International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 3

A Competent Multi-Keyword Exploration Scheme over Encrypted Data in Cloud

Saranya K, Manohari J, Haripriya H[#]
[#]Assistant Professor
B. E, Computer Science and Engineering, Prince Shri Venkateshwara Padmavathy Engineering College, Chennai, Tamil Nadu, India

ABSTRACT

Cloud computing is an emerging technology in the present scenario which enhances enormous data storage and its processing. Various issues are present in maintaining the cloud environment. Some among them include data privacy and data confidentiality faced by the data owners while storing their private data onto the cloud. The domain of Data mining mainly helps in efficient retrieval of the content stored in the public space using various retrieval methods. Generally, in cloud, the data privacy is provided by outsourcing the files in encrypted form before being uploaded. This creates more demand in server side data utilization i.e., searching on encrypted files from the cloud. As of now the retrieval is made productive through keyword based search. Most recent works focus on the single keyword provided as queries in the search process. In the keyword ranked search, it had been difficult to store when the number of keywords are more. In the work carried out below, an efficient and reliable multi-keyword search method called K-Gram search is developed. Using K-gram it is possible to retrieve more than one file i.e., related files to the keyword searched, productively relative to misspelled keywords too. In addition the user behavior is too monitored from the start for enhancing keyword search and it also ensures proper confidentiality via access confirmation through secret keys. The system thus tends to provide more chances for tackling data privacy and confidentiality issues. It can also achieve sub linear search time thereby increasing the query processing and accuracy in keyword search.

Keywords: *Cloud computing; Data mining; Multi-keyword search; K gram; Data privacy; Sub-linear search time*

INTRODUCTION

As there is an enormous growth in the technology in today's world, simultaneously the growth of data is too large. Traditionally data owners had their own storage. As a result of which they were able to access their data with more privacy but they were unable to access them from any location. Also the data rate is huge nowadays, they found it difficult to store them in their own directory and access them. So they moved on to cloud storage.

Cloud storage provide more easier access from any location(location independent resource pooling) and had a large amount of space for storage since it is highly scalable (resource elasticity). Also cloud helps to achieve on-demand quality data services and usage-based pricing. It is reliable, reduces the cost of managing data and storage spending. Even though cloud was able to tackle the huge data, it too had issues regarding data privacy and confidentiality like information leakage. This is because any user availing the public key was able to access the files being stored in the cloud. In order to ensure that the information is being preserved, file encryption was carried out using certain cryptography techniques. This depicts the process of encrypting the file before they are stored in the cloud. By doing so more security was ensured for the files being stored in the cloud. File encryption process in addition resulted in serious issue in data retrieval from cloud storage.

Even though efficient data mining procedures are available it was difficult to perform the retrieval efficiently. In order to increase the performance and enhance the retrieval, keyword based search was used. Keyword based search using a single keyword enhanced the accuracy in search but there was only possibility of retrieval using the exact keyword. The user had a single possibility of searching the file using the keyword as that of the exact filename. Thus the existing system won't be able to provide any search result using misspelled keywords. The exact keyword search provided more accuracy based on Boolean keyword matching system and ought to provide data privacy since the user was able to retrieve the file only when he is aware of the exact file name. Later multi-keyword based search came to existence where user can search the file using various possibilities of keywords. Even this approach dealt with various issues in arranging the keywords and accessing them. For which they used vector space model, tree based indexes.

In the proposed work, search is enhanced by generating a set of keywords based on k-gram algorithm. The main advantage of doing so is the user can even carry out the search by using misspelled keywords too if he is unaware of the exact filename. This enables the easier and efficient access of the appropriate file by the user when there is a need. The query is processed efficiently due to keyword based search for enabling effective retrieval of the file.

The paper covers the information about the related work in section II followed by the proposed system in section III. In section IV, the system architecture is explained briefly followed by various modules and its description in section V. In section VI, the algorithms used mainly for encryption and search based techniques are explained. Finally we have the conclusion followed by a set of references.

