ISSN: 2321-7782 (Online)
Volume 3, Issue 11, November 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Efficient Multi-Keyword Ranked Search over Public Cloud Storage

Khan Yasmeen Ab Quddus[1]
Department of Computer Engineering
Shree Ramchandra College Of Engineering, SRCOE,
Pune, India

Thombare Baban H[2]
Department of Computer Engineering
Shree Ramchandra College of Engineering, SRCOE,
Pune, India

Abstract: ***Now a days Cloud computing is become more popular as cloud infrastructure provide storage facilities for data owners. Data owner stored their data on public cloud for convenience high availability, easy accessibility and reduced cost in data management. Clouds storage allows users to store data and can be access anywhere, any time by using any technology. Sensitive information such as business related documents, medical records and product information may be stored in a cloud.***

As sensitive data are encrypted before outsourcing to cloud, traditional keyword search techniques are useless. Security and privacy are becoming challenging for cloud infrasture. To keep user data confidential from an untrusted Cloud Service Provider and third parties, a simple way is store data on cloud in encrypted form. The data decryption key should be known to only authorized user. Any Users can search their files using keywords in the cloud. For consumers, they want to find the most relevant products or data, which is highly desirable in the "pay-as-you use" .Many schemes Existing search approaches over encrypted cloud data support only exact or fuzzy keyword search, but not semantics-based multi-keyword ranked search. Therefore, it challenging for researchers for how to enable an effective searchable system that support ranked search and find exact results from encrypted cloud data. This paper proposes an effective approach to solve the problem of multi-keyword ranked. A practically efficient and flexible searchable scheme that supports both multi-keyword ranked search and synonym based search. Vector Space Model use to address multi-keyword search and result ranking.

Vector Space Model (VSM) is used to build index on users document, i.e., each document is expressed as a vector where each dimension value is calculated by the Term Frequency (TF) i.e. number of occurrences of the keyword. One vector is also generated in the incoming query . The vector has the similar dimension with document index and its each dimension value is the Inverse Document Frequency (IDF) weight which is obtained by dividing the total number of documents by the number of files containing the term. Then cosine measure is used to compute similarity of one document to the search query. To improvise search efficiency, a tree-based search index structure which is balance binary tree is used.

Keywords: Ranked search, multi-keyword search, Synonym based search, TFIDF, Fuzzy keyword

I. INTRODUCTION

Now a days, consumer centric cloud computing is a new model in IT infrastructure providing the on-demand high quality applications and services from a shared pool of computing resources. The Cloud Service Provider (CSP) has full control over the outsourced data; So there is need of privacy and security arises.For security purpose sensitive data is encrypted before outsourcing to the cloud server. However Traditional plaintext search methods become useless as data is in encrypted form. The simple and awkward method is downloading all data and decrypt it locally is obviously impractical, because the consumers want to search only the relevant data rather all the data. Therefore it is essential to have an efficient and effective search service over encrypted outsourced cloud data. The existing search approaches like ranked search, multi-keyword search enables the cloud customers to find the most relevant data quickly. Ranked search also reduces the network traffic just by sending the most

relevant data to user request. But In real search scenario it might be possible that user searches with the synonyms of the predefined keywords not the exact or fuzzy matching keywords, due to lack of the user's knowledge about the data. Existing approaches supports only exact or fuzzy keyword search. That is there is no facility of synonym substitution and/or syntactic variation which are the typical user searching behaviors. Therefore synonym based multi-keyword ranked search over encrypted cloud data remains a challenging problem.To overcome this problem of effective search system this paper proposes an efficient and flexible searchable scheme that supports both multi-keyword ranked search and semantic based search. The Vector Space Model is used to address multi-keyword search and result ranking. By using VSM document index is build i.e. each document is expressed as vector where each dimension value is the Term Frequency (TF) weight of each corresponding keyword. Another vector is generated for user query . It has same dimension as that of document index and its each dimension value is the Inverse Document Frequency (IDF) weight. Then cosine measure is used to calculate the similarity between the document and the search query. To enhance the efficiency of the search method we use the extended keyword set with semantic words or natural language words for the keywords. This will ultimately support data retrieval on querying semantic query. User can try searching it by its meaning in natural language Even he doesn't know exact or synonym of keywords of encrypted data.

II. LITERATURE SURVEY

Zhangjie Fu, Xingming Sun, Nigel Linge and Lu Zhou proposes an effective approach to solve the problem of multi-keyword ranked search over encrypted cloud data supporting synonym queries. To address multi-keyword search and result ranking, Vector Space Model (VSM) is used to build document index, that is to say, each document is expressed as a vector where each dimension value is the Term Frequency (TF) weight of its corresponding keyword. A new vector is also generated in the query phase. The vector has the same dimension with document index and its each dimension value is the Inverse Document Frequency (IDF) weight. Then cosine measure can be used to compute similarity of one document to the search query. To improve search efficiency, a tree-based index structure which is a balance binary tree is used [1].

J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, uses the Fuzzy keyword search method that enhances system usability by returning the matching files containing exact match of the predefined keywords or the closest possible matching files based on keyword similarity semantics, when *exact* match fails. They exploit edit distance to quantify keywords similarity and develop an advanced technique on constructing fuzzy keyword sets, which greatly reduces the storage and representation overheads [2].

C. Wang, N. Cao, J. Li, K. Ren, and W. Lou proposes the Ranked search that enhances system usability by returning the matching files in a ranked order regarding to certain relevance criteria (e.g., keyword frequency). It gives a straightforward yet ideal construction of ranked keyword search under the state-of-the-art searchable symmetric encryption (SSE) security definition, and demonstrates its inefficiency. To achieve more practical performance, they propose a definition for ranked searchable symmetric encryption, and give an efficient design by properly utilizing the existing cryptographic primitive, order-preserving symmetric encryption (OPSE) [3].

N. Cao, C. Wang, M. Li, K. Ren, and W. Lou designed a system that solves the challenging problem of privacy-preserving multi-keyword ranked search over encrypted cloud data (MRSE), and establish a set of strict privacy requirements for such a secure cloud data utilization system to become a reality. Among various multi-keyword semantics, they choose the efficient principle of "coordinate matching", i.e., as many matches as possible, to capture the similarity between search query and data documents, and further use "inner product similarity" to quantitatively formalize such principle for similarity measurement [4].

W. Sun, B. Wang, N. Cao, M. Li, W. Lou, and Y. T. Hou present a privacy-preserving multi-keyword text search (MTS) scheme with similarity-based ranking to address this problem. To further enhance the search privacy, they propose two secure index schemes to meet the stringent privacy requirements under strong threat models. In particular, to support multi-keyword

queries and search result ranking functionalities, they proposes to build the search index based on the vector space model, i.e., cosine measure, and incorporate the *TF* × *IDF* weight to achieve high search result accuracy[6].

III. MULTIKEYWORD RANKED SEARCH

The large number of user data and documents store on cloud.,It is crucial for the search service to allow multikeyword query and provide result similarity ranking to meet the effective data retrival .Existing searchable encryption fouces on single keyword search or Boolean keyword search.Existing search approaches cannot accommodate requirement like ranked search,Multikeyword search, semantics based search etc rank search enable cloud customer to find the most relevant information quickly.Ranked search can also reduce network traffic as cloud server sends back only most relevant data Multi keyword search is also very important to improve serach result accuracy as single keyword serach often return coarse search result.In the real world search scenario.it is quite commaon that cloud customer's searching input might to synonyms of predefined keyword ,not the exact of fuzzy matching i.e existing searchable encryption support only exact or fuzzy matching.

To meet the above challenges ,this paper proposes Vector Space Model (VSM which is a practically flexible and efficient search technique .This technique support multi keyword ranked and synonyms que ry processing that uses Vector Space Model (VSM).VSM is used to build document index, i.e., each document is expressed as a vector where each dimension value is calculated by the Term Frequency (TF) i.e. number of occurrences of the keyword. A new vector is also generated in the query phase.. To improve search efficiency, a tree-based search index structure which is balance binary tree is used. Advantages of proposed system includes: a)It Resolve the problem of Synonym based multi-keyword ranked search over encrypted cloud data.b) Using Ranked Search which reduces network traffic.C) Improve Search results accuracy as single keyword search.

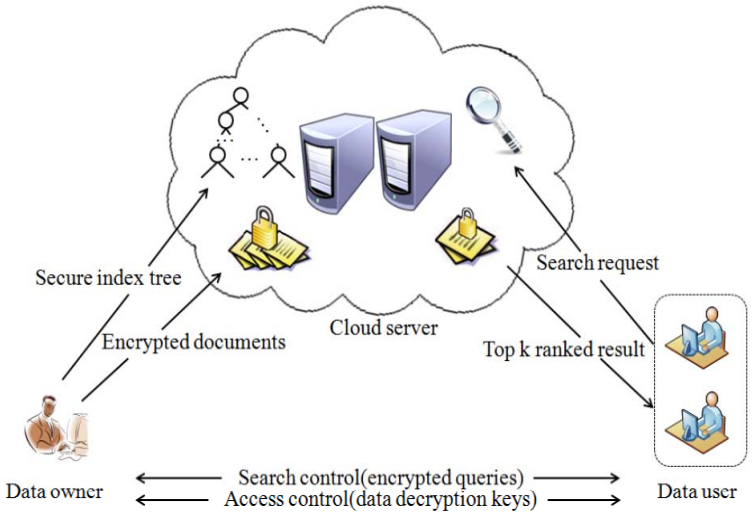

Fig1: Framework of search over encrypted cloud data

Framework of search over encrypted cloud data consist of three different entities: the data owner, the data user and the cloud server, as illustrated in Fig.1. The data owner, individual or enterprise, has a document collection DC which will be outsourced into the cloud. The data owner encrypts DC in the form of C before outsourcing to the cloud. And for the purpose of searching interested data, the data owner will also generate a searchable index I based on a set of distinct keywords W extracted from DC. Then, the encrypted file collection C and searchable index I will be outsourced to the system will generate an encrypted search trapdoor based on the keywords or the synonyms of the predefined keywords entered by the user (has been authorized by data owner).

Given the trapdoor, the cloud server will search the index I and then return search results to the user. The search result is a set of encrypted documents containing the entered keywords,and they are well-ranked according to similarity measures. An additional feature provided by the system is that it can return a certain number of documents instead of all relevant documents. By sending a parameter k together with the search query, the user can get top-k most relevant documents.

IV. THE CONSTRUCTION OF SYNONYM KEYWORD SET IN CLOUD SERVER IMPROVED KEYWORD EXTRACTION TECHNIQUE:

To hunt just the intrigued information as opposed to all the information, magic words are separated firstly from cloud information before outsourcing. This incorporates another content element weighting strategy which serves to include another weighting component that mirrors the noticeability of the pivotal word on the size of unique TF and IDF (term recurrence backwards archive recurrence) technique. The proposed instrument serves to concentrate magic words which can all the more precisely and successfully speak to the elements of content.

a) ***Development of Keyword Set Extended by Synonyms:***

To accomplish a superior semantic quest for outsourced information, the essential word set is reached out by basic equivalent word. Firstly, a typical equivalent word thesaurus is constructed as taking after, (1) Selecting the comparative or normal magic words; (2) Selecting the decisive words which can be semantically substituted. Secondly, the watchword set is stretched out by the developed equivalent word thesaurus.

b) ***Rank function:***

In data retrieval, a ranking function is usually used to evaluate accurate scores of matching files to a user request. Among many of ranking functions, the "TF×IDF" rule is most widely used, where TF (term frequenc y) denotes the occurrence of the term appearing in the document, and IDF (inverse document frequency) is obtained by dividing the total number of documents by the number of files containing the term. That means, TF refers to the importance of the term in the document and IDF shows the importance or rate of distinction in the whole document collection DC. Hence, Rank function provides an efficient Ranked search scheme which protects the sensitive frequency information from leakage.

This Design helps in achieving:

1. Index and Query confidentiality:

The main contrast between the customary plan and proposed plan is that few dummy essential words are acquainted in proposed plan with accomplish better likeness scores. In spite of the fact that sham decisive words bring about amplified measurement of related vectors and matrices. The cryptographic strategy utilized is same as a part of two plans. Subsequently, the proposed plan can secure record privacy and question secrecy in both the danger models.

2. Query unlinkability:

For the dummy keywords that are used, the arbitrarily chosen number i will allow the proposed arrangement to convey distinctive similitude scores notwithstanding for the same interest catchphrases. To control the level of progress the estimation of i can be adjusted hence the level of unlinkability is delivered. Be that as it may, since access illustration is not guaranteed adequately speaking to efficiency, the results are come back from the same requesting will constantly achieve some likeness which could be abused with some fit quantifiable examination by the cloud server. it goes about as a tradeoff that one needs to make amidst capability and security .

c) Searchable Index Tree:

With a specific end goal to expand exactness of inquiry, a tree-based record structure is utilized. Tree-based record structure is an offset double tree. A tree of searchable list is built with the assistance of archive record vectors. With the assistance of produced record tree the related archives can be discovered effortlessly by navigating the tree.

d) Tree Based Search Algorithm:

The sequence of search process for keywords is as follows: The search sequence starts from the root node and when arrives at an internal node, if at least a keyword w is present, it continues to search both subtrees, otherwise stops searching in the

current subtree because it means none of the leaf node contains keyword in search query. When the search process reaches at a leaf node, the process computes the cosine value between the document index vector stored in the leaf node and the query vector as the similarity score. Finally the number of documents that contains the keyword in the search query are returned to the root node. The search complexity for this sequential search is O (r logm) as the height of a balance binary tree with m leaf nodes is logm+1.

V. MODULE DESCRIPTION

Efficient Multi-Keyword Ranked Search system consists of following modules:

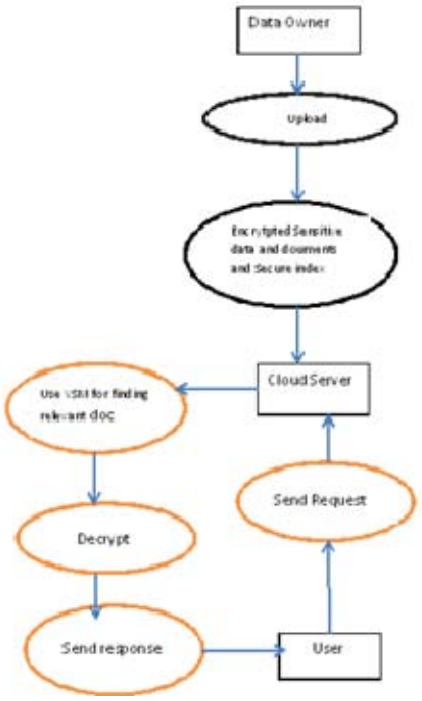

Fig 3: System flow

e) Synonym Expansion :

Synonyms are words with the same or similar meanings. In order to improve the accuracy of search results, the keywords extracted from outsourced text documents need to be extended by common synonyms, as cloud customers' searching input might be the synonyms of the predefined keywords, not the exact or fuzzy matching keywords due to the possible synonym substitution and/or her lack of exact knowledge about the data.

f) Owner Upload Data:

This module is used to help the server to encrypt the document using RSA algorithm and to convert the encrypted document to the zip file with public key and then private key send to user for download.Also it implements following features:

