Multi-Keyword Ranked Search and Data Indexing in Secure Cloud Environment

Shrthi S V, Jagadeesha R

Department of Computer Science Engineering, Kalpataru Institute of Technology, Tiptur, India

E-mail: shrupriya.sv09@gmail.com

Abstract

A Keyword Based Data Indexing and Receiving in Protected Cloud Environment due to the increasing popularity of cloud computing, more and more data owners are motivated to outsource their data to cloud servers for great convenience and reduced cost in data management. However, sensitive data should be encrypted before outsourcing for privacy requirements, which obsoletes data utilization like keyword-based document retrieval. In this paper, we have a tendency to gift a secure multi-keyword hierarchical search theme over encrypted cloud knowledge that at the same time supports dynamic update operations like deletion and insertion of documents. Specifically, the vector house model and, therefore, the widely-used tf×df model are combined within the index construction and question generation. We have a tendency to construct a special tree-based index structure and propose a "greedy depth-first search" algorithmic rule to supply economical multi-keyword hierarchical search. The secure knn algorithmic rule is employed to cypher the index and question vectors, and in the meantime guarantee correct connexion score calculation between encrypted index and question vectors. So, as to resist applied mathematics attacks, phantom terms are extra to the index vector for glaring search results.

Keywords: *Searchable encryption, multi-keyword ranked search, dynamic update, cloud computing*

INTRODUCTION

Cloud computing has been thought-about as a brand new model of enterprise it infrastructure, which may organize immense resource of computing, storage and applications, and change users to get pleasure from present, convenient and on-demand network access to a shared pool of configurable computing resources with nice potency and token economic overhead

[1]. Attracted by these appealing options, each people and enterprises square measure impelled to source their information to the cloud, rather than getting computer code and hardware to manage the information themselves.

Despite of the varied blessings of cloud services, outsourcing sensitive info (such as e-mails, personal health records, company finance information, government documents, etc.). To remote servers brings privacy considerations. The cloud service suppliers (csps) that keep the data for users could access users' sensitive information while not authorization. A general approach to shield the information confidentiality is to inscribe the information before outsourcing [2]. However, this can cause an enormous price in terms of information usability. Let us say, the present techniques on keyword-based info retrieval, that area unit wide used on the plaintext information, cannot be directly applied on the encrypted information. Downloading all the information from the cloud and decode regionally is clearly impractical. In order to address the above problem, researchers have designed some general-purpose solutions with fully homomorphism encryption or oblivious rams [3, 4]. However, these strategies do not seem to be sensible because of their high process overhead for each the cloud sever and user. On the contrary, additional sensible special-purpose solutions, comparable to searchable encoding (se) schemes have created specific contributions in terms of potency, practicality and security. Searchable encoding schemes change the shopper to store the encrypted information to the cloud and execute keyword search over cipher text domain. So far, verdant works are planned below completely different threat models to realize varied search practicality, comparable to single keyword search, similarity search, multi-keyword mathematician search, hierarchical search, multi-keyword hierarchical search, etc. Recently, some dynamic schemes are projected to support inserting and deleting operations on document assortment. These are important works because it is extremely attainable that information homeowners ought to update their data on the cloud server. However, few of the dynamic schemes support economical multi-keyword hierarchic search. This paper proposes a secure tree-based search scheme over the encrypted cloud data, which supports multi-keyword ranked search and dynamic operation on the document collection. Specifically, the vector space model and the widely-used "term frequency (tf) ×

inverse document frequency (idf)" model are combined in the index construction and query generation to provide multi-keyword ranked search.

In order to get high search potency, we have a tendency to construct a tree-based index structure and propose a "greedy depth-first search" algorithmic rule supported this index tree. Because of the special structure of our tree-based index, the planned search theme will flexibly deliver the goods sub-linear search time and trot out the deletion and insertion of documents. The secure knn algorithmic rule is employed to write in code the index and question vectors, and in the meantime guarantee correct connectedness score calculation between encrypted index and question vectors. To resist different attacks in different threat models, we construct two secure search schemes: the basic dynamic multi-keyword ranked search (bdmrs) scheme in the known cipher text model, and the enhanced dynamic multi-keyword ranked search (edmrs) scheme in the known background model.

