International Journal of Research
Available at
https://edupediapublications.org/journals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 03
March 2017

Privacy Preserving Multi Keyword Ranked Search Over Encrypted Cloud data

K Aparna [1], V. Ganesh Dutt [2]

[1]PG Scholar, Dept of CSE, Sri Sunflower College of Engineering and Technology, Lankapalli, Krishna Dist, AP.

[2] Assistant Professor, Dept of CSE, Sri Sunflower College of Engineering and Technology, Lankapalli, Krishna Dist, AP.

Abstract: *Cloud computing is modern technology as a new computing model in number of business domains. Large numbers of large scale departments are startingto shift the data on to the cloud environment. With the advantage of storage as a service many enterprises are moving their valuable data to the cloud, since it costs less, easily scalable and can be accessed from anywhere any time. Improved dynamic multi-keyword ranking search scheme with top key via encrypted cloud data that simultaneously supports dynamic update operations as deleting and inserting documents. Greedy depth first search algorithm is provided for efficiency multi keywords on place and index structure. Cryptography is one of the establishing trust models. Searchable security is a cryptographic method to provide security. In number of researchers have been working on developing security and efficient searchable encryption methods. We take new effective cryptographic techniques based on data structures like CRSA and B-Tree to enhance the level of security. We propose new multi-keyword search query over encrypted cloud data in retrieving top k scored documents.The vector space model and TFIDF model are used to construct index and query generation. This paper focuses on multi keyword search based on ranking over an encrypted cloud data (MRSE). The search uses the feature of similarity and inner product similarity matching. We propose to support the top-k Multi-full-text search security and performance analysis show that the proposed scheme guarantees a high safety and practicality and dynamic update operations, such as deleting and inserting documents. The experimental results show that the overhead in computation and communication is low.*

Index Terms: Advanced Symmetric Encryption Certified Authority, Cloud data, -Multi keyword Retrieval, Cloud data, Data security, Ranked Search, Similarity Matching.

1. INTRODUCTION

Cloud computing is a term used to describe a set of IT services that are provided to a customer over a network on a leased basis and with the ability to scale up or down their service requirements. Clouds are large pools of easily usable and accessible virtualized resources. These resources can be dynamically reconfigured to adjust to a variable load (scale), allowing optimum resource utilization. It" s a pay-per-use model in which the Infrastructure Provider by means of customized Service Level Agreements (SLAs)[1] offers guarantees typically exploiting a pool of resources. Organizations and individuals can benefit from mass computing and storage centers, provided by large companies with stable and strong cloud architectures.Security and privacy concerns have been the major challenges in cloud computing. The hardware and software security mechanisms like firewalls etc. have been used by cloud provider. These solutions are not sufficient to protect data in cloud from unauthorized users because of low degree of transparency [2]. Since the cloud user and the cloud provider are in the different trusted domain, the outsourced data may be exposed to the vulnerabilities [2] [3]. Thus, before storing the valuable data in cloud, the data needs to be encrypted [5]. Data encryption assures the data confidentiality and integrity. To preserve the data privacy we need to design a searchable algorithm that works on encrypted data [6]. To protect data privacy, confidentiality, and data security, sensitive data like personal health records, e-mails, tax documents, photo albums, financial transactions, and so on, have to be encrypted by data owners before outsourcing to the public cloud [7]. However, the traditional plaintext keyword search data utilization service is obsolete. Downloading all the data and decrypting at the data user side is trivially impractical duo to huge amount of bandwidth cost is required in cloud scale systems. The data can be easily searched and utilized otherwise no purpose of storing data in the cloud. Thus, exploring effective and privacy preserving search over encrypted cloud data is of most prominence. This is a very challenging issue; it degrades performance of system usability and scalability. It is very difficult to meet the requirements of system usability, performance, and scalability, by considering the large number of on-demand data users and large number of outsourced data documents in the cloud. To meet effective data retrieval, the huge amount of documents demands the cloud server to perform relevance ranking as a result, instead returning all result documents. Such ranking system facilitates data users to find the most relevant data rapidly, rather than burdensome sorting through every match in the content collection [8]. Ranked search eliminates the unnecessary network traffic by sending only most relevant documents, which is very much desirable in "pay-as-you-use" cloud model. Such ranked systems should not leak any keyword related information to the cloud server for privacy protection. Such ranking systems should support fuzzy keyword and multi-keyword search, as a single keyword search often gives far too harsh results. A general approach to protect the confidentiality of data is to encrypt the data before outsourcing. However, this will result in an enormous cost in terms of data, ease of use. For example, the existing techniques for keyword-based information retrieval, which are widely used on the plaintext data, cannot be applied directly to the encrypted data. Download all data in the cloud and to decrypt locally is obviously impractical. To solve the above problem, researchers have some general purpose solutions with fully homomorphic encryption or blind RAMs [9] constructed. These methods are not practical due to their high computational cost for both the cloud Sever and users. Proposed scheme to achieve flexible search sub-linear search time and deal with the deletion and insertion of documents. The safe kNN algorithm is used to encrypt the index and query vectors, and in the meantime to ensure precise meaning score calculation between encrypted index and query vectors [10]. To various attacks reflected in different threat models we construct two secure search rules: the dynamic top-k multi-keyword space search procedure in the known cipher text model and the improved dynamic Top-k multi-keyword space search procedure in the known background model.

