International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

Available online: www.ijmrset.com

e-ISSN: 2394-8752, p-ISSN: 2394-8791

Volume-02, Issue-12, December 2015

A Secure Multi-Keyword Search Based On Ranking Using Selection Algorithm in Clustering

Manivel T[1], Uma V[2]

PG Student, Dept. of I.T, Muthayammal Engineering College, Rasipuram, Tamilnadu, India

Associate Professor, Dept. of C.S.E, Muthayammal Engineering College, Rasipuram, Tamilnadu, India

ABSTRACT:The advent of distributed systems, data owners are motivated to outsource their complex data management systems from local sites to commercial public cloud for great flexibility and economic savings. But for protecting data privacy, sensitive data has to be encrypted before outsourcing which obsoletes traditional data utilization based on plaintext keyword search. Thus, enabling an encrypted cloud data search service is of paramount importance. Considering the large number of data users and documents in cloud, it is crucial for the search service to allow multi-keyword query and provide result similarity ranking to meet the effective data retrieval need. Related works on searchable encryption focus on single keyword search or Boolean keyword search and rarely differentiate the search results. In this paper, for the first time to define and solve 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, 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. This paper propose a basic MRSE scheme using secure inner product computation and then significantly improve it to meet different privacy requirements in two levels of threat models. Thorough analysis investigating privacy and efficiency guarantees of proposed schemes is given and experiments on the real-world dataset further show proposed schemes indeed introduce low overhead on computation and communication.Proposed a multi-keyword search based on ranking using selection algorithm in clustering method. Double Encryption is used to secure data. AES & DES algorithm is used for encryption.

KEYWORDS: Privacy preserving, Multi-Keyword search, Selection algorithm, Double Encryption

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of evaluating data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs or both [8]. Data mining software is one of a number of analytical tools for evaluating data. It allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationships identified.

Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.The Cloud data owners prefer to outsource documents in an encrypted form for the purpose of privacy preserving. The volume of data in data centre has experienced a dramatic growth. Therefore it is essential to develop efficient and reliable cipher text search techniques. A hierarchical clustering method is used for fast cipher text search within a big data environment. To verify the authenticity of search results, a structure called minimum hash sub-tree is designed [1]. Proposed a multi-keyword search based on ranking using selection algorithm in clustering method. Double Encryption is used to secure data. AES & DES algorithm is used for encryption.

II. RELATED WORK

Information Retrieval and secure search

A new framework for confidentiality preserving rank-ordered search and retrieval over large document collections. The proposed framework not only protects document/query confidentiality against an outside intruder but also prevents an untrusted data centre from learning information about the query and the document collection. We present practical techniques for proper integration of relevance scoring methods and cryptographic techniques such as order preserving encryption [2], to protect data collections and indices and provide efficient and accurate search capabilities to securely rank-order documents in response to a query. The proposed methods thus form the first steps to bring together advanced information retrieval and secure search capabilities for a wide range of applications including managing data in government and business operations, enabling scholarly study of sensitive data and facilitating the document discovery process in litigation.

Boolean search

To protect data privacy, sensitive clouddata has to be encrypted before outsourced to the commercial public cloud, which makes effective data utilization service a very challenging task [4]. Although traditional searchable encryption techniques allow users to securely search over encrypted data through keywords, they support only Boolean search and are not yet sufficient to meet the effective data utilization need that is inherently demanded by large number of users and huge amount of data files in cloud. In this paper, this paper to define and solve the problem of secure ranked keyword search over encrypted cloud data [3]. Ranked search greatly enhances system usability by enabling search result relevance ranking instead of sending undifferentiated results and further ensures the file retrieval accuracy. Specifically, we explore the statistical measure approach, i.e. relevance score from information retrieval to build a secure searchable index and develop a one-to-many order-preserving mapping technique to properly protect those sensitive score information

Symmetric-Key Cryptography

The problem of searching on encrypted data and provide proofs of security for the resulting crypto systems. Our techniques have a number of crucial advantages. They are provably secure.They provide provable secrecy for encryption, in the sense that the untrusted server cannot learn anything about the plaintext when only given the cipher text, they provide query isolation for searches meaning that the untrusted server cannot learn anything more about the plaintext than the search result, they provide controlledsearching, so that the untrusted server cannot search for an arbitrary word without the user's authorization, they also supporthidden queries, so that the user may ask the untrusted server to search for a secret word without revealing the word to the server[5]. The algorithms we present are simple, fast (for a document of length, the encryption and search algorithms only need stream cipher and block cipher operations) and introduce almost no space and communication overhead, and hence are practical to use today.

