I J C T A, 9(15), 2016, pp. 7205-7210

Privacy Preserving Similarity Based File Retrieval Through Blind Storage

C.M. Manoj* and G.K. Sandhia**

ABSTRACT

The main objective of this paper is to eradicate the confidentiality and privacy concerns of cloud computing and to provide a better search performance over encrypted data. The reason behind using cloud computing is to outsource the data to external cloud for scalable data storage. The data that is to be outsourced might contain confidentiality and privacy concerns. Therefore it becomes mandatory to encrypt these data before outsourcing into the external cloud. Though this solution provides security, it leads to a new problem of performing an accurate search over encrypted data. To overcome this issue, the proposed system supports multi-keyword search over encrypted data. It uses TF - IDF technique to improve the search efficiency by ranking the searched content. And a good level of privacy is provided by the concept of blind storage where the encrypted data are divided into fixed size blocks and stored in random locations in cloud server.

1. INTRODUCTION

Cloud Computing is a technique of using a network as a remote server for hosting services, processing and storing data, instead of having the local servers or using a personal computer. Cloud computing eradicates the hardware limitations of local systems by making use of its scalable and reliable resources. In general, the users makes use of cloud computing for outsourcing data, e.g., when user upload some private data to cloud, there are chances for it to contain sensitive information like personal information, images, etc., leading to confidentiality and privacy issues[1] thus requiring high level of security and protection. So it makes it necessary to encrypt such data before uploading to cloud. But when other users tries to access such data by searching, it leads to salient difficulties due to the complications involved in performing search over encrypted data. So in recent years many research have been conducted on performing search over encrypted data[2]-[5].

Recently number of researches are done on multi-keyword search over encrypted data for the betterment of search results. Cao et al. [6] have proposed a scheme for performing ranked search over encrypted data. Naveed et.al. [7] proposed a scheme for performing searchable encryption through blind storage in order to avoid users from knowing access pattern of search results.

In general, a good search system should have the following characteristics to meet the practical requirements.

- A search system should not only search with the help of keywords entered by the user but also the synonyms of the keywords should be considered.
- The search system should support multi-keyword search like that of google search.
- It should not only consider name of the file searched but also the content present in it.
- A good search system should be capable of performing search over encrypted data[8][9].

* Post Graduate Student, *Email: cm.manoj@yahoo.in*

** Assistant Proffesor Department of Computer Science and Engineering SRM UNIVERSITY Chennai-603203, India.

- Similarity based retrieval of results should be used [10]. Where the most matching results are shown.
- Relevance based ranking order should be followed while retrieving the searched data. Where, relevance ranking shows how frequently a term is occurring in a file and based on this term frequency, ranking of the search result is shown to the user.