2. RELATED WORK

Many searching techniques over encrypted cloud data have proposed. S. Deshpande [11] suggested a technique searching over encrypted cloud data using fuzzy keywords. They used Edit distance to quantify keyword similarity and developed two techniques on constructing fuzzy keyword sets to achieve optimized storage and representation overheads. Cong wang et al. [12] Has proposed a method ranked keyword search over encrypted cloud data using keyword frequency and order preserving encryption. It supports only single keywords at a time. Is the keyword frequency deciding document file score. Rank given to every file based on the relevance score of that file. Top ranked files have sent to users instead all files. To enrich search functionality N. Cao et al. [13] Have proposed a scheme supporting conjunctive keywords search. It is privacy – preserving multi-keyword ranked search technique using symmetric encryption. M. Chou et al. [14] proposed a solution for fuzzy multi-keyword search over encrypted cloud data using privacy aware Bed Tree.They used a co-occurrence probability approach to identify useful multi-keywords for publishing data, documents and relevant fuzzy keyword sets constructed using edit distance. They constructed index tree for all data, documents, where each leaf node having the hash value of a keyword, one or two data vectors that represents n- gram of that keyword and bloom filters for each edit distance value.Chi Chen, has proposed a hierarchical clustering method to more search support semantics and the demand for fast passphrase -Search meet in a big data environment [15] .The proposed hierarchical approach clusters the documents on the basis of minimum relevance threshold , and then partitions The resulting hierarchical cluster is reached , therestriction on the maximum size of the cluster [15] . In the search phase can achieve a linear computational complexity compared to an exponential increase in the size of document collection this approach. To verify the authenticity of the search results, a structure called minimum hash sub tree is designed in this paper. The proposed method has an advantage over the traditional method in the Rank Privacy as relevant documents.