Our contributions are summarized as follows:

1) We style a searchable secret writing theme that supports each the correct multi-keyword graded search and versatile dynamic operation on document assortment.

2) Due to the special structure of our tree-based index, the search quality of the projected theme is essentially unbroken to index. And in observe, the projected theme is able to do higher search potency by death penalty our "greedy depth-first search" algorithmic rule. Moreover, parallel search may be flexibly performed to additional scale back the time price of search method.

RELATED WORK

Searchable coding schemes alter the shoppers to store the encrypted knowledge to the cloud and execute keyword search over cipher text domain. Thanks to totally different cryptography primitives, searchable coding schemes may be made victimisation public key based mostly cryptography or centrosymmetric key based mostly cryptography [5–7]. Song *et al.* proposed the first symmetric searchable encryption (sse) scheme, and the search time of their scheme is linear to the size of the data collection. Goh proposed formal security definitions for sse and designed a scheme based on bloom filter. The search time of goh's scheme is *o* (*n*), where *n* is the cardinality of the document collection. Curtmola *et al.* proposed two schemes (sse-1 and sse-2) which achieve the optimal search time [5]. Their sse-1 scheme is secure against chosen-keyword attacks (cka1) and sse-2 is secure against adaptive chosen-keyword attacks (cka2). These early works area unit single keyword mathematician search schemes, that area unit terribly straightforward in terms of practicality. Afterward, plentiful works are projected beneath completely different threat models to attain numerous search practicality, love single keyword search, similarity search, multi-keyword mathematician search graded search and multi-keyword graded search etc.

Multi-keyword Boolean search permits the users to input multiple question keywords to request appropriate documents. Among these works, conjunctive keyword search schemes solely come the documents that contain all of the question keywords. Separative keyword search schemes come all of the documents that contain a set of the question keywords. Predicate search schemes area unit planned to support each conjunctive and separative search. Of these multi-keyword search schemes retrieve search results supported the existence of keywords that cannot offer acceptable result ranking practicality.

Ranked search will alter fast search of the foremost relevant knowledge. Causing back solely the top-k most relevant documents will effectively decrease network traffic. Some early works have realised the hierarchical search victimisation order-preserving techniques, however, they are designed just for single keyword search. Cao *et al.* realised the primary privacy-preserving multi-keyword hierarchical search theme, within which documents and queries area unit drawn as vectors of lexicon size. With the “coordinate matching”, the documents area unit hierarchical ac-cording to the amount of matched question keywords. However, Cao *et al*. theme does not take into account the importance of the various keywords, and so is not correct enough. In addition, the search efficiency of the scheme is linear with the cardinality of document collection. Sun *et al.* presented a secure multi-keyword search scheme that supports similarity-based ranking [7].

The authors constructed a searchable index tree based on vector space model and adopted cosine measure together with tf×idf to provide ranking results. Sun *et al.* search algorithm achieves better-than-linear search efficiency but results in precision loss. Orencik *et al.* proposed a secure multi-keyword search method which utilized local sensitive hash (lsh) functions to cluster the similar documents

[2]. The lsh algorithm is suitable for similar search but cannot provide exact ranking. Zhang *et al.* proposed a scheme to deal with secure multi-keyword ranked search in a multi-owner model [2]. In this theme, totally different knowledge use different secret keys to code their documents and keywords whereas approved knowledge users will question while not knowing keys of those different knowledge owners. The authors projected AN "additive order protective function" to retrieve the foremost relevant search results. However, these works do not support dynamic operations.