2. PROBLEMS PERTAINING IN CLOUD SERVER

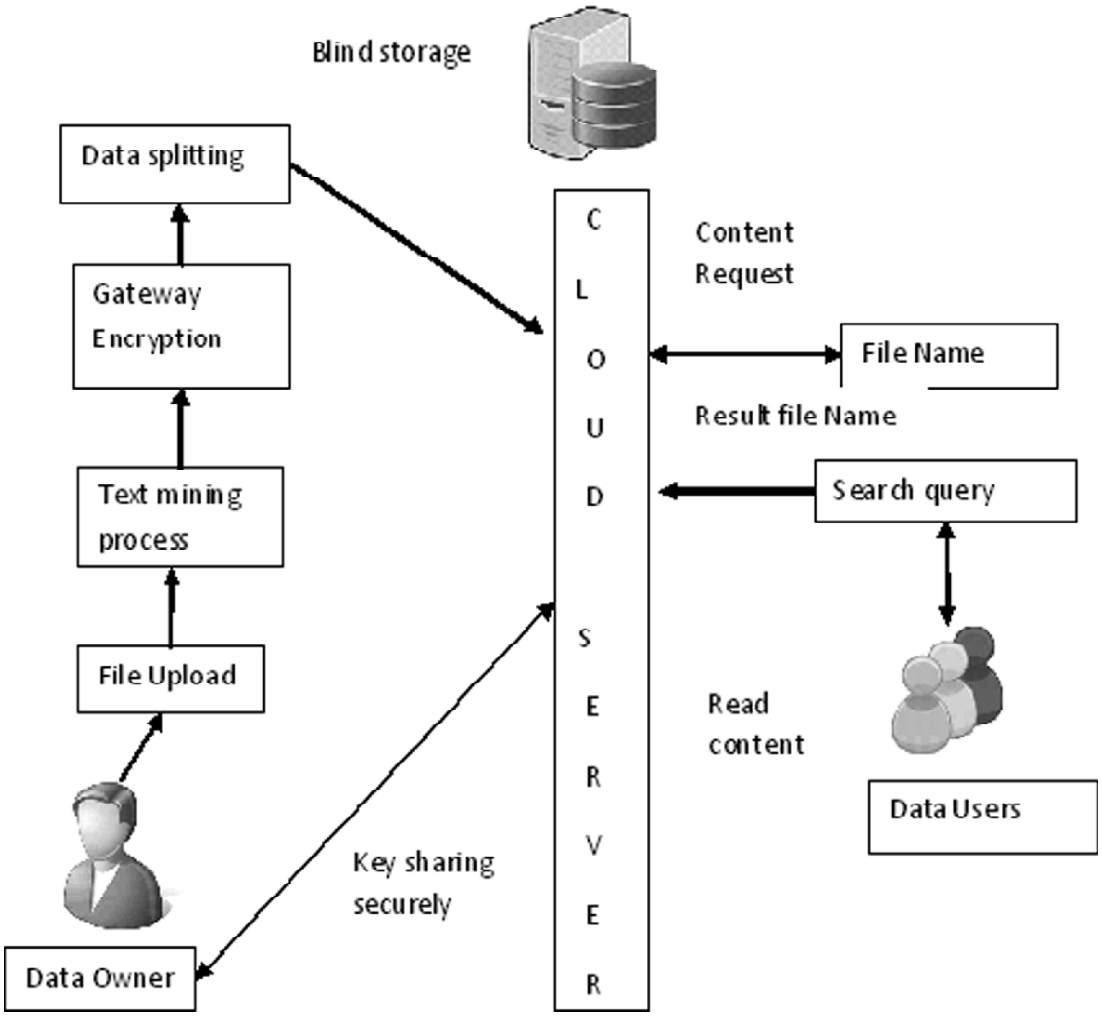
Though cloud server provides good range of user service, it has a interest in knowing what is actually present inside the data uploaded. So it leads to threat for the data shared in cloud server.

Based on the content stored in the cloud server, W.Sun et al proposed two different threat models. They are (i) known ciphertext model, (ii) known background model [2].The main difference between these models are that, in known background model more information are stored in the cloud server compared to known ciphertext model.

But every time we cannot be sure of how much data will be stored in cloud server. So in our paper we use a method known as blind storage. Blind storage prevents the server from gaining knowledge about what kind of data is uploaded or how and where it is stored by splitting the files and randomly ordering it in server.

3. SYSTEM MODEL AND SECURITY REQUIREMENTS

3.1. System Model



The system model shows the multi-keyword ranked search that consists of following entities. They are data owner, data user, and cloud server.

Data owner is the person who actually owns and uploads the files. When a data owner wishes to share a file with other users, the owner uploads it to the cloud server. But before uploading is done, the file is encrypted using a asymmetric encryption scheme and further the encrypted file is split into random number of pieces based on the size of the file. The information about the these split ups are stored in a table in order to rebuild the file when needed.

Cloud server is the storage location, where the data uploaded by the data owner is stored. Actually the cloud server does not have any knowledge about the file being uploaded since it is encrypted, split and stored in random order in the server. This helps in securing the file uploaded by the data owner and also privacy concerns are also satisfied.