RELATED WORK

Traditional searchable encryption[1] [2] has been widely studied as a cryptographic feature, mainly focusing on security and efficiency improvements. In the existing system[1], Song et al. was the first one to propose the traditional single keyword searchable encryption scheme. To search over the encrypted files with a sequential scan, the first practical symmetric searchable encryption (SSE) scheme employs a 2-layered encryption structure. In this searchable scheme each word is encrypted independently and the user has to go through the entire document to search a particular keyword. Unless the appropriate trapdoors are given via secret key the document contents are hidden to the server. But its drawback is only fixed length output is supported and the search time was linear to the size of data collection.

Ranked search enables quick search of the most relevant data. Wang et al.[6] and Cao et al.[2] did their research on secure ranking on single and multi-keyword respectively. In reference [2], the 'first privacy preserving symmetric multi-keyword ranked search' scheme was proposed against two threat model (which is called MRSE), in which the documents and queries was represented as vectors of dictionary size following vector model and secure inner product to realize the high efficiency of search. By using 'coordinate matching' the documents were ranked according to the number of matched keywords. But it failed to provide more accuracy and the search efficiency of the scheme was linear with the cardinality of the collected document.

Abundant researches were been carried out and various schemes were proposed to achieve quite more search functionality. They include similarity search [5] [8] [10] and multi-keyword ranked search [2] [9]. In the existing work[5] supporting Boolean keyword search, an important issue of fuzzy keyword search was proposed by Li et al, to address the spelling errors and was improved in later work. It aimed at tolerance of both minor typos and format inconsistencies in the user search input.

One among the search functionality is multi-keyword Boolean search which allows the retrieval of only the suitable documents related to the multiple query keywords. It includes 'conjunctive keyword search' scheme which has the possibility of returning the documents containing all the query keywords and the 'disjunctive keyword search' scheme which has the possibility of returning the documents containing a subset of the query keywords. It also possess both the conjunctive and disjunctive schemes proposed together as 'predicate search scheme'. Somehow all the multi-keyword search schemes provide search results based on the keywords existing but fail to provide acceptable keyword ranking functionality.

The work carried out in reference[3] presents a secure multi-keyword ranked search scheme employing VSM (vector space model) and the widely used TF-IDF mode, which comprises of the dynamic update

operations like insertion and deletion of the documents.

In reference [9], Sun et al proposed a secure multi-keyword search technique with greater efficiency by adopting MDB tree supporting similarity based ranking. A searchable index tree was constructed based on vector space model and adopted cosine measure (compare source and query) together with (Term Frequency)*(Inverse Document Frequency) to provide accurate ranking results. It achieves better-than-linear search efficiency but results in precision loss. However most of the existing searchable encryption schemes support only exact keyword search, thereby affecting data usability and user's experience. Later a semantic keyword search scheme based on stemming algorithm was proposed which retrieves documents having semantically close keywords related to the search query.

PROPOSED SYSTEM

Numerous works have been proposed in order to carry out the search process like single keyword, multi-keyword based search attaining more importance due to their realistic implementation in file retrieval. Even though they gained more significance, the relative computational overhead they caused led to performance degradation.

The process of searchable file encryption enables the client to collect the encrypted file from cloud and execute the keyword search above cipher text domain. File encryption is performed by the use of various cryptography techniques. In the proposed system will help to develop a secure and efficient search method called K-gram search, which is used over encrypted data and supports multi-keyword search for processing a range of documents stored in cloud. K-gram search technique searches all possibilities of keywords including misspelled words too in order to retrieve the appropriate file.

The working of the proposed system is elucidated as follows.

- Validation of the login credentials.
- The files to be outsourced is uploaded to the cloud storage after encrypting it using an efficient cryptographic technique called 'Blow-fish' algorithm.
- Automatic keyword set generation for the encrypted file is done comprising of all possibilities of misspelled words.
- User searches the appropriate file using multi-keyword search approach
- Search keyword is encrypted and is compared with collection of key set generated for the original encrypted file.
- If both the keywords match, the particular file is searched from the stored file list and successfully the files of attention are been retrieved from the cloud server through sub-linear search time.
- Before the retrieval of the particular file, access confirmation will be made with the data owner in order to ensure data privacy being achieved.

This improves the chances to attain data privacy, and provides the search result in sub-linear search.