Practically, the info owner might have to update the document assortment when he transfers the gathering to the cloud server. Thus, these schemes area unit expected to sup-port the insertion and deletion of the documents. There are many dynamic searchable encoding schemes. Within the work of Song *et al*., the every document is taken into account as a sequence of fastened length words, and is singly indexed [7]. This theme supports straight-forward update operations, however, with low potency. Goh projected a theme to come up with a sub-index (bloom filter) for each document supported keywords. Then, the dynamic operations can be easily realized through updating of a bloom filter along with the corresponding document.

However, goh's scheme has linear search time and suffers from false positives. In 2012, Kamara *et al.* constructed an encrypted inverted index thatcan handle dynamic data efficiently. But, this scheme is very complex to implement. Subsequently, as an improvement, Kamara *et al.* proposed a new search scheme based on tree-based index, which can handle dynamic update on document data stored in leaf n-odes. However, their scheme is designed only for single-keyword boolean search. In Cash *et al.* Presented a data structure for keyword/identity tuple named "t-set".

Then, a document can be represented by a series of independent t-sets. Based on this structure, Cash *et al.* proposed a dynamic searchable encryption scheme. In their construction, newly added tuples are stored in another database in the cloud, and deleted tuples are recorded in a revocation list. The final search result is achieved through excluding tuples in the revocation list from the ones retrieved from original and newly added tuples. Yet, Cash *et al.* dynamic search scheme does not realize the multi-keyword ranked search functionality.

EXISTING SYSTEM

In the existing techniques on keyword-based data retrieval, that are wide used on the plaintext information, cannot be directly applied on the encrypted information. Downloading all the info from the cloud and decode regionally is clearly impractical. Of these multi keyword search schemes retrieve search results supported the existence of keywords, that cannot offer acceptable result ranking practicality. However, sensitive information ought to be encrypted before outsourcing for privacy necessities that obsoletes information utilization like keyword-based document retrieval.

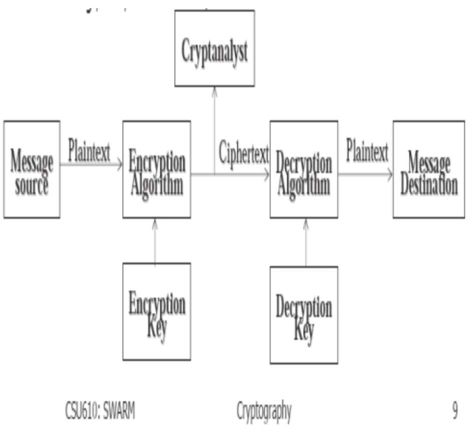

***Fig. 1:** Existing System Algorithms.*

Specifically, the vector space model and the widely-used tf×df model are combined in the index construction and query generation.

PROPOSED SYSTEM

A secure and dynamic multi-keyword stratified search theme over encrypted cloud information we tend to construct a special tree-based index structure and propose a "greedy depth-first search" algorithmic program to produce economical multi-keyword stratified search. The planned theme is able to do sub-linear search time and manage the deletion and insertion of documents flexibly. In depth experiments square measure conducted to demonstrate the potency of the planned theme.

- Abundant works are planned beneath completely different threat models to realize numerous search practicality.
- Recently, some dynamic schemes are planned to support inserting and deleting operations on document assortment.
- This paper proposes a secure tree-based search scheme over the encrypted cloud data, which supports multi keyword ranked search and dynamic operation on the document collection.

Proposed System Algorithms

- Algorithm to provide efficient multi-keyword ranked search .
- The secure knn algorithm is utilized to encrypt the index and query vectors.

- Propose a "greedy depth-first search" algorithm based on this index tree.
- Algorithm achieves better-than-linear search efficiency but results in precision loss.
- The lsh algorithm is suitable for similar search but cannot provide exact ranking.
- {i′s ; ci} ← genupdateinfo (sk; ts; i; up type)) this algorithm generates the update information {i′s; ci} which will be sent to the cloud server.

Advantages

Despite of the various advantages of cloud services, outsourcing sensitive information such as e-mails, personal health records, company finance data, government documents, etc.