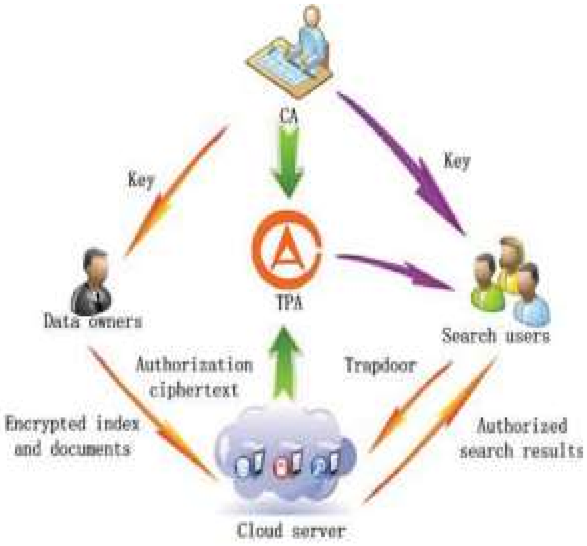

Fig No 1. Hash Sub Tree Designed

3. SYSTEM MODEL

We considered a cloud computing system model involves three different entities. Those are Data Owner, Cloud Service Provider and Data. The responsibility of each entity is as follows: Data Owner (DO): DO has a collection data documents DC= {d1, d2…, dm} with sensitive information to be outsourced to the cloud server. To provide data privacy, the documents are encrypted before outsourcing. DO creates a dictionary based on keywords extracted from the all m documents based on Term Frequency Inverted Document Frequency (TFIDF) [16] which is described in section 4. The dictionary includes synonyms of each keyword from the thesaurus [17]. The dictionaryishavingand keywords, and for each keyword may have t synonyms, so that the dictionary size is n × t. DO creates an index vector for each document based on the keywords extracted from the document. The size of the index vector is equal to the number of keywords in the dictionary that is an. Each dimension in the index vector stores sum of the frequency of keyword and corresponding synonyms in the dictionary is denoted as term frequency (TF) in our system. Index vectors of all documents are encrypted before outsource to the cloud. DO create query vector based on keywords entered by Data user. To provide user privacy, query vector encrypted, as Trapdoor and send to Data user. The data owner sends search access control to the authorized data user.

A. Data users:

Data users are the users who accessing sensitive data from the cloud. The cloud server searches keywords or synonyms related to documents, which are interested to data user and sends to the data owner. The data user receives trapdoor and searches access control of data owner and sends trapdoor and access control to the cloud server to retrieve required documents from the cloud.

B. Cloud Service Provider (CSP):

Cloud server receives encrypted documents and encrypted index vectors from data owner and stores into data owner's cloud storage. Cloud server having the capability to take the data request from user and check the search access control of the user. It will retrieve the documents from cloud storage depending upon the privileges to access number of documents. To increase the document retrieval accuracy from cloud server, the top scored (ranked) documents return to data user from the cloud server. The architecture of multi-keyword synonym query over encrypted cloud data.

C. Threat model:

The cloud server is measured as "honest-but-curious" [18] in our proposed scheme. The cloud server follows the proposed method specification and also examines data in its cloud storage and data which are received from data user during the processing to learn additional information. We consider one threat model for our system with different attack capabilities that is as follows: Known ciphertext model: In this model, the cloud server knows only encrypted documents and encrypted index vectors, which are outsourced from data owner.

