International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878, Impact Factor (2015): 3.791

A Secure Multi-Keyword Ranked Search by Using Conditional Random Field and Blind Storage over Encrypted Mobile Cloud Data

Megha Baraskar[1], Dr. Kishor Kolhe[2]

[1]PG Student, Department of Information Technology, MITCOE Pune, India
[2]Department of Information Technology, MITCOE Pune, India

Abstract: ***Due to the ability of providing access the data anywhere and anytime all over the globe the cloud computing is becoming trend in IT industry. As an accomplishment in flexibility of data the problem gets with safety as in term of data security and privacy. Cloud is being straightforward still considered as interested in the client data. To solve above problems in cloud, blind storage techniques with searchable encryption is used. For keyword extraction system propose a new CRF method. Utilization of Cosine similarity algorithm as replacement for Euclidean distance algorithm is done. Cosine Similarity algorithm is used for calculating the similarity between documents and Conditional Random Field (CRF) for finding the important keywords. The indexes of the blocks are kept which are created by dividing the files in various parts by blind storage. These indexes are not visible to cloud provider by which the cloud provider is totally unaware of data. By analysis we shows that the derived results are improved as well as those are meaningful of keyword minimizing the traffic in overall network and hence server provides more specific application productivity.***

Keywords: Searchable encryption, cloud computing, attributes based encryption, CRF, Cosine Similarity, Blind Storage

1. Introduction

Cloud computing provides networks, servers, storage, applications as well as services which is universal, reliable, on-request network access to shared Cloud computing. This all are provided quick with minimum effort of management and without interference of service provider. A remote server called as cloud provides services by internet for businesses with utilizing computer software. This functionality mitigates expenses spent by user annually or monthly subscription. Cloud is a centred server and causes all the sensitive information like emails, personal health records, private videos as well as photos, company finance data, government documents etc. is stored in centralized manner. The cloud server may not be fully trusted in a cloud environment due to a number of reasons [10], [11]. The cloud server leak data, information to the unauthorized entities or data may be hacked. Data is only confidential as well as secure when it encrypted before outsourced. This gives end-to-end information security in cloud. Information encryption is very difficult task on huge-scale but encryption is efficient way to information used in cloud. Along, data owners shares data with huge numbers of user. But User seeks only limited information which they want access form cloud within a particular session. Searching specific information in large amount of data is very time consuming. There is a technique to search information faster than normal search i.e. keyword-based search. This method helps to user for finding particular information with plain text and avoids non related information from cloud. But, data encryption is one restriction for users' capability to find out information on the basis of keyword search because of need of security to keyword or data denied the search on the basis of plain text to access encrypted cloud information. There is another way of search called ranked search. Raked search hugely enhances system performance through common matching information. In ranked search information is arranged on the basis of some related criteria such as keyword frequency.

Another way is to search the information on the basis of Euclidean distance algorithm. But Euclidean distance have disadvantage is that the focus on circulated about the specimen mean in a round way. Were the appropriation to be determinedly non-round, for case ellipsoidal, then would expect the likelihood of a 'test point' fitting in with the set to depend not just on the separation from the specimen mean additionally on the direction. Qin Liu [6] points privacy issues and efficiency and decrease the communication cost. Author introduce scheme called efficient information retrieval (IR) to reduce query cost for ranked query. This is based on aggregation and distribution layer (ADL) a middleware layer between the data user and the cloud.

Conditional Random Field (CRF) model is specially works on specific features of document. Numerous features of documents are sufficiently as well as effectively utilized by CRF i.e. a state of art sequence labelling method for efficient keyword extraction. Contents of document analysed and comprehended initially within processing of manual assignment of keyword to a document and determined keywords which describe the meaning of document.

Cosine Similarity algorithm for ascertaining similitude in the middle of documents and CRF for finding the vital keywords are used. The indexes of the blocks are kept which are created by dividing the files in various parts by blind storage. These indexes are not visible to cloud provider by which the cloud provider is totally unaware of data.

The remainder of this paper is organized as follows. In Section 2, the system model, security requirements and design goal are formalized. In Section 3, we recap trapdoor generation, relevance scoring, TF-IDF, CRF. In Section 4, we propose the system Performance evaluation in Section 5, respectively. In Section 6, we present related work. Finally, we conclude this paper in Section 7.

