,* top-*k* documents sorted by their similarity with f*W*.

B. Privacy Requirements for MRSE

The privacy requirements for MRSE involved data privacy, index privacy, search privacy. Dealing with data privacy the traditional symmetric key cryptography is used by the data owner for encrypt the data documents. In index privacy to prevent server from leak datas from the documents this type of privacy should be involved so searchable index should be constructed in such a manner. The query procedure involved various search privacy requirements which are complex and difficult to tackle as follows.

1) *Keyword Privacy:* Most of the users require the privacy ie,they want to keep the information regarding what keywords are used in the search query from the server. So the corresponding trapdoor need to hide from the server. The cryptographic way is used to protect generated trapdoor from the server.The server use document frequency scheme for finding the probability of keywords in the query.
2) *Trapdoor Unlinkability:* The nature of trapdoor function should be randomized. The relationship between the given trapdoor is couldn't deduce by the server. The trapdoor unlinkability has the function that [6] it produce sufficient no determinacy into the trapdoor generation procedure.
3) *Access Pattern:* The access pattern consists of sequence of search result. The search result is a set of document with ranked result. The search result for query keyword consists of id list of all documents ranked by their order of relevance. The private information retrieval (PIR) technique is used for a few searchable encryption works [7].

IV. PRIVACY-PRESERVING AND EFFICIENT MRSE

The multi keyword ranked search can be efficiently achieved with the help of "inner product similarity measure" and "coordinate matching". The binary data vector for a document *Fi* is denoted by *Di* where each bit *Di[j]* □ {0, 1} which represent the existence of keyword *Wj* in the document. The inner product of binary column vectors is denoted by $Di \cdot Q$. The cloud server find out ranking by comparing the similarity of documents with keywords in search query document. The inner product of $Di \cdot Q$. where *Di* data vector and *Q* is the query vector is not exposed to the server for preserving system wise privacy.

A. MRSE I: Privacy-Preserving Scheme in Known Ciphertext Model

1) *Secure kNN Computation:* The euclidean distance concept is used in k-nearest neighbor(kNN) scheme.In this method a database record pi and a query vector q is considered and the euclidean distance between them is used for finding k nearest neighbor documents. The combination of one *(d+1)*-bit vector is assigned to S and two $(d + 1) \times (d + 1)$ is assigned to invertible matrices as *{M1,M2}* .The *pi* represent data vector and *q* as query vector are extend to *(d + 1)*-dimension vectors as

$\vec{}pi$ and $\vec{}q$.The dimension of the (d + 1)- is set to $-0.5||pi2||$ and 1. A random number $r > 0$ is scaled by query vector $\vec{}q$ *as* (rq, r).Then $\vec{}pi$ and $\vec{}q$ are split into two random vectors as $\{\vec{}pi', \vec{}pi''\},\{\vec{}q', \vec{}q''\}$.This is encrypted as $\{MT1\ \vec{}pi', MT2\ \vec{}pi''\}$ and $\{M-11\ \vec{}q', M-12\ \vec{}q''\}$.The product of data vector pair and query vector pair is represented as $-0.5r(||pi||2-2pi \cdot q)$.The Euclidean distance between the selected k nearest neighbors is represented as $(||pi||2-2pi \cdot q+||q||2)$.

2) *MRSE I Scheme:* In the case of more advanced design a new random number t is assigned for each query vector's extended dimension.This is a randomness function and this will increase the difficulty to server for understanding the relationship between the received trapdoors.In the search result two things carefully calibrated they are keyword privacy requirement and randomness .To achieve ranked search over multiple keyword search use the follows.

a) *Setup a (n + 2):* bit vector is represented as S and two $(n+2)\times(n+2)$ are represented as invertible matrices $\{M1, M2\}$. A 3-tuple representation is$\{S, M1, M2\}$, formed by secret key SK.

b) *BuildIndex(F, SK) :* For every document Fi data owner generates a binary data vector Di.The corresponding keyword

c) Wj appears in the document Fi is represented by each binary bit Di[j].

d) *Trapdoor(fW):* With fW as input and search query consists of t keywords of interest this will generate one binary vector Q. Each bit Q[j] represented boolean values true or false.

e) *Query(TfW, k, I):* T*f*W represents trapdoor,the cloud server returns the result of top-k ranked files with their id list depending on the query which entered by the owner.