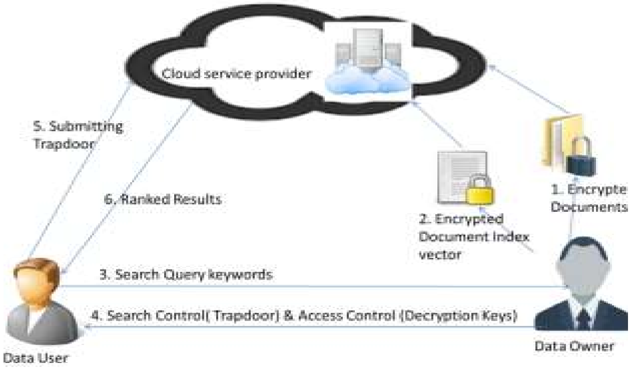

Fig No 2 System Model

4. PROPOSED SYSTEM

The first symmetrical searchable encryption (SSE) scheme and the search of the scheme is linear in the size of the data collection. Proposed formal security definitions for SSE and developed a system based on Bloom filter. It is proposed that two systems (SSE - 1 and 2) that the optimal search time is reached. Your SSE 1 scheme is secure against attacks Chosen- Keyword (CKA1) and SSE -2 is secure against adaptive chosen- keyword attacks (CKA2). These early works are single keyword Boolean search schemes that are very simple in terms of functionality. After plenty of plants have been proposed under different threat models to search various search functions, such as single keyword search, similarity search more keyword Boolean search space and multi keyword search on place, etc. Multi - keyword Boolean search allows achieve the user to enter multiple query keywords to request appropriate documents. Among these works, combining keyword search systems give only the documents that contain all of the query keywords. Disjunctive Keyword Schemes return all documents that contain keywords proposed [19] .Predicate search schemes a subset of the query, both connecting divisive to support search. All these schemes More Keyword retrieve search results based on the presence of keywords, which can provide not acceptable result ranking functionality [20]. Proposed guide can achieve sublinear search time flexible and deal with the deleting and inserting documents. The safe kNN algorithm used to encrypt the index and query vectors, in the meantime accurate relevancy score calculation between encrypted index and query vectors [21] .Ensure to withstand various attacks in different threat models, build two secure search systems: the dynamic top k multi-keyword search scheme selected in the known ciphertext model, and improved dynamic top k multi- keyword space search procedure in the known background model.For our system, we choose the B-tree as indexing data structure to identify the match between search query and data documents. Specially, we use inner data correspondence, i.e., the number of query keywords appearing in document, to evaluate the similarity of that document to the search query. Each document is converted to a balanced B-tree according to the keywords and encrypted using CRSA. Whenever user wants to search, he/she creates a trapdoor for the keywords. Our aim is to design and analyse the performance of multiple keywords ranked search scheme using Commutative RSA algorithm and B-tree data structure for searchable index tree.

1. Commutative Encryption (CRSA):

The RSA cryptosystem is one of the optimum public key cryptography approaches. However, its overall robustness gets limited due to one way encryption and majority of existing RSA schemes suffer from reorder issues. Therefore, in order to make this system least complicated and more efficient, an approach called Commutative RSA has been proposed. In this scheme, the order in which encryption has been done would not affect the decryption if it is done in the same order. Encryption is the standard method for making a communication private. With the many cryptographic approaches, our system follows the commutative RSA algorithm. The mathematical scheme for performing this encryption is described by a pseudo algorithm.

2. BMS Tree Index Construction:

In the process index tree construction, we generate node for each document in the document collection. These nodes are act as leaf nodes in the tree. The internal nodes are formed based on these leaf nodes. The index tree construction process is described in the algorithm 1. An example of BMS index tree for our scheme which is constructed on plaintext. The data structure of the node is defined as (ID, F, child [], DID), where ID is a unique id generated using GenID() function, F is index vector, child[] is pointers to children of the node and DID is a document ID. In the algorithm, we used two variables Current Node Collection and Temp Node Collection to store collection of nodes. Current Node Collection stores the set of currently processing nodes which have no parents and Temp Node Collection stores set of newly formed nodes. Fu[i] always stores the biggest TF value of wiamong its children. The possible largest relevance score of its children is estimated using this technique.

Algorithm 1 Build BMS Index Tree(DC)

```
For each data document Ddid in DC do
    Construct leaf node l for Ddid l.ID=GenID(), l.child[i]=null for i=1,…, b; l.DID=DID, and F[i]=TFDdid,ki for i=1,…, n;
    Insert l to CurrentNodeCollection;
End for While the number of nodes in CurrentNodeCollection is more than 1 do
    For each five of nodes u1, u2, u3, u4, and u5 in
    CurrentNodeCollection do
        Generate a parent node u for u1, u2, u3, u4, and u5with u.ID=GenID(), u.child[i] = uifor i = 1to 5; u.DID = 0, and D[i] = max{ui.F[j] for i=1 to 5} for each j=1 to n;
        Insert u to TempNodeCollection;
    End for
    The remaining nodes (less than 5 nodes) in CurrentNodeCollection generate a parent node u like above;
    Insert u to TempNodeCollection;
    Replace CurrentNodeCollection with TempNodeCollection and then free the TempNodeCollection;
End while
Return only one node, left in the CurrentNodeCollection called the root node;
```