The data user is the one who actually tries to access the file uploaded by the data owner. When a user needs a specific file then the search over encrypted data is performed by the user using the interested keywords. As a result, a list of most relevant files are provided to the user which are present in the cloud server from which user can choose the specific file that is needed. Finally, in order to rebuild, decrypt and access the file the user will need an access key.

3.1.1. Working of System Model

Initially, the data owner chooses the file that has to be shared with search users. Once the file is chosen, the different keywords present in the file are extracted and stored in table along with the synonyms of those keywords which can be used at the time of retrieval. After extracting the keywords, in order to improve its security the file is encrypted using an asymmetric algorithm. Further, to prevent the server from knowing about the file it is split into number of fixed size blocks. The blocks are then numbered and details about every file and its split ups are stored in an index table to rebuild the file when needed. These are blocks are finally stored in cloud server in a random order so as to prevent server from accessing it. Once the file is uploaded, it can be searched and accessed by the user. In order to access it, the user performs a search that lists a set of most relevant files from which the user can choose and access the needed file, provided that user has the access key to decrypt and rebuild the file.

3.2. Security Requirements

The multi-keyword ranked search system comprises the use of cloud server in order to store the files being uploaded by the data owners. Though the cloud server simply provides a place for storing files, it always has the curiosity in knowing what is stored in it. So as to prevent server from attaining knowledge about the server we perform various security related operations to the file before and at the time of uploading to the cloud server. The different steps involved in securing the file are as follows.

- The first step for security is encrypting of the file that is being uploaded with the use of RSA algorithm which is an asymmetric algorithm.
- The next step to make the file more securely store in cloud server is, splitting up of the file into number of fixed size blocks.
- The final step is to store these number of blocks in the cloud server in a random order so that it becomes completely impossible for anyone to rebuild the file.

4. DESIGN GOALS

To enable an efficient multi-keyword ranked search over encrypted data stored in cloud server through the concept of blind storage, various design strategies are followed. They are,

- *Multi-keyword Search:* To fulfill the needs in practical users and to provide better performance to the user, the MKRS (Multi-Keyword Ranked Search) should support multi-keyword search on encrypted data that are stored in a cloud server and should also provide relevance based ranking result.
- *Search Efficiency:* When the search is performed over a sizeable document[11] it practically should perform with best efficiency[12].This is achieved by extracting and storing the keywords in the separate index table that point to the file. So when a user performs a search the keywords are compared with the information in this table and results are produced.
- *Ranked Search:* While performing search with multiple keyword, the document or file that is having maximum number of occurrences of the search keyword is given more preference[13],[14].

Similarly, D.X.Song et al proposed different practical techniques for performing search over encrypted data[9].

5. TECHNIQUES USED

5.1. Relevance Ranking

In searchable asymmetric encryption methods, since a large number of documents are processed, the results of the search has to be retrieved in a most relevant order based on the keywords search. Ranking is a general way to assess the relevancy of the file. There are number of relevance ranking techniques available, from which we adopt TF-IDF weighting suggested in [15][16]. In TF-IDF technique, $tf_{t,f}$ refers to the number of occurrences of the term t in the file f and df_t refers to the number of documents that holds the term t. N denotes the total number of documents in the database. Then we calculate idf_t=log, where idf_t denotes the inverse document frequency. Finally, $tf_{t,f}$*idf_t is calculated to find the weighting of the term t in the file f.

5.2. Blind Storage

In order to overcome different attacks and security threats, cloud has to provide high security for the data being uploaded by the data owners. The technique of blind storage[3] enables the owner to securely store the data and such data is made visible only to the data owners. The blind storage concept enables to store data in remote servers such that the server does not gain any knowledge about the actual content of the file uploaded. This concept supports uploading new files, modifying or deleting them. The server can know only the name of the file being uploaded. So blind storage leaks only very little information to the cloud server. Through such a technique the data owner can prevent the users from knowing about the content of their file. Here each files are split into number of fixed size blocks. These block information are indexed in a table for rebuilding of the file.