3) *Analysis:* The analysis consists of Functionality and Efficiency analysis and Privacy analysis.

4) *Functionality and Efficiency:* For find out the similarity score introduce a random factor εi.Sometimes sorting the result based on the similarity score is not accurate for the original scheme.The search accuracy can be introduced by εi follow a normal distribution N(μ, σ2), where the flexible trade-off parameter is represented by standard deviation σ.

5) *Privacy:* For data privacy traditional encryption schemes are used where in case of index privacy secret key SK is used confidentially.

B. MRSE II: Privacy-Preserving Scheme in Known Background Model

In case of server known the background information then there is a possibility for leak information regarding which keyword is used for searching. This is possible in case of known background model scheme. In the known background model privacy can be preserved by two methods 1) Scale Analysis Attack 2) MRSE II Scheme.

V. PERFORMANCE ANALYSIS

The performance of this technique is evaluated by tradeoff between privacy , precision and Efficiency.

A. Precision and Privacy

Every search query consists of certain dummy keywords. So the similarity score of the document is not exactly accurate. Based on this keywords server return the top- k document. This documents list may consists of documents out of range for searching due to dummy keywords presence. To evaluate the purity of document retrieved propose a precision factor $Pk = k'/k$ where k' is number of real top-k documents returned by server as a result of search query entered by the owner. The standard deviation σ of the random variable ε is affected by the precision of documents retrieved.

The performance evaluation of this technique is shown in below graph.

B. Efficiency

1) *Index Construction:* In data set F for each document Fi to built a corresponding searchable subindex Ii.The first step here is to map the keyword set from document Fi to data vector Di.The size of directory is find with the help of dimensionality of data vector.The index and subindex are related to time cost required for it.

2) *Trapdoor Generation:* The number of keywords present in the directory is help to find the time required for generating the trapdor.The generation of trapdoor required two things.First one is required two multiplications of a matrix and second one is a split query vector.In two proposed schemes the dimension of the matrix or query vector is different. This will increase if the size of directory is large.

3) *Query:* Query execution in cloud server consists of two steps first one is computing of data set second one is similarity measure for all data documents.The number of keywords in the query document has impact on cost of trapdoor generation.If the number