1. Login: In this module owner has to get login by giving valid name and password. User details: Owners will add the registered user details.
2. Upload Files: Data owner generate a public key, auto generate a rank, using keyword extract synonym and encrypt all files
3. Keyword ranking:Data owner can view file ranking, title, public key and file content.
4. User request:Data owner can view requested user and send a public key to requested user t hrough e-mail.
5. User types:Creates user types.
6. User master: Creates user master.

g) Client Module:

This module is used to help the server to view details and upload files with the security .Admin uses the login key at login time. Before the admin logout, he change the login key. The admin can change the password after login and view user downloading details and counting of file request details.It also perform following operation:

1. Key word and search: by using keyword we can search and download the file. Also user can change the password.
2. Request key: User can request a key to the admin. Admin view the secret key and sends public key through email to the user.
3. Download files: using public key user can download the files

h) Rank function module:

In data retrieval, a ranking function is usually used to evaluate accurate scores of matching files to a user request. Among many of ranking functions, the "TF×IDF" rule is most widely used, where TF (term frequenc y) denotes the occurrence of the term appearing in the document, and IDF (inverse document frequency) is obtained by dividing the total number of documents by the number of files containing the term. That means, TF refers to the importance of the term in the document and IDF shows the importance or rate of distinction in the whole document collection DC. Hence, Rank function provides an efficient Ranked search scheme which protects the sensitive frequency information from leakage.

i) Cryptography Module:

1. Data Encryption: This module is used to help the Data owner to encrypt the document using RSA Algorithm and to convert the encrypted document to the Zip file with activation code and then activation code send to the user for download.
2. Data Decryption: Decrypt file using public key user can decrypt the given or downloaded file.

VI. CONCLUSION

The overall performance of the proposed schemes is calculated by implementing the search system on a cloud storage server. The performance of the scheme is evaluated regarding the accuracy of the proposed solution keyword extraction method. The main contributions of proposed system design are summarized in two aspects: multi-keyword ranked search to produce more accurate search results and semantic-based search to support for synonym based queries. Also Extensive experiments on real-world dataset were performed to verify and validate the approach, showing that the proposed system solution is very effective and efficient for multi-keyword ranked searching in a cloud environment. The next focus is to research semantics-based search approaches over encrypted cloud data that support syntactic transformation, anaphora resolution and processing of other natural language technology. The aim is that cloud consumers can find the most relevant and accurate products or data by using the proposed methods of search system.

ACKNOWLEDGEMENT

We would like to thank all the people involved in research in Achieving effective cloud search services and I like to thank my guide Prof. Thombare B.H. Assistant professor of Computer Engineering Department

References

1. Zhangjie Fu, Xingming Sun, Nigel Linge and Lu Zhou, "Achieving Effective Cloud Search Services: Multi-keyword Ranked Search over Encrypted Cloud Data Supporting Synonym Query", IEEE Transactions on Consumer Electronics, Vol. 60, No. 1, February 2014.
2. J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy keyword search over encrypted data in cloud computing," Proceedings of IEEE INFOCOM'10 Mini-Conference, San Diego, CA, USA, pp. 1-5, Mar. 2010.
3. C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure ranked keyword search over encrypted cloud data," Proceedings of IEEE 30th International Conference on Distributed Computing Systems (ICDCS), pp. 253-262, 2010.
4. N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," Proceedings of IEEE INFOCOM 2011, pp. 829-837, 2011

5. Q. Chai, and G. Gong,"Verifiable symmetric searchable encryption for semi-honest-but-curious cloud servers,"

6. W. Sun, B. Wang, N. Cao, M. Li, W. Lou, and Y. T. Hou, "Privacy preserving multi-keyword text search in the cloud supporting similarity based ranking," ASIACCS 2013, Hangzhou, China, May 2013, pp. 71-82, 2013.

7. Sara Paiva, "A Fuzzy Algorithm for Optimizing Semantic Documental Searches", International Conference on Project Management / HCIST 2013.

8. I. H. Witten, A. Moffat, and T. C. Bell, Managing gigabytes: Compressing and indexing documents and images, Morgan Kaufmann Publishing: San Francisco, May 1999, PP. 36-56.