SYSTEM ARCHITECTURE

The system architecture for our proposed system shown in fig 2 is as follows,

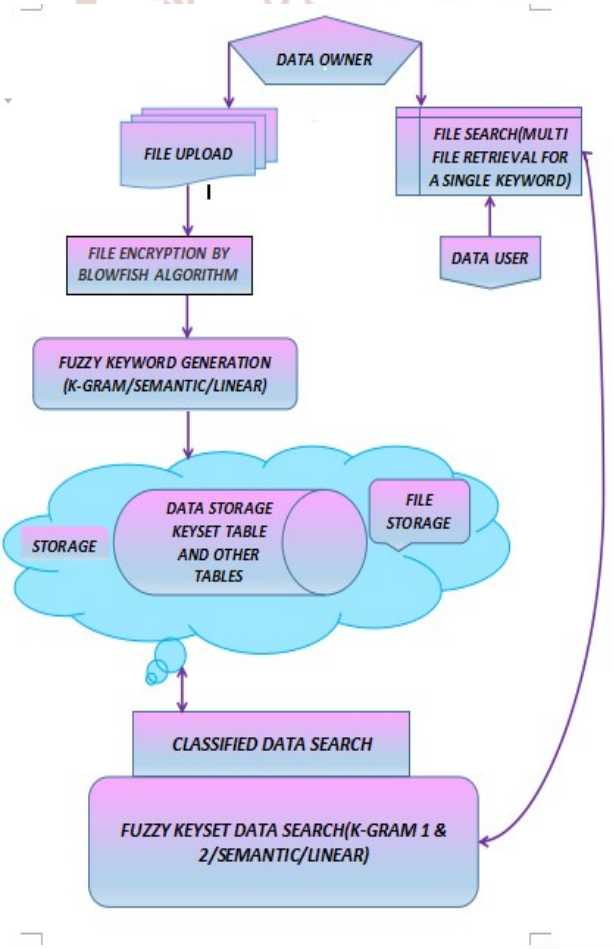

Fig 1: SYSTEM ARCHITECTURE

The system architecture shown in fig 1 comprises of processing of the outsourcing file in both admin and user side. The admin or the data owner possess major functionalities such as file uploading and key set generation while the user searches for a particular file through multi-keyword based search and if a match

occurs between the search keyword and those available in the key data set, then the user accessing the cloud server is authenticated via access confirmation performed by the data owner. The user behavior is continuously monitored by the admin throughout the system to enhance the search process according to the user's interest, thus resulting in efficient file search and its retrieval.

A. ADMIN SIDE

The process starts from the admin. Initially admin is the one who logs in to the system. The prime activity done by the admin is outsourcing his data on to the cloud storage. After packing up their data from local storage on to cloud, the data owners felt less difficulty in handling large amount of data they had. In addition the data is to be outsourced in the cloud server, the file need to be encrypted in order to ensure its privacy and confidentiality, thereby avoiding data leakage. Once the files are uploaded, the admin triggers the generation of key data sets implemented through K-gram techniques and in addition can also add some additional keywords to the data set in order to enable more efficient and quick retrieval. The generated key sets are clustered using support vector machine and are stored in the cloud. The admin side processing of the file to be outsourced ends after the keyword generation is completed.

B. USER SIDE

The second part of the process includes the user side processing of the encrypted file. The user too has certain login credentials (user name,password,mail id) in order to enter into the system. The main activity for the user is to retrieve a particular file required using keyword-based search techniques. In the proposed system, k-gram based search is carried out. The keyword to be searched is encrypted, and is compared with the collection of key data set generated for the original encrypted file. If a match is found, the particular file is searched from the stored file list and it is successfully retrieved from the cloud server through sub-linear search time. The advantage in proposed system is that the files can be retrieved even using misspelled words too. This may lead to an issue regarding preserving data privacy. In order to achieve this, the access confirmation is made regarding the rights to be granted to the user is provided from the admin side. The process of access confirmation is accompanied by sharing the secret private key to the appropriate user through the user's registered mail Id. With the use of the private key shared from admin, the user decrypts the file necessary from the files of attention being listed and finally downloads it from the cloud.