2. System Model, Security Requirements and Design Goal

2.1 System Model

As shown in Fig. 1, the system model in the EMRS consists of three entities: data owner, search users and cloud server. The data owner keeps a large collection of documents which will be uploaded on a cloud server in an encrypted form. In the system, It is the one who shares the private key with end user for decryption of the document. The data owner will create a document index and will upload its encrypted version on cloud.

The system achieves the following:

1. The document privacy in such a way that even cloud should not get the actual data.
2. The efficient search mechanism.

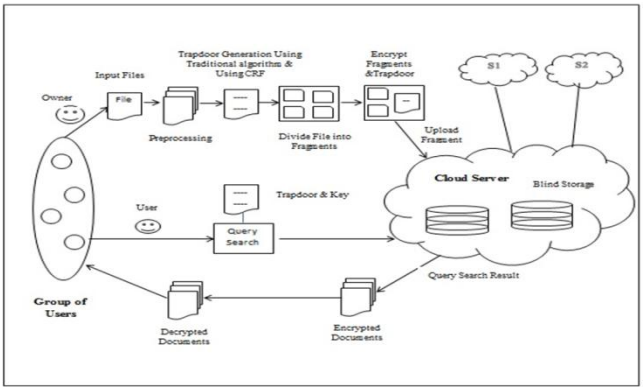

Figure 1: System Model

When a search user wants to do search over the encrypted documents, He /she receive the secret key from the data owner. Then, she chooses a keywords which contains l interested keywords .Finally, the search user sends request to the cloud server to request the most relevant results.

Upon receiving search from the search user, the cloud server uses the index in the blind storage and computes the relevance scores with the encrypted query vector. Then, the cloud server sends back top documents that are most relevant to the searched keywords. The search user can use these descriptors to access the blind storage system to retrieve the encrypted documents. An access control technique, e.g., attribute-based encryption is implemented to manage the search user's decryption.

2.2 Security Requirements

In the Multikeyword ranked search, we consider the cloud server may be curious but honest that means it executes all the tasks assigned by the data owner and the user correctly. Still it is very curious about the uploaded data in its storage and the received trapdoors to obtain additional information.

System aims to provide the following four security requirements:

Confidentiality of Documents and Index: Documents and index should be encrypted before being uploaded to a cloud server.

Trapdoor Privacy: search user would like to keep his/her searches from being disclosed to the cloud server, the cloud server should be prevented from knowing the exact keywords contained in the trapdoor of the search user.

Trapdoor Unlinkability: The trapdoors should not be linkable, which means the trapdoors should be totally different even if they contain the same keywords. Ie the trapdoors should be randomized. The cloud server cannot realise any associations between two trapdoors.

Concealing Access Pattern of the Search User: Access pattern is the sequence of the searched results. In the system the access pattern should be totally concealed from the cloud server. Specially, the cloud server should not know the total number of the documents stored on it nor the size of the searched document even when the search user retrieves this document from the cloud server.

2.3 Design Goal

To provide efficient and secure multi-keyword ranked search over encrypted mobile cloud data via blind storage system, system has following design goals:

Multi-Keyword Ranked Search: multi-keyword search over encrypted mobile cloud data support better experience and achieve relevance-based result ranking.

Search Efficiency: It achieve linear search with better search efficiency.

Con dentiality and Privacy Preservation: To prevent the cloud server from learning any additional information about the documents and the index, and to keep search users' trapdoors secret, system cover all the security requirements.

3. Preliminaries

CRF:

(1) CRF model training:

The input is a set of feature vectors in above step. CRF model that is used to label the keyword type. In particular CRF model, a word or phrase can be used as an example, and the keyword is annotated by one kind of labels, such as ‘KW_B’, ‘KW_I’, ‘KW_S’, ‘KW_N’, and ‘KW_Y’. The tagged data are used to training the CRF model in advance. The output is a CRF model file in the CRF++.

(2) CRF labeling and keyword extraction:

The features are extracted from preprocessed documents. Then, prediction is done according to the keyword type by using the CRF model. And the keywords from the document are extracted.