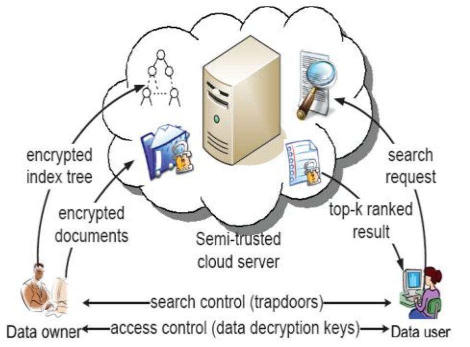

Fig. 1. The architecture of ranked search over encrypted cloud data

***Fig. 2:** System Architecture.*

Tree-based index with the document collection. 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.

Proposed Schemes

Modules

- ➢ Data Owner.
- ➢ Data User.
- ➢ Dynamic Multi-Keyword Ranked Search.
- ➢ Search Efficiency.
- ➢ Privacy-Preserving:
- o Index Confidentiality and Query Confidentiality.
- o Trapdoor Unlinkability.
- o Keyword Privacy.
- ➢ Dynamic Update Operation.

Algorithm

- ➢ Edmrs Scheme.
- ➢ Tree-based Search Algorithm.
- ➢ Secure Scheme.
- ✓ Our proposed search scheme achieves multi-keyword ranked search over encrypted data with high efficiency and search result accuracy.
- ✓ We propose a secure dmrs scheme which meets privacy requirements in the known ciphertext model.
- ✓ Benefiting from tree-based index structure, our search scheme supports dynamic update operation (like deletion and insertion) on documents, which caters to real-world needs and is

superior to most current static schemes.

To alter economical, secure and dynamic multi-keyword hierarchic search over outsourced encrypted cloud knowledge beneath the said models, our system style ought to at the same time bring home the bacon the subsequent style goals.

Dynamic Multi-Keyword Ranked Search

To design a hunt theme over encrypted information that provides not solely effective multi-keyword question and correct result ranking, however, additionally dynamic update on document collections.

Search Efficiency

Our search theme aims to realize higher sensible search potency than linear search by exploring a tree-based index structure associate degreed an economical search algorithmic rule.

Privacy-Preserving

To prevent the cloud server from learning additional information from the dataset, the index tree, and the queries. The specific search privacy requirements are summarized as follows:

Index Confidentiality and Query Confidentiality

The underlying plaintext information (including keywords in the index and query, keywords' tf values stored in the index, and idf values of query keywords) should be protected from cloud server.

Trapdoor Unlinkability

The cloud server should not be able to determine whether two encrypted queries (trapdoors) are generated from the same search request.

Keyword Privacy

The cloud server could not identify the specific keyword in query, index or dataset.

Search Algorithm

The search process of our dmrs scheme starts from the root node with a recursive procedure upon the tree in a special depth-first manner, which is called as "greedy depth-first traverse strategy". Specifically, if the node's similarity score is a smaller amount than or up to the minimum similarity score of the presently selected top-documents, search method returns to the parent node, otherwise, it goes all the way down to examine the kid node. The similarity score of every node is calculated as formula (1), i.e., the dot product of question vector and knowledge vector. This procedure is dead recursively till the objects with top- scores are selected. The search may be done terribly with efficiency, since solely a part of the index tree is visited thanks to the comparatively correct most score prediction.

Index Confidentiality and Query Confidentiality

In dmrs, and are obfuscated vectors, which means the cloud server cannot infer the original vectors or without the secret key. Therefore, index confidentiality and query confidentiality are well protected.

Query Unlinkability

The trapdoor of question vector is generated from random cacophonic operation, which implies same search requests would be reworked into completely different question vectors (trapdoors), therefore, the question unlinkability is protected. However, equipped with capability on chase visited nodes with corresponding similarity scores, the cloud server may be ready to link a similar search requests consistent with a similar similarity scores. Underneath this circumstance, the question unlinkability is unobtainable.