V. MODULES DESCRIPTION

The proposed system is explained in detail via the following set of modules,

- User registration
- Upload File
- Search
- Mail alert process
- File download process

1. USER REGISTRATION

In this module, the login process is elucidated. In our system, registration process is carried on both the admin and user side. Their login credentials are stored in the storage for validating them later when they enter the system. Usually, the user account name and appropriate password of that account are sufficient to do the authentication and login process is validated.

2. UPLOAD FILE

In this module, the following activities are done.

- Load the input document from the owner's storage.
- Read the document file being uploaded and encrypt it using Blow fish algorithm.
- Pre-process the uploaded file and trigger the automatic generation of the key data set for the encrypted file.

3. SEARCH

3.1. Frequent Search

In this search, we get the non-stop words as input and calculate the count of words and find the repeated occurrence of each and every word from the non-stop words.

3.2. Similarity Search

In this search, from the maximum frequent words we find the weightage of each and every word and from them the similarity between the words is found, based on which the words are grouped into clusters.

3.3. Linear search

In this search, we are going to perform search regarding the exact keywords. Each cluster has n number of similar words as keywords and using which we find the file for that cluster with the help of lexical analysis tool.

4. MAIL ALERT PROCESS

As soon as the file is retrieved using keyword based search and after the user is authenticated by the data owner, the secret key is shared to the corresponding authorized user to their respective mail Id (specified by the user during the registration process) to carry out the decryption. The purpose of using secret key is to preserve the privacy for the outsourced data by the data owner. In the constructed system, anyone with the public key can write the data stored on the cloud server but only authorized users with the private key can carry out the search process.

5. FILE DOWNLOAD PROCESS

The downloading process carried by the user must be privacy preserving one, which is achieved by providing proper access confirmation. It is a two-step process as mentioned below,

i. Getting the secret key from the corresponding user email id.
ii. Apply the secret key to the encrypted file retrieved from the storage for performing decryption using Blow-fish algorithm and download the encrypted file for viewing it in the user's own data store.

VI. ALGORITHMS USED

In the proposed system, the main issue that is dealt with is data privacy (by the use of misspelled keywords too for enhancing file retrieval in the later part, which is solved by providing access confirmation), provided through encryption of the file to be outsourced onto the cloud. The encryption and decryption process is achieved through an efficient cryptographic technique i.e., Blow-fish algorithm.

- ***Blow-fish*** is a symmetric (uses the same key for both encryption and decryption process) block (code is divided into fixed 64-bits length blocks) cipher algorithm, designed by Bruce Schneider in the year 1993. Blow-fish can be used as a drop-in replacement for DES (Data Encryption Standard) algorithm. The Blow-fish algorithm consists of 2 parts; ***key expansion*** (input key is converted to several sub-key arrays, a total of 4168 bytes) and ***data encryption***.

PSEUDOCODE

The Blow-fish algorithm follows the steps below,

1. Divide the input X into two blocks,XL and XR.
2. Consider for a range of available blocks i= 1 to16,
   2.1. XL=XL XOR P1 (first 32 bits of plain text)
   2.2 XR= Function ,F(XL) XOR XR
   2.3 Swap XL,XR
   2.4 Increment i
3. Swap XL,XR.
4. XL=XL XOR P18 (last 32 bits of plain text).
5. Combine XL and XR back into X.

ADVANTAGES

Blow-fish algorithm is an efficient cryptographic technique since it has the following features when compared to other cryptographic techniques.

- Fast, as data is processed on 32-bit microprocessors at a rate of 26 clock cycles per byte.
- Compact, since it requires <5KB of memory space for processing.
- Simple, as it carries out simple operations using 32 bit operands.
- Secure, because it possess variable key length, in the range of 32-448 bits; default is 128 bits.

The search technique being implemented to enhance the file retrieval process is K-gram based search. It is been proposed in order to retrieve the files productively using All possibilities of misspelled keywords too.

- K-gram search technique is used to generate and search keywords comprising of all the possibilities of misspelled keys relative to the file name.

K-GRAM DESCRIPTION

K-grams are used to limit the terms for which the edit distance need to be computed to the query term by accessing k-gram index. K-gram index is used to retrieve the terms that have many grams in common with the query. The retrieval process is that of a single scan through the postings for the k-grams in the query string q.

In the proposed system, we mainly make use of k-gram 1 and k-gram 2 search technique in addition to linear search (exact keyword) method in order to address the retrieval difficulties.

K-gram 1 search is used to generate all possibilities of keywords with only one misspelled letter in it as shown in fig 2.1. For instance, a given filename is analyzed and all chances of misspelling one letter in it is generated and stored in the cloud server for later retrieval.

Similarly, K-gram 2 is used to generate all possibilities of keywords with combination of any two letters being misspelled in it as shown in fig 2.2. Similarly a given filename is analyzed and all chances of misspelling two letters in it is generated and stored in the cloud server for later retrieval.

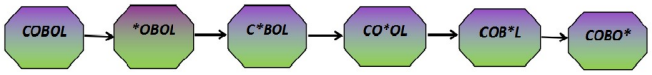

Fig: 2.1: KEYWORDS GENERATED USING K-GRAM 1

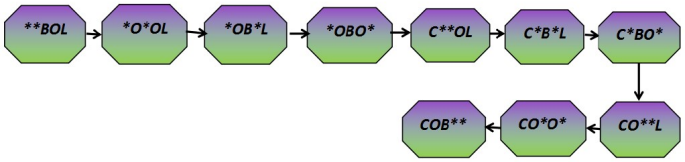

Fig 2.2: KEYWORDS GENERATED USING K-GRAM 2

CONCLUSION

The keyword-based search is a widely used data retriever in many storage and information retrieval systems but its traditional processing methods cannot be directly applied for processing the encrypted data. In the above work, in order to improve the search efficiency, multi-keyword based search scheme using K-gram technique is proposed, where there is no need to give exact filename to download the file. This improves the ability of handling the privacy breaches, accuracy and speeds up the query processing. The recommended system thus provides the efficient search result for the input search queries in sub-linear time.

FUTURE ENHANCEMENTS

For the future work, further more analysis can be done to come up with an advanced and more efficient search procedure to retrieve the files with enhanced accuracy and data privacy being achieved.

REFERENCES

1. *X. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data", in Security and Privacy, 2000. SP 2000.Proceedings. 2000 IEEE Symposium on , May 2000, pp. 44–55.*
2. N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data", IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 1, pp. 222–233, 2014.
3. Z. Xia, X. Wang, X. Sun, and Q. Wang, "A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data", IEEE Transactions on Parallel and Distributed Systems, vol. 27, no. 2, pp. 340–352, 2016.
4. W. M. Liu, L. Wang, P. Cheng, K. Ren, S. Zhu, and M. Debbabi, "Pptp: Privacy-preserving traffic padding in web-based applications", IEEE Transactions on Dependable and Secure Computing, vol. 11, no. 6, Nov 2014.
5. J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy keyword search over encrypted data in cloud computing", in INFOCOM, 2010 Proceedings IEEE, 2010, pp. 1–5.
6. *Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure ranked keyword search over encrypted cloud data", in Distributed Computing Systems (ICDCS), IEEE 30th International Conference on,June 2010, pp. 253–262.*
7. H. Li, Y. Yang, T. H. Luan, X. Liang, L. Zhou, and X. Shen, "Enabling fine-grained multi-keyword search supporting classified sub-dictionaries over encrypted cloud data", IEEE Transactions on Dependable and Secure Computing, vol. 13, no. 3, pp. 312–325, May- June 2016.
8. B. Wang, S. Yu, W. Lou, and Y. T. Hou, "Privacy-preserving multi-keyword fuzzy search over encrypted data in the cloud", in INFOCOM, 2014 Proceedings IEEE, 2014, pp. 2112–2120.
9. W. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li, "Privacy preserving multi-keyword text search in the cloud supporting similarity-based ranking", in Proceedings of the 8th ACM SIGSAC Symposium on Information, ser. ASIA CCS '13. ACM, 2013, pp. 71–82.
10. C. Wang, K. Ren, S. Yu, and K. M. R. Urs, "Achieving usable and privacy assured similarity search over outsourced cloud data", in INFOCOM, 2012 Proceedings IEEE, 2012, pp. 451–459.