5.3. Public Key Encryption using RSA Algorithm

In this paper RSA algorithm is implemented which is a public key encryption method. This algorithm converts the actual data into unreadable form so that intruders cannot gain any information even if they break into the system. It generally generates two keys, that are public and private keys. Where public key is used for encryption and private key is used for decryption. There are three phases in RSA algorithm, first phase is key generation which is used in the process of encryption and decryption. Second phase is encryption, where the actual data is converted into cipher text. Decryption is the third and final phase of RSA, where the encrypted data is converted back to actual data.

5.4. Construction of Blind Storage

The blind storage technique is considered to be an secured storage technique for string data in remote cloud server [17]. The construction of blind storage includes different steps in it. The first step starts with analyzing

the file uploaded, where the name of the file is noted in a table. The next step is to split the encrypted file into number of fixed size blocks. Then the divided blocks are randomly numbered. The numbers of each blocks and order of these blocks information are stored in the same table next to the name of the file. This numbering and storing of information will be used at the time of rebuilding the file. Finally the blocks of the file are stored in the remote cloud server in a random order.

6. RELATED WORK

Searching over encrypted data is a propitious technique that enables searchable encryption in cloud data. Mainly there are two types of searchable encryption technique. They are, searchable asymmetric encryption and searchable symmetric encryption. The concept of searchable asymmetric encryption is proposed by Boneh et al. [18] that supports single keyword search while searching over encrypted cloud data. To support subset, range and conjunctive search queries over encrypted data, the search was extended in [19]. The above search technique should match all keyword at same time but it is unable to return result in specific order. Mask matrix technique is adopted in ranked search scheme by Liu et al [12] to achieve cost effectiveness by using fully homomorphic encryption. Yu et al [15] proposed multi-keyword retrieval scheme to return top-k relevant documents. The technique of attribute based encryption is used in [20],[8] to get search authority in searchable asymmetric encryption.

Though searchable public key encryption have good search functionalities since it includes more asymmetric cryptographic operation, it is not efficient. Song et al [9] first proposed searchable symmetric encryption technique which supports single keyword and builds searchable encrypted index in symmetric way. Further security of searchable symmetric encryption was improved in [21] by Curtmola et al. The introduction to a basic approach of using keyword related index to enable quick search of documents has led to various subsequent works, like [22],[7] and [11]. In order to improve user experience in searching in [13] and [14] some proposals are introduced to enable ranked results rather than producing undifferentiated results by proposing relevant scoring in searchable encryption. In order to improve the user experience, performing fuzzy keyword search on the encrypted data is developed in [3] and [23]. The technique of k-nearest neighbors (kNN) in searchable encryption is adopted in privacy preserving multi keyword search proposed by Cao et al. Though this proposal achieve rich functionalities, it incurs a large computation in cloud server. By extending the work of Cash et al. [22] in [11] real-world data sets are implemented to achieve multi-keyword search in large datasets. But ranked result of [22] is excluded in [11].

7. CONCLUSION

In this paper, multi-keyword searching over encrypted data is proposed to enable searching and sharing of files in a secured manner. The TF-IDF technique is adopted for ranking the search results to enable the similarity based text retrieval, which considers the number of occurrences of the keywords in each document and gives most relevant results. To make the process more secured, asymmetric encryption is implemented using RSA algorithm and the concept of blind storage is also adopted, which prevents the cloud server from gaining knowledge about the data being uploaded.

REFERENCE

[1] H. Li, Y. Dai, L. Tian, and H. Yang, ‘‘Identity-based authentication for cloud computing,’’ in Cloud Computing. Berlin, Germany: Springer-Verlag, 2009, pp.157-166.

[2] W. Sun, et al., ‘‘Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking,’’ in Proc. 8th ACM SIGSAC Symp. Inf., Comput. Commun. Secur., 2013, pp. 71–82.