4. Proposed Scheme

The architecture of the system is depicted in depth in figure 1. The owner of data will upload the document or collection of documents over cloud. Before outsourcing, the data must be encrypted and this will be taken consideration by owner of data. In between, the trapdoors will likewise be produced for documents to empower the result and index table will be created.
Documents will be then outsourced to the cloud. The distributed storage architecture will be taken after on cloud which is known as blind storage. Every document will be broken into blocks of interval say 1KB. These blocks will be stored on various locations or on distinctive blind nodes. The index table will be kept up for these documents so these can be reassembled later.

While in other side the end user will have the secrete keys which will be used for decrypting the received documents. Before that user will search for important files and will produce search trapdoors. After the request cloud will revert the most suitable document depending upon the TF-IDF concept and the outcome will be ranked already. Proposed implementing system is capable of multi keyword search and priority of results will be depending on order of keywords presented by user.

In this section, we propose the detailed multikeyword ranked search. Since the encrypted documents and index both are stored in the blind storage system,

Modules 1: Data Owner

a. Registration & Login:

First fill up the information like first name, last name, mail id and so on .Then create your login id .Then use valid username and password for logging.

b. File Upload & Download:

In file uploading, select the File or Folders .After selection of files, it performs the pre-processing and uploads the encrypted blocks at cloud server. In Download, enter the search keywords and get similar files.

c. File Processing and use of CRF(Conditional Random Fields):

This is statistical modelling method which can applied often in a pattern recognition and machine learning, they are used for structured prediction. Whereas an ordinary classifier forecasts a label for a single sample without regard to "neighbouring" samples, a CRF can take context into account; e.g., the linear chain CRF popular in natural language processing guesses sequences of labels for sequences of input samples.

d. File Index Generation and Trapdoor Generation

Generate the TF-IDF of File and Index using Index algorithm generation and send the data as Trapdoor to end users securely.

e. File Encryption & Decryption:

In Encryption, select file and encrypt that file using AES key for security purposes. In decryption, select encrypted file and decrypt that file using same key for getting original file.

Module 2: Data User

a. Registration & Login:

Register with valid details and with organization details, then login using appropriate username and password.

b. Multi Keyword Search

When user wants information about something then he performs multi keywords search over encrypted data stored in cloud server.

c. File Decryption:

In decryption decrypt the file using valid key.

Module 3.Cloud Server

d. Blind Storage of Data

In Blind storage of data, server does not have idea about the how the data or blocks are stored (random storage of data).
Mathematical Model

Given mathematical model is basically deals with complete data sharing model in group. Each step represent as described the below.

1) Document similarity

$$Sim(d1,d2) = \frac{\sum_{i=1}^{n} d1_i * d2_i}{\sqrt{\sum_{i=1}^{n}(d1_i)^2} * \sqrt{\sum_{i=1}^{n}(d2_i)^2}}$$

d1and d2: are document vector

The cosine similarity, Sim(d1,d2) is represented using a dot product and magnitude.

CRF:
a) Length of the word

$$\frac{Len(Word)}{Max\_Len}$$

b) Part-Of-Speech of word or phrase,

$$pos = \begin{cases} pos{=}1 & \text{if one of word in a phrase is n} \\ \text{pos} = 0 & \text{otehrwise} \end{cases}$$

c) TF*IDF

$$\frac{Freq(Word)}{Max\_Freq} \times \log_2 \frac{N+1}{n+1}$$

d)The position of the first appearance of the word

$$\#(Word)/\sum Word_i$$

A. Experimental Setup:

The system is built using Java framework (version jdk 1.8) on Windows platform, MySQL 5.0 and SqlYog 5.0 for database .Net beans (8.0.2) is use as development tool. The current system doesn't require any specific hardware to run. Any standard machine is capable of running the application.

5. Performance Evaluation

The fig. A shows the time comparison between document search using Euclidean Distance, cosine similarity and document search using CRF &cosine similarity.

Table 1: Comparative readings for document search

	Euclidean Distance	Cosine Similarity	CRF & Cosine Similarity
Time (in ns)	990000000	700000000	60000000

The document search using CRF &co-sine similarity take less time than document search using Euclidean Distance and document search using cosine similarity, in-crease the performance.

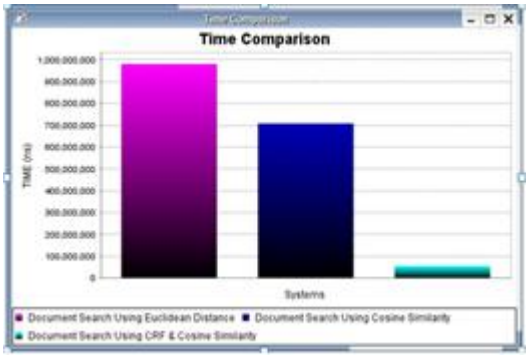

Figure a: Time Comparison Graph

Table 2: Comparative readings for document search

	Existing system	Proposed System
Memory Space (in bytes)	340	145

The figure B shows the data storage comparison between existing system (Traditional Trapdoor) and proposed system (Trapdoor using CRF). The data storage in proposed system (Trapdoor using CRF) in low.

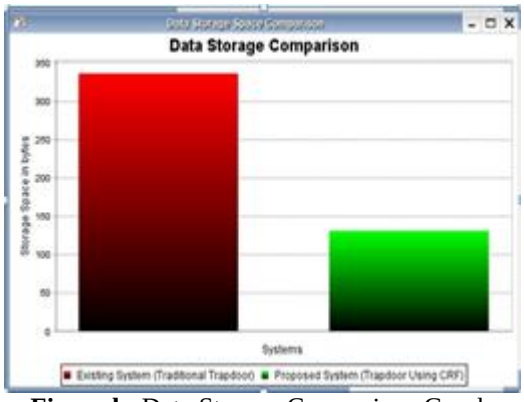

Figure b: Data Storage Comparison Graph

Table 3: Comparative document search for 4 keywords

	Number of Keywords (Dictionary Words)			
	500	750	1000	1500
Euclidean Distance (Time in Ms)	1512	2020	2143	2527
Cosine Similarity (Time in Ms)	1469	1999	2096	2502
CRF and Cosine Similarity (Time in Ms)	1306	1529	1590	1738

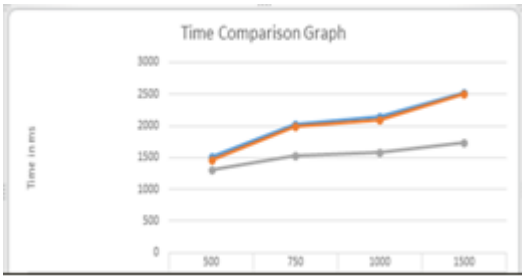

Figure c: Time Comparison Graph

The above fig C graph shows time required to search same number (w = 4) of keywords from different size of keyword dictionary with the same number of documents, In Fig.1 X-axis shows Number of Keywords while Y-axis shows time required to search the keywords in ms.

Table 4: Time comparison Graph (different keywords)

	Number of Keywords (Dictionary Words)			
	5	10	15	20
Euclidean Distance (Time in Ms)	2527	2540	2596	2602
Cosine Similarity (Time in Ms)	2462	2512	2569	2588
CRF and Cosine Similarity (Time in Ms)	1412	1521	1591	1602

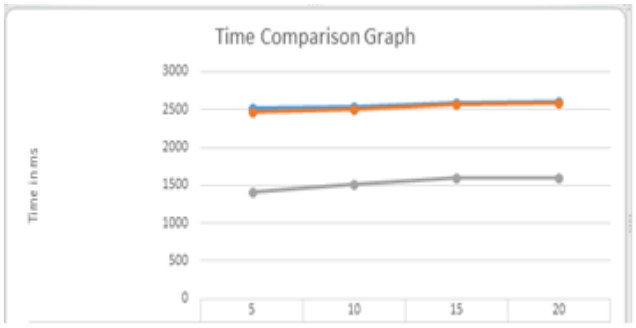

Figure d: Time Comparison Graph having same number of Dictionary Words

The above graph, figure D shows time required to search different number of keywords from same size of keyword dictionary (1500 keywords) with the same number of documents, In Fig.2 X-axis shows Number of Keywords while Y-axis shows time required to search the keywords in ms.

6. Related Work

In paper [1] for achieving effective retrieval of remotely stored encrypted data for mobile cloud computing, authors designed and implemented new ranked fuzzy keyword search method. A test on a real database set (RFC) is performed and demonstrated the proficiency of their suggestion and a correlation with the latest searchable encryption algorithm given a critical upgrade of the index development speed of plan.

Authors proposes [2] the 1st choas based searchable encryption approach which additionally permits both ranked and fuzzy keyword searches on the encoded information put away in the cloud. Given methodology ensures the security and secrecy of the client. This plan is executed and assessed utilizing two databases: RFCs and the Enron database. Tests have been performed to demonstrate the proficiency of the proposition.

Author developed a Searchable Encryption CP-ABE (SE-CP-ABE) access control method by mixing of both security holomorphic encryption algorithm with traditional CP-ABE algorithm and provided security observation as well as examined observation for the method. Outcomes of technique demonstrate that SE-CP-ABE method is assures security of CP-ABE as well as deploys ciphertext retaining and mitigates time of retaining [3].In paper [4] author presented cryptographic method to overcome the issue of searching over encrypted data as well as given evidences of security. This method gives confidentiality for encryption in case of unreliable server which is unknown related with the plaintext if just ciphertext is provided.

Author proposed [5] an effective index to enhance the search performance and accept the blind storage method to hide access pattern of the searching user. Observations of method show that method is able to gain secrecy of documents as well as dex, trapdoor confidentiality, trapdoor unlinkability and hiding access pattern of user.

In [7], [8] authors proposed search authority in SPE using attribute-based encryption technique.

7. Conclusion

The proposed system is used a multi-keyword ranked search scheme for accurate, efficient and secure search on encrypted mobile cloud data. Proposed scheme can effectively achieve user's confidentiality of documents as well as index, trapdoor unlink-ability, trapdoor privacy, and concealing access pattern of the search on the basis of security analysis. Proposed scheme can achieve better efficiency on the basis of extensive performance evaluations shown in terms of the functionality and computation overhead as compared with existing ones.

By using CRF algorithm, there is reduction in the time consumption and Cosine similarity algorithm is used to increase the accuracy of proposed system.

References

[1] Abir Awad, Adrian Matthews, Brian Lee," Secure Cloud Storage and Search Scheme for Mobile Devices", 17th IEEE Mediterranean Electro technical Conference, Beirut, Lebanon, 13-16 April 2014.
[2] Abir Awad, Adrian Matthews, Yuansong Qiao, Brian Lee," Chaotic Searchable Encryption for Mobile Cloud Storage", IEEE Transactions on Cloud Computing,, no. 1, pp. 1, July2015.
[3] An-Ping Xiong, Qi-Xian Gan, Xin-Xin He, Quan Zhao, "A searchable encryption of CP-ABE scheme in cloud storage", Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2013 10th International Computer Conference on Dec. 2013.
[4] Dawn Xiaodong Song David Wagner Adrian Perrig, ''Practical Techniques for Searches on Encrypted Data,'' Defense Advanced Research Projects Agency under DARPA contract N6601-99- 28913
[5] B. Zhang and F. Zhang, ''An efficient public key encryption with conjunctive-subset keywords search,'' J. Netw. Comput. Appl., vol. 34, no. 1, pp. 262–267, Jan. 2011.
[6] Q. Liu, C. C. Tan, J. Wu, and G. Wang, ''Efficient information retrieval for ranked queries in cost-effective cloud environments,'' in Proc. IEEE INFOCOM, Mar. 2012, pp. 2581–2585
[7] W. Sun, S. Yu, W. Lou, Y. T. Hou, and H. Li, ''Protecting your right: Attribute-based keyword search with fine-grained owner-enforced search authorization in the cloud,'' in Proc. IEEE INFOCOM, Apr./May 2014, pp. 226–234.
[8] Q. Zheng, S. Xu, and G. Ateniese, ''VABKS: Verifiable attribute based keyword search over outsourced encrypted data,'' in Proc. IEEE INFOCOM, Apr. 2014, pp. 522–530.
[9] Song DX, Wagner D, Perrig A (2000), "Practical techniques for searches on encrypted data." In: Proceedings of the IEEE Symposium on Security and Privacy, IEEE, pp 44–55.
[10] Korkmaz, T. Tek, S. "Analyzing Response Time of Batch Signing." Journal of Internet Services and Information Security, Vol. 1, 2011, No. 1, pp. 70-85.
[11] Fukushima, K. Kiyomoto, S. Miyake, Y. "Towards Secure Cloud Computing Architecture - A Solution Based on Software protection Mechanism". Journal of Internet Services and Information Security, Vol. 1, 2011, pp. 4-17