3. Search Process using DFST:

The search process of MSRQE scheme is the recursive function upon the BMS tree name as Depth First Search Technique algorithm. We create a result documents as RankedList, whose element is denoted as (Score, DID). Here, the score is the relevance score between Fdid and query vector Q, which is calculated using formula(1). The RankedList stores top k scored documents to query. The elements of RankedList are in descending order according to score function during the search process. The DFST algorithm is presented in algorithm 2. Kth score is a smallest relevance score in RankedList.

Algorithm 2 DFST(Index Tree Node u)

```
If the node u is not a leaf node then
    If Score(Fu, Q) >kth score then
        Sort the children of u in descending order according to scores of children
        For i=1 to the number of children of u do
            GDFS(u.child[i]);
        End for
    Else
        Return;
    End if
Else
    If Score(Fu, Q) >kth score then
        Delete the element with a smallest relevance score from RankedList;
        Insert a new element (Score (Fu, Q), u.ID) and sort all elements of RankedList in descending order;
    End if
    Return;
End if.
```

Designing searchable encryption schemes, the structures obtained with sub- linear search time, In particular, dynamic search is simple, highly parallel and can easily handle updates. Specifically, for n documents on m keywords and with p cores (processors) indicated available and it supports complex queries and multi- client settings using SSE as a black box. To ensure that our data structures and related business activities which is support for dynamic databases and it support terabyte -Scale databases in these richer / complex encrypted search settings. The dynamic expansion maintains the optimum index size and only basic size information.

4. MRSE FRAMEWORK

For easy presentation, operations on the data documents are not shown in the framework since the data owner could easily employ the traditional symmetric key cryptography to encrypt and then outsource data. With focus on the index and query, the MRSE system consists of four algorithms as follows

1. **Setup(ℓ)** Taking a security parameter ℓ as input, the data owner outputs a symmetric key as SK.

2. **BuildIndex(F, SK**) Based on the dataset F, the data owner builds a searchable index I which is encrypted by the symmetric key SK and then outsourced to the cloud server. After the index construction, the document collection can be independently encrypted and outsourced.

3. **Trapdoor(fW**) With t keywords of interest in fW as input, this algorithm generates a corresponding trapdoor TfW.

4. **Query(TfW, k, I**) When the cloud server receives a query request as (TfW, k), it performs the ranked search on the index I with the help of trapdoor TfW, and finally returns FfW, the ranked id list of top-k documents sorted by their similarity with fW.

The representative privacy guarantee in the related literature, such as searchable encryption, is that the server should learn nothing but search results. With this general privacy description, we explore and establish a set of strict privacy requirements specifically for the MRSE framework. As for the data privacy, the data owner can resort to the traditional symmetric key cryptography to encrypt the data before outsourcing, and successfully prevent the cloud server from prying into the outsourced data.

5. RESULTS AND DISCUSSION

The proposed scheme, data users can achieve different requirements on search precision of privacy by the standard deviation of adjustment that can be treated as a compensation parameter. The comparison of systems with a recent work that achieves high search efficiency.

BDMRS scheme calls the search results by exact calculation of document vector and query vector. Thus, top- k search accuracy of BDMRS scheme is 100 %. But based and similarity Multi- keyword square search pattern, the basic scheme in suffering from loss of precision due to the accumulation of sub-vectors with the index construction . The test is repeated 16 times, and the average accuracy of 91 %. During the search, when the relevance of the node is greater than the minimum relevance in results Rlist, examines the cloud server, the children of the node; otherwise it returns. So many nodes not accessed during a real search. We denote the number of leaf nodes that contain one or more keywords in the query. It is generally greater than the number of documents required k, but far less than the cardinality of the document collection n. As a balanced binary tree, the height of the index n is log will be maintained, and the complexity of the calculation is ranked relevance O (m).