III. PROPOSED ALGORITHM

A. *Design Considerations:*

- Search efficiency
- Retrieval accuracy
- Integrity of the search result
- Privacy requirements

B. *Description of the Proposed Algorithm:*

In this proposed system to define and solve 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. Fig.1,among various multi-keyword semantics [8], we choose the efficient principle of "coordinate matching".

- Double Encryption - AES & DES algorithm
- Clustering method - Multi-Keyword Search based on Ranking using Selection Algorithm
- Even though large amount of data privacy is preserved in cloud
- To improve security and efficiency
- Keyword and privacy ranking method used to retrieve relevance documents

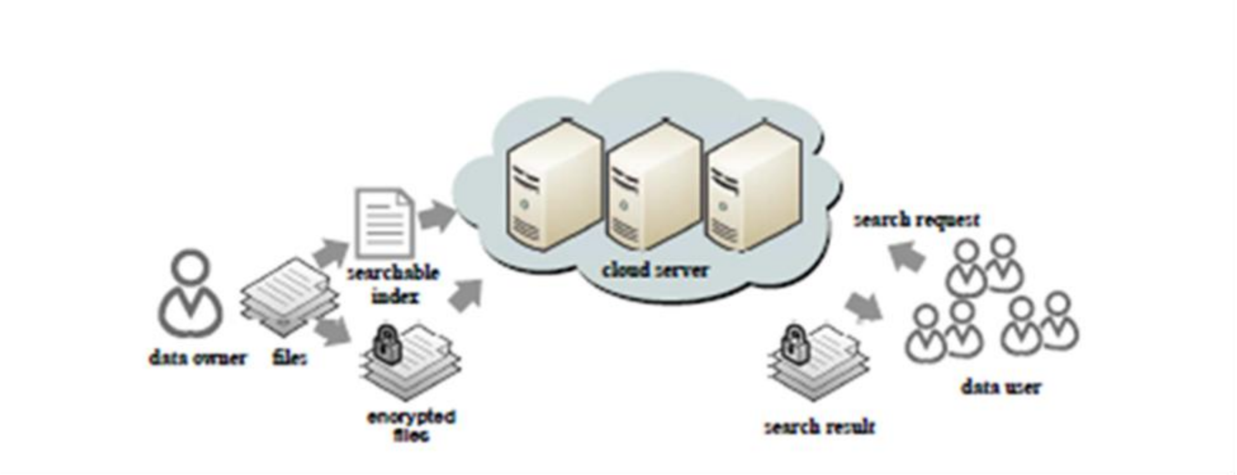


Fig.1 System Architecture

A data flow diagram (DFD) is a graphical representation of the “flow” of data through an information system, modelling its process aspects. A DFD is often used as preliminary step to create on overview of the system which can later be elaborated. DFDs can be used for visualization of data processing (structured design). FIG.2 DFD shows what kind of information will be input to and output from the system, where the data will come to and from, and where the data will be stored [1]. It does not show the information about the timings of process or information about whether the process will operate in sequence or in parallel (which is shown on a flow chart).

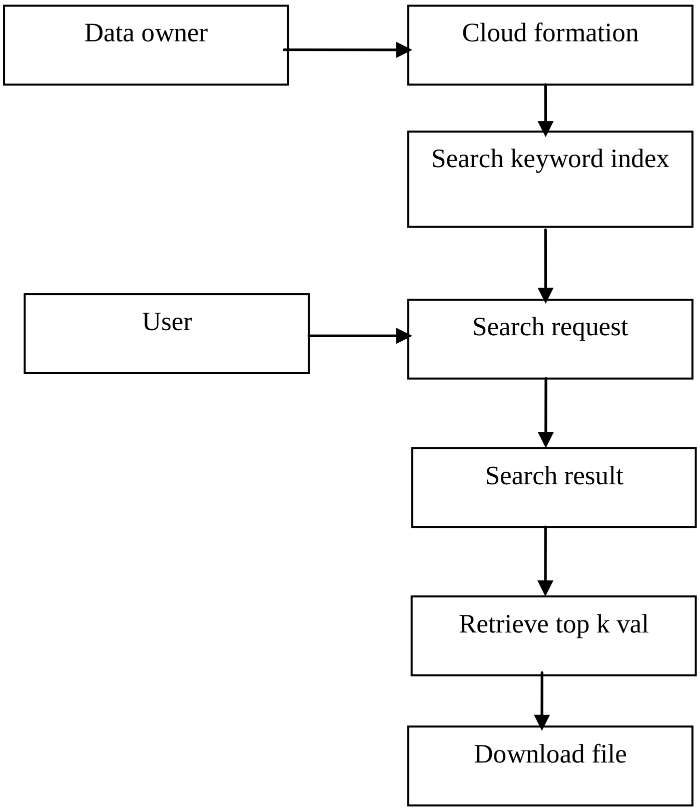


Fig. 2.1 Data Flow Diagram

The proposed algorithm is consists of the following definitions: The basic information of documents and queries are inevitably leaked to the honest-but-curious server since all the data are stored at the server and the queries submitted to the server [6]. Moreover, the access pattern and search pattern cannot be preserved in MRSE-HCI as well as previous searchable encryption. Some basic definitions are there:

Definition 1 (Size Pattern) Let D be a document collection. The size pattern induced by a q-query is a tuple a(D; Q) = (m; jQ1j; _ _ _ ;jQqj) where m is the number of documents and jQij is the size of query Qi.
Definition 2 (Access Pattern) Let D be a document collection and I be an index over D. The access pattern induced by a q-query is a tuple b(D;Q) = (I(Q1); ; I(Qq)), where I(Qi) is a set of identifiers returned by query Qi, for 1 _ i _ q.
Definition 3 (Search Pattern) Let D be a document collection. The search pattern induced by a q-query is am_q binary matrix c(D;Q) such that for 1 _ i _ m and 1 _ j _ q the element in the ith row and jthcolumn is 1, if an document identifier idiis returned by a query Qj.
Definition 4 (known cipher text model secure) Let = (Keygen; Index;Enc; Trapdoor; Search; Dec) be an index-based MRSE-HCI scheme over dictionary Dw, n 2 N, be the security parameter, the known cipher text model secure experiment PrivKkcm
A; _ (n) is described as follows.
1) The adversary submits two document collections D0 and D1 with the same length to a challenger.
2) The challenger generates a secret key fsk; kg by running Keygen(1l(n)).
3) The challenger randomly choose a bit b 2 f0; 1g, and returns Index(Db; skb) !Iband Enc(Db; kb) ! Eb to the adversary.
4) The adversary outputs a bit b0
5) The output of the experiment is defined to be 1
If b0 = b, and 0 otherwise.
We say MRSE-HCI scheme is secure under known ciphertext model if for all probabilistic polynomial time adversaries A there exists a negligible function negl(n) such that Pr(Privkkcm
A;_ = 1) _ 1=2 + negl(n)
Proof The adversary A distinguishes the document collections depending on analyzing the secret key, index and encrypted document collection. Then we have equation 10, where Adv(AD(fsk; kg)) is the advantage for adversary A to distinguish the secret key from two random matrixes and two random strings, Adv(AD(I)) is the advantage to distinguish the index from a random string and Adv(AD(E)) is the advantage to distinguish the encrypted documents from random strings.

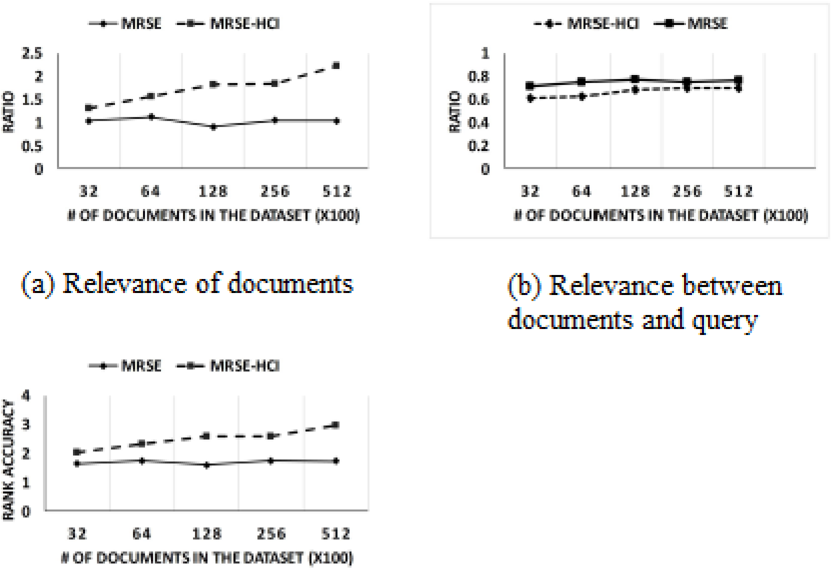
Pr(PrivKkcm
A;_ (n) = 1) = 1=2+
Adv(AD(sk; k)) + Adv(AD(I)) + Adv.(AD(E))

The elements of two matrixes in the secret key are randomly chosen from f0; 1gl(n), and the split indicator S and key k are also chosen uniformly at random from f0; 1gl(n). Given f0; 1gl(n), A distinguishes the secret key from two random matrixes and two random strings with a negligible probability. Then there exits a negligible function negl1(n).

IV. Simulation Results

The Fig.3 describes search accuracy by utilizing plaintextsearch as a standard. Fig.3 (a) illustrates the relevance of retrieved documents. With the number of documents increases from 3200 to 51200, the ratio of MRSE-to-plaintext search fluctuates at 1, while MRSEHCI- to-plaintext search increases from 1:5 to 2. From the Fig.3 (a), we can observe that the relevance of retrieved documents in the MRSE-HCI is almost twice as many as that in the MRSE, which means retrieved documents generated by MRSE-HCI are much closerto each other. Fig.3(b) shows the relevance between query and retrieved documents[7]. With the size of document set increases from 3200 to 51200, the MRSEto-plaintext search ratio fluctuates at 0:75. MRSE-HCIto plaintext search ratio increases from 0:65 to 0:75accompanying with the growth of document set size.

From the Fig.3 (b), we can see that the relevance between query and retrieved documents in MRSEHCI is slightly lower than that in MRSE. Especially, this gap narrows when the data size increases since a big document data set has a clear category distribution which improves the relevance between query and documents. The tradeoff parameter a, is set to 1, which means there is no bias towards relevance of documents or relevance between documents and query[1].



(a) Relevance of documents

(b) Relevance between documents and query

(c) Overall evaluation

Fig.3 Search precision

From the result, we can conclude that MRSEHCI is better than MRSE in rank accuracy. Fig. 4 describes the rank privacy [9]. In this test, no matter the number of retrieved documents, MRSE □ HCI has better rank privacy than MRSE. This mainly caused by the relevance of documents introduced into search strategy.

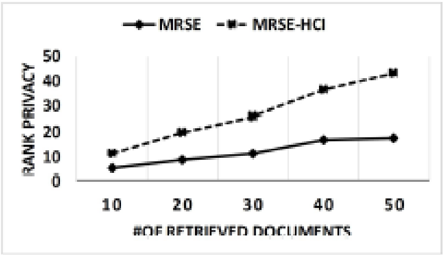


Fig.4 Rank privacy

V. Conclusion and Future Work

In our future work, the software executes successfully by fulfilling the objectives of the project. Further extensions to this system can be made required with minor modifications [7]. This project presents to design and develop an efficient service to protect users' data privacy is a central question of cloud storage.The invention can be implemented in digital electronic circuitry or in computer hardware, firmware, Software or in combinations of them. Apparatus of the invention can be implemented in a computer program product tangibly embodied in a machine-readable storage device for execution by a programmable processor and method steps of the invention can be performed by a programmable processor executing a program of instructions to performfunctions of the invention by operating on input data and generating output. 1

REFERENCES

1. Chi Chen, Xiaojie Zhu, IEEE, PeisongShen, IEEE, J.Hu, S.Guo, Z.Tari, and Albert Y. Zomaya, "An Efficient Privacy-Preserving Ranked Keyword Search Method," DOI 10.1109/TPDS.2015.2425407, IEEE Transactions on Parallel and Distributed Systems.
2. Confidentiality-Preserving Rank-Ordered Search A. Swaminathan,†Y. Mao,† G.- M. Su,† H. Gou,†A. Varna,† S. He,†M. Wu,† and D. Oard,‡ *IEEE J. ACM*, vol. 45, no. 6, pp. 965–982,1998.
3. Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data Cong Wang, Student Member, IEEE, Ning Cao, Student Member, IEEE, KuiRen, Senior Member, IEEE,Wenjing Lou, Senior Member, IEEE
4. Privacy Preserving Keyword Searches on Remote Encrypted DataYan-Cheng Chang and Michael Mitzenmacher Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
5. D. X. D. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in Proc. S & P, BERKELEY, CA, 2000, pp. 44-55.
6. S. Grzonkowski, P. M. Corcoran, and T. Coughlin, "Security analysis of authentication protocols for next-generation mobile and CE cloud services," in Proc. ICCE, Berlin, Germany, 2011, pp. 83-87.
7. Y. C. Chang, and M. Mitzenmacher, "Privacy preserving keyword searches on remote encrypted data," in Proc. ACNS, Columbia Univ, New York, NY, 2005, pp. 442-455.
8. C. Wang, N. Cao, K. Ren, and W. J. Lou, Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data, IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 8, pp. 1467-1479, Aug. 2012.
9. S. Zerr, D. Olmedilla, W. Nejdl, and W. Siberski, "Zerber+R: top-k retrieval from a confidential index," in Proc. EDBT, Saint Petersburg, Russia, 2009, pp. 439-449.