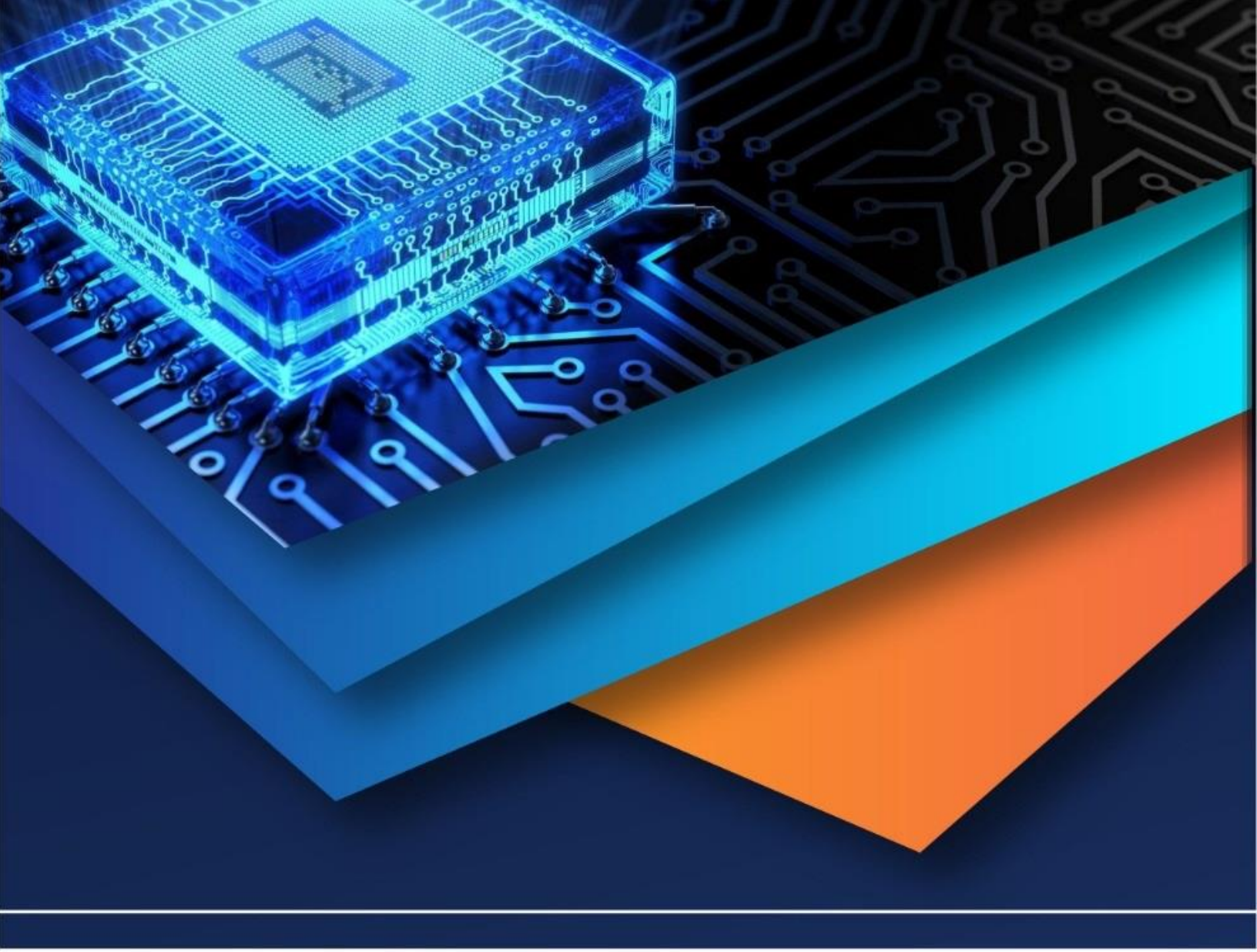
of keywords in the query increases it will increases overhead.

VI. CONCLUSION

This paper define and solve the problem of multi keyword ranked search over encrypted cloud data with different privacy requirement. Effective similarity measure such as "coordinate matching" means many matches as possible and "inner product similarity" measure are used. The weighted query is not work properly by this method this is the limitation of this work. Meeting the challenges of Strong privacy requirements two threat models are used.Consider the case of computation and communication low overhead are produced by this scheme. This is the advantage of this scheme.

REFERENCES

[1] Ning Cao, Cong Wang, Ming Li, Kui Ren, and Wenjing Lou, "Privacy Preserving Multi-Keyword Ranked Search Over Encrypted Cloud Data" ,IEEE transactions on parallel and distributed systems,vol. 25, pp. 222-233, November 2014.
[2] A. Singhal, "Modern information retrieval: A brief overview", IEEE Data Engineering Bulletin, vol. 24, no. 4, pp. 35–43, 2001.
[3] I. H. Witten, A. Moffat, and T. C. Bell, "Managing gigabytes: Compressing and indexing documents and images," Morgan Kaufmann Publishing San Francisco, May 1999.
[4] R. Curtmola, J. A. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: improved definitions and efficient constructions",in Proc. of ACM CCS, 2006.
[5] S. Yu, C. Wang, K. Ren, and W. Lou, "Achieving secure, scalable, and fine-grained data access control in cloud computing," in Proc. of INFOCOM, 2010.
[6] W. K. Wong, D. W. Cheung, B. Kao, and N. Mamoulis, "Secure knn computation on encrypted databases," in Proceedings of the 35th SIGMOD international conference on Management of data, 2009, pp 139–152.
[7] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, "Cryptography from anonymity," in Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, 2006, pp. 239–248.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY