# Position Based Sentence Search for Encrypted Unstructured Data in Cloud Environment

Muhammad Zaman Fakhar[1] and Dr. Shoab Ahmad Khan[2] and Madiha Waris[3]
[1,2,3]National University of Sciences and Technology, Islamabad, Pakistan
[1]zamanfakhar11@ce.ceme.edu.pk, [2]shoabak@ceme.nust.edu.pk, [3]madihawaris11@ce.ceme.edu.pk

***Abstract*— Over recent years cloud computing has attained foremost commercial success. Cloud computing minimizes resource wastage risk by reducing the entrance barrier for cloud service providers. By extensive usage of cloud services unstructured data volume is increasing over it. Therefore security considerations to save data from hackers are also becoming a necessary aspect. To avoid illicit use of data placed on the cloud different encryption techniques and standards have been proposed by the researchers. Data searching becomes a challenging task by adopting the existing techniques. In this paper a new technique Position Based Sentence Search (PBSS) has been proposed. This technique facilitates user with sentence searching for unstructured data from the original document in cloud environment. PBSS provides an efficient ranked sentence search by means of preprocessed indexes. There is no need of decrypting documents during the search process as in existing systems. Decryption is done on the retrieval of the documents only. Nobody has the prior knowledge about contents placed on the cloud therefore PBSS achieves sentence privacy.**

***Index Terms*—PBSS, indexing unstructured data, searchable encryption, searchable cloud environment, ranking based encrypted search, position based sentence search.**

## I. Introduction

With the commencement of cloud computing composite data management systems from local sites are transformed to viable public cloud. Data owners are encouraged to outsource the data management systems to public cloud to achieve flexible and commercial benefits [1]. Cloud computing is all about transferring services, applications and data. Also attaining commercial assistances, location transparency, and centralized facilitation are the significant resources in cloud computing. On a shared collective platform like cloud data can be retrieved easily with less revenue and improved assistance [2].

Cloud storage has the capability to save a bulk of data for a large number of users. This minimizes the storage capacity problem. To provide different competences multiple isolated applications and services are disseminated over the internet in cloud environment [3]. When sensitive data storage is done on the cloud, existence of large number of users can cause cloud security to be affected. Thus for achieving data privacy complex data has to be outsourced on the cloud after encryption. Therefore to hide data from hackers and malicious attackers a protected system is needed.

Searchable encryption is a technique by which the outsourced data placed on cloud can be kept private. Searchable encryption will let this data to be difficult to hack when searched. With searchable encryption techniques encrypted data is placed on the cloud server on which search can be performed. Processing of encrypted data placed on cloud server is done without decrypting it. The encrypted data is placed on the cloud in the form of code words which are difficult to hack by untrusted user or hackers [4]. The encrypted data will be accessible by authenticated and authorized users only. These users will be able to perform search on this data and retrieve desired results.

This paper presents a search technique Position Based Sentence Search (PBSS) for unstructured data in cloud environment. Using this technique on the basis of sentence being searched the user is able to retrieve the corresponding document which contains that sentence. In PBSS during search a sentence match is done by finding the positions of the specific keywords searched in the document index. The close positions indicate the sentence match. PBSS retrieves documents with high frequency of occurrences and minimum standard deviation of positions on the top. Unstructured data is indexed at first by collecting distinct keywords words from the original document with their specific positions. After index generation the data is placed on the cloud server in the form of codewords. The codewords are generated using hash algorithms which enhance the security and privacy of the document index.

To generate secure searchable encrypted index of unstructured data documents bloom filters are used. Bloom filter is a data structure which is used for fast set membership test with the possible false positives. A Bloom filter is stored as an array of bits. All the bits are initialized to ‗0‘. After the addition of an element different hash functions are performed on that element. The input to each hash function is the element to be added in the array and the output from each hash function is the index into the array which is different for each hash. After the calculation of the hash algorithms each bit in the filter at the indexes which are specified by the hash outputs is set to

1. For checking a specific element in the bit array it is tested that if any of the bit is set to ‗0‘. If it is set to ‗0‘ it means that this particular element is not present in the bit array [5].
As an example in the Fig.1 the bloom filter using three hash functions as shown.

Proc. Of the International Conference on Cloud Computing and eGovernance 2013 – ICCCEG 2013. Edited by Manikandan Ayappan. © Organizers of ICCCEG 2013 [iccceg@iccceg.org]. Published by Association of Scientists, Developers and Faculties, HQ, India. ISBN : 978-81-925233-2-3 || DOI : 10.ASDFOI/925233.002

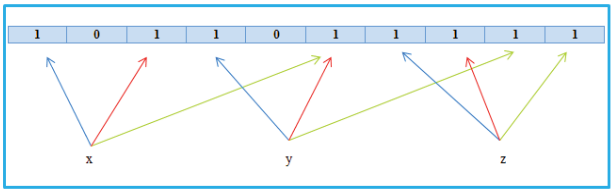

Fig.1. Bloom filter using three hash functions

In Fig.1 three words x, y and z have been added as members to the array in such a way that three hash functions have been applied on each of the word.
The summarized contributions are as follows:

- Distinct words from each document are extracted and indexed.
- The retrieved searched documents are in the ranked order based on term frequency and position standard deviation.
- Bloom filters are used for distinct words indexing to ensuring security.
- The paper introduces a new technique named PBSS which facilitates user to have position based sentence searching with punctuation marks.
- Overhead of decryption before searches is reduced.
- The content identity is not revealed to any of data user or cloud server.
- PBSS provides efficient and accurate sentence search as compared to the keyword based search.

The remaining paper is structured in such a way that section II states the problem statement. Section III describes the literature review. Proposed framework has been shown in section IV. In Section V the proposed technique along with its working scenario has been described. Section VI comprises conclusion and future work.

## II. Problem Statement

For sentence based search the encrypted documents located on the cloud server are difficult to search proficiently. From these documents search can only be performed after downloading and decryption of these documents. The motive to implement PBSS technique is to retrieve documents in the descending order of their ranks from the whole set of encrypted documents placed on the cloud. The problem statement is as follows:

- How sentence based search can be achieved using positions of keywords in the documents?
- How to index unstructured data for sentence based searching?
- How the proposed technique achieves data privacy?

## III. Literature Review

Critical complex data when placed on the cloud server can undergo security complications. Therefore this data when searched by users should be present in encrypted form to vanish the hacking inference. In literature different techniques for data protection, indexing, and searching have been discussed. These techniques serve as a foundation towards the implementation of proposed technique for sentence based searching.

N. Cao et.al proposed a technique to enable searchable encryption having secured ranked search [6]. Accurate retrieval of results have been obtained using relevance ranking of search results. The technique has used a statistical measuring approach from the point of information retrieval to building secure index. Keyword privacy has been ensured providing no information leakage on cloud server. Experimental evaluation has been done showing improved efficiency as compared to the previous techniques. No implementation details have been shown in this study and only mathematical proofs have been given to validate the results. The technique is comprised of two phases which are setup phase and retrieval phase. In setup phase key generation and index generation is done followed by the retrieval phase. In retrieval phase trapdoor generation and index search is performed. This search is based on single keyword search only. Only theoretical details have been given. No analysis is performed on security and performance of keyword to be searched.

Curtmola et.al proposed improved search techniques on the basis of literature reviewed in searchable encryption domain [7]. New security definitions have been discussed by authors along with highlighting limitations in the existing literature for security definition. The authors presented such constructive definitions which are comparatively efficient w.r.t to existing searching techniques. Key generation, build index, trapdoor generation and index searching are basic steps which have been proposed to carry out search. The proposed technique has been constructed on the basis of combination of a lookup table and an array. A linked list is generated for saving the list of document identifiers in which the word is found. Th problem associated with the implementation of this technique is a need to update the array and the trapdoor whenever the document is added or removed.

Park et.al proposed a technique comprising two approaches which are efficiency and searching in cloud data center [8]. At first for efficiency and search two techniques of index search-I and II have been proposed. Secondly for these two techniques the analysis has been performed. The encrypted database has also been evaluated for efficiency. The basic steps involved in the proposed technique are SysPara, KeyGen, IndGen, DocEnc, TrapGen, Retrieval and Dec. The indexes produced as a result of this technique are not secure and the index contents may be deduced by hackers. Any third party malicious user can trace the common keywords from any of two documents.

Im-Yeong Lee and Sun-Ho Lee presented a technique based on re-encryption concept [9]. Searchable indexes have been generated in order to share data safely. In the first step key generation is performed. Then the re-encryption keys are generated. In the second step keyword search is done by generating trapdoor with secret key. Authenticated users can only decrypt the data. There is not technical and experimental proof for the validation for this technique. Only theoretical evaluation has been done.