[3] B.Wang, S.Yu, W.Lou, and Y.T.Hou, ‘‘Privacy-preserving multi-keyword fuzzy search over encrypted data in the cloud,’’ in Proc. IEEE INFOCOM, Apr./May 2014, pp. 2112–2120.

[4] E. Stefanov, C. Papamanthou, and E. Shi, ''Practical dynamic searchable encryption with small leakage,'' in Proc. NDSS, Feb. 2014.

[5] Y. Yang, H. Li, W. Liu, H. Yang, and M. Wen, ''Secure dynamic searchable symmetric encryption with constant document update cost,'' in Proc. GLOBECOM, Anaheim, CA, USA, 2014

[6] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, ''Privacy-preserving multikeyword ranked search over encrypted cloud data,'' IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 1, pp. 222–233, Jan. 2014

[7] M. Naveed, M. Prabhakaran, and C. A. Gunter, ''Dynamic searchable encryption via blind storage,'' in Proc. IEEE Symp. Secur. Privacy, May 2014, pp. 639–654.

[8] Q. Zheng, S. Xu, and G. Ateniese, ''VABKS: Veriûable attribute based keyword search over outsourced encrypted data,'' in Proc. IEEE INFOCOM, Apr. 2014, pp. 522–530.

[9] D. X. Song, D. Wagner, and A. Perrig, ''Practical techniques for searches on encrypted data,'' in Proc. IEEE Symp. Secur. Privacy, May 2000, pp. 44–55.

[10] H. Pang, J. Shen, and R. Krishnan, ''Privacy-preserving similarity-based text retrieval,'' ACM Trans. Internet Technol., vol. 10, no. 1, p. 4, 2010.

[11] D. Cash et al., ''Dynamic searchable encryption in very-large databases: Data structures and implementation,'' in Proc. NDSS, Feb. 2014.

[12] 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.

[13] C. Wang, N. Cao, K. Ren, and W. 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.

[14] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, ''Secure ranked keyword search over encrypted cloud data,'' in Proc. IEEE 30th Int. Conf. Distrib. Comput. Syst. (ICDCS), Jun. 2010, pp. 253–262.

[15] J. Yu, P. Lu, Y. Zhu, G. Xue, and M. Li, ''Toward secure multi keyword top-kretrievaloverencryptedclouddata,''IEEETrans.DependableSecure Comput., vol. 10, no. 4, pp. 239–250, Jul./Aug. 2013.

[16] Menaka S, Radha N ''Text Classification using Keyword Extraction Technique,''International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013.

[17] D. Kavitha, S.Hemavathy, ''A Survey on Cloud Computing Security Issues And Multi-Keyword Ranked Data Search Efficiency in Blind Storage,'' International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 9, September 2015.

[18] D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, ''Public key encryption with keyword search,'' in Proc. EUROCRYPT, 2004, pp. 506–522.

[19] D. Boneh and B. Waters, ''Conjunctive, subset, and range queries on encrypted data,'' in Proc. TCC, 2007, pp. 535–554.

[20] W. Sun, S. Yu, W. Lou, Y. T. Hou, and H. Li, ''Protecting your right: Attribute-based keyword search with ûne-grained owner-enforced search authorization in the cloud,'' in Proc. IEEE INFOCOM, Apr./May 2014, pp. 226–234.

[21] R.Curtmola, J.Garay, S.Kamara, andR.Ostrovsky, ''Searchable symmetric encryption: Improved deûnitions and efûcient constructions,'' in Proc. 13th ACM Conf. Comput. Commun. Secur., 2006, pp. 79–88.

[22] D. Cash, S. Jarecki, C. Jutla, H. Krawczyk, M.-C. Roºu, and M. Steiner, ''Highly-scalable searchable symmetric encryption with support for Boolean queries,'' in Proc. CRYPTO, 2013, pp. 353–373.

[23] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, ''Fuzzy keyword search over encrypted data in cloud computing,'' in Proc. IEEEINFOCOM, Mar. 2010, pp. 1–5.