Keyword Privacy

In the known cipher text model, the cloud server is supposed to only know the encrypted document set, index tree and trapdoor. Therefore, without other background information, the cloud server is unable to deduce keywords or tf/idf values from the result similarity scores. However, in enhanced threat model, the cloud server may be equipped with more knowledge like document/keyword frequency statistics of the dataset. Then the cloud server could launch statistical attack to deduce or even identify specific keywords in the query. As an improvement, our future work aims to design a secure scheme that meets all the privacy requirements above even in enhanced threat model.

Dynamic Update Operation

Since our dmrs scheme is designed on a red-black tree data structure, the dynamic operations (like insertion or deletion of a document) could be executed efficiently through structural update on the index tree. Furthermore, since the documents are directly related to the leaf nodes, the whole structure of index tree would change little.

CONCLUSION

In this paper, a secure, economical and dynamic search theme is planned, that supports not solely the correct multi-keyword stratified search, however, additionally the dynamic deletion and insertion of documents. We tend to construct a special keyword balanced binary tree because the index, and propose a “greedy depth-first search” rule to get higher potency than linear search. Additionally, the parallel search method will be dole out to any scale back the time value. The safety of the theme is protected against 2 threat models by victimisation the secure knn rule. Experimental results demonstrate the potency of our planned theme.

There is a unit still several challenge issues in bilaterally symmetrical se schemes. Within the projected theme, the info owner is liable for generating change information and causation them to the cloud server. Thus, the info owner must store the unencrypted index tree and also the information that area unit necessary to cipher the Israeli Defense Force values. Such an energetic knowledge owner might not be terribly appropriate for the cloud computing model. It might be a pregnant, however, tough future work to style a dynamic searchable encoding theme whose change operation are often completed by cloud server solely, in the meantime reserving the flexibility to support multi-keyword graded search. Additionally, because the most of works regarding searchable encoding, our theme principally considers the challenge from the cloud server. Actually, there is a unit several secure challenges in a very multi-user theme. Firstly, all the users sometimes keep an equivalent secure key for trapdoor generation in a very bilaterally symmetrical se theme.

In this case, the revocation of the user is massive challenge. If it is required to revoke a user during this theme, we want to construct the index and distribute the new secure keys to any or all the approved users. Secondly, rhombohedral schemes sometimes assume that everyone the information users area unit trustworthy. It is not sensible and a dishonest knowledge user can result in several secure issues. Let us say, a dishonest knowledge user might search the documents and distribute the decrypted documents to the unauthorized ones. Even more, a dishonest knowledge user might distribute his/her secure keys to the unauthorized ones. Within the future works, we are going to attempt to improve this theme to handle these difficult issues

REFERENCES

1. K. Ren, C.Wang, Q.Wang et al. Security challenges for the public cloud. *IEEE Internet Computing*. 2012; 16(1): 69–73p.
2. S. Kamara, K. Lauter. Cryptographic cloud storage. *In Financial Cryptography and Data Security. Springer*. 2010; 136–149p.
3. C. Gentry. A fully homomorphic encryption scheme. Ph.D. Dissertation, Stanford University; 2009.
4. O. Goldreich, R. Ostrovsky. Software protection and simulation on oblivious rams. *Journal of the ACM (JACM)*. 1996; 43(3): 431–473p.
5. D. Boneh, G. Di Crescenzo, R. Ostrovsky, G. Persiano. Public key encryption with keyword search. *In Advances in Cryptology- Eurocrypt.* 2004; 506–522P.
6. D. Boneh, E. Kushilevitz, R. Ostrovsky, W. E. Skeith. Public key encryption that allows pir queries. *In Advances in Cryptology-CRYPTO*. 2007; 50–67p.
7. D. X. Song, D. Wagner, A. Perrig. Practical techniques for searches on encrypted data. *In Security and Privacy, 2000. S&P 2000. Proceedings*. 2000; 44–55p.