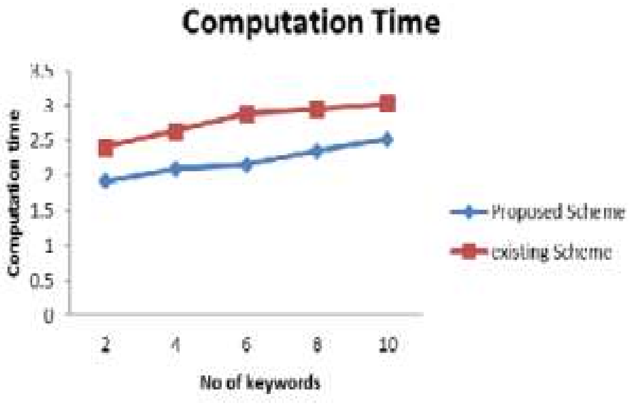

Figure 3: Time Comparison

The graph the comparison of the search computation time in seconds of our proposed system against the RSA based system. For two keywords search, the time taken by the RSA based scheme is approximately 2.5 seconds, whereas our proposed system takes approximately 0.5 seconds less. As the number of keywords increased for search, the computation time for search also increases linearly in both schemes. But CRSA based scheme is found to perform better. Thus it is evident that encryption algorithm CRSA with B Tree as index tree performs better than RSA and B tree Combination

6. CONCLUSION

We intend to provide feasible solutions for multi-keyword synonym ranked query problems over encrypted cloud data while preserving strict system wise privacy in cloud computing paradigm. The first one multi-keyword search, the second synonym based search, third similarity ranked search and the last is efficient data retrieval with BMS tree and DFST searching algorithm. Our example illustration further shows efficient and accurate top k documents retrieval of proposed scheme with sub-linear time complexity. Multi rank keyword search scheme is proposed, which not only supports true multi-keyword search on space, but also the dynamic deletion and insertion of documents. We build a special keyword balanced binary tree as the index. In addition, the search process may be performed in parallel to reduce the time, cost. The security of the system is protected against two threat models through secure top-k retrieval algorithm. The experimental results show the effectiveness of our proposed scheme. Thorough analysis investigating privacy and efficiency guarantees of proposed schemes is given, and experiments on the real-world dataset shows our proposed scheme introduces low overhead on both computation and communication.

7. FUTURE WORK

The future work would concentrate on using Elliptic Curve Cryptography (ECC) encryption technique for better performance. Further, we intend to analyze the behavior of our proposed system(s) for multiuser environment.The dynamic operation such as updating and deletion has to assume with privacy and security policies.

8. REFERENCES

[1] K. Ren, C. Wang, Q. Wang et al., “Security challenges for the public cloud,” IEEE Internet Computing, vol. 16, no. 1, pp. 69–73, 2012.

[2] Cloud Security Alliance, ‘Security Guidance for Critical Areas of Focus in Cloud Computing,’ http://www.cloudsecurityalliance.org, 2009.

[3] R. Brinkman, ‘Searching in encrypted data,’ in University of Twente, PhD thesis, 2007.

[4] S. Kamara and K. Lauter, ‘Cryptographic cloud storage,’ in RLCPS, January 2010, LNCS. Springer, Heidelberg.

[5] A. Singhal, ‘Modern information retrieval: A brief overview,’ IEEE Data

Engineering Bulletin, vol. 24, no. 4, pp. 35–43, 2001.

[6] W. K. Wong, D. W. Cheung, B. Kao, and N. Mamoulis, 'Secure knn computation on encrypted databases,' in Proc. of SIGMOD, 2009.

[7] S. Kamara and K. Lauter, "Cryptographic Cloud Storage," Proc. 14th Int'l Conf. Financial Cryptography and Data Security, Jan. 2010.

[8] Zhihua Xia, Xinhui Wang, Xingming Sun, and Qian Wang, "A Secure and Dynamic Multi-keyword RankedSearch Scheme over Encrypted Cloud Data," IEEE Transactions on Parallel and Distributed Systems, 2015

[9] KawserWazedNafi, TonnyShekhaKar, SayedAnisulHoque, Dr. M. M. A Hashem, "A Newer User Authentication, File encryption and Distributed Server Based Cloud Computing security architecture "Lecturer, Stamford University, Bangladesh, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 10, 2012

[10] Chinua Xia, Xinhui Wang," A Secure and Dynamic Multi-keyword Ranked Search Scheme over Encrypted Cloud Data", DOI 10.1109/TPDS. 2401003, IEEE Transactions on Parallel and Distributed Systems, 2015.

[11] S. Deshpande, "Fuzzy keyword search over encrypted data in cloud computing",World Journal of Science and Technology, vol. 2, no. 10, (2013).

[12] D.X.Song,D. Wagner and A.Perrig,"Practical techniques for searches on encrypted data, In Security and Privacy", 2000. S&P 2000,IEEE, (2000).

[13] N. Cao,C. Wang, M. Li, K. Ren, and W. Lou,"Privacy-preserving multi-keyword ranked search over encrypted cloud data", INFOCOM, 2011 Proceedings IEEE, (2011).

[14] C. Wang, KuiRen, Shucheng Yu, Urs, K.M.R "Achieving usable and privacy-assured similarity search over outsourced cloud data", INFOCOM, 2012 Proceedings IEEE, (2012)

[15] Chi Chen, Xiaojie Zhu, "An Efficient Privacy-Preserving Ranked Keyword Search Method", Member, IEEE, IEEE DOI 10.1109/TPDS.2425407, IEEE Transactions on Parallel and Distributed Systems, 2015.

[16] Yi Yang, Hongwei Li, Wenchao Liu, Haomiao Yao, Mi Wen," Secure Dynamic Searchable Symmetric Encryption with Constant Document Update Cost", School of Computer Science and Engineering, University of Electronic Science and Technology of China, Globecom - Communication and Information System Security Symposium, 2014.

[17] Chi Chen, Xiaojie Zhu, "An Efficient Privacy-Preserving Ranked Keyword Search Method", Member, IEEE, IEEE DOI 10.1109/TPDS.2425407, IEEE Transactions on Parallel and Distributed Systems, 2015.

[18] Hongwei Li, Dongxiao Liu, Kun Jia, and XiaodongLinss"Achieving Authorized and Ranked MultikeywordSearch over Encrypted Cloud Data" School of Computer Science and Engineering, University of Electronic Science and Technology of China. IEEE ICC - Communication and Information Systems Security Symposium,2015

[19] Zhangjie Fu, KuiRen, JiangangShu, Xingming Sun "Enabling Personalized Search over Encrypted OutsourcedData with Efficiency Improvement", DOI 10.1109/TPDS.2506573, IEEE Transactions on Parallel and Distributed System, 2015

[20] Wenhai Sun, Bing Wang, Ming Cao,"Privacy-preserving Multi-keyword Text Search in the Cloud SupportingSimilarity-based Ranking "asia ccs'13, May 8– 10, Hangzhou, China. Copyright 2013 acm 978-1-4503- 1767-2/13/05, 2013.

[21] KawserWazedNafi, TonnyShekhaKar, SayedAnisulHoque, Dr. M. M. A Hashem, "A Newer UserAuthentication, File encryption and Distributed Server Based Cloud Computing security architecture "Lecturer,Stamford University, Bangladesh, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 10, 2012