N. Cao et.al proposed a searchable encryption technique for facilitating multiple keyword searches from the data placed on the cloud server [1]. Accurate retrieval of results is obtained by using relevance ranking concept. The coordinate matching approach has been used to achieve multi keyword based search. Similarity measure has been calculated by using inner product similarity concept. In order to ensure security two techniques have been proposed for carrying out multi keyword ranked search. These techniques are as follows: Privacy preserving scheme in known cipher text model and privacy preserving scheme in known background model. Basic process for performing both of the techniques is same. This process constitutes four steps which are Setup, BuildIndex, Trapdoor and Query. Mathematical designs and statistical proofs have been given in order to validate the results. The proposed technique has less communication and computation overhead. Existing Boolean keyword searchable encryption schemes do not support multi keyword ranked search over encrypted cloud data while preserving privacy. The limitation of this technique is the linear traversing of the whole index of all the documents for every search request.

Tamboli et.al proposed a technique of fuzzy keyword search for retrieving exactly matched documents [10]. It provides the ability to search the closest possible matching document if the exact match does not exist. For quantification of keyword similarity the idea of edit distance has been used. This system provides the facility to encrypt the text files, image files, and video files. Security has been ensured by the encryption of the data to be searched by the user. Various algorithms have been proposed for the creation of fuzzy set. The basic technique follows the process such that at first the keyword is taken as an input and in a database fuzzy set is maintained. The documents which are matched are returned by the database. A private key id is used for the decryption of searched document. The technique provides an efficient search and concept of exact search is also introduced. Experimental results have been shown which are not truly demonstrating the analytical scenario of the experiments.

Traditional schemes for searchable encryption have only permitted the Boolean search without checking the relevant files during search [7], [12], [13], [14], [15]. These schemes lack the ability of large data set searching and exact match search. The user had to have pre knowledge of encrypted data placed on the cloud. Therefore accurate file retrieval was a major issue in traditional searchable systems. Hence in this paper a new model has been proposed for position based sentence searching based on relevant ranking using bloom filters and hashing algorithms.

## IV. Proposed Framework

The proposed framework for PBSS is shown in Fig 1. There entities are involved in complete operations 1) Data Owner, 2) Data User and 3) Cloud Server. For all documents of the data owner indexing is performed. When data owner uploads the document to the server its encrypted index is created and the original document after encryption (asymmetric or symmetric) is uploaded to the cloud. Now the cloud has encrypted indexes of documents and the encrypted documents. During search of sentences two possible scenarios can be used to achieve security and privacy of search sentence. The sentence which data user want to search can be sent directly to cloud after converting to codewords similar to those generated in indexing step. For this case the data owner has to share the codeword generation steps with data user and the secret keys as well. In second case the data users can send the searched sentence to the data owner who will convert them to codewords and will send to cloud server to perform search.

When sentence in form of codewords is sent to the cloud the cloud server check the documents indexes and select those documents which contain the codewords. After the data selection from indexes PBSS algorithm is performed and matches the searched sentences in the selected documents data. On basis of standard deviation (SD) and term frequency (TF) the documents are ranked. The ranked documents are then decrypted and original documents are returned to data user on request.

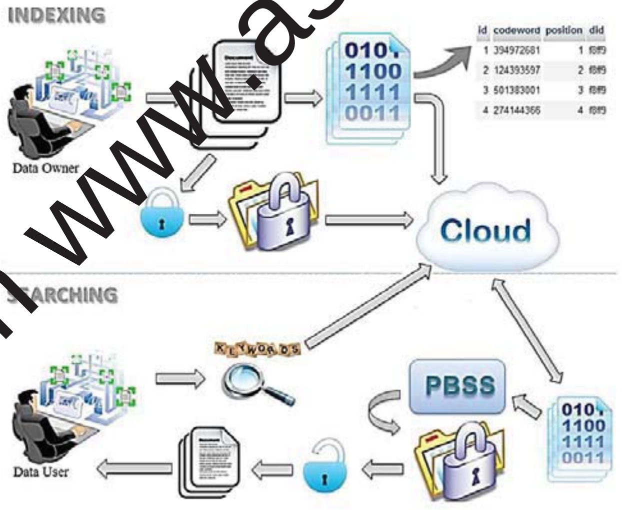

Fig.1. Proposed Framework for PBSS

## V. Proposed Technique

Proposed technique for PBSS comprises of two steps: A) Indexing and B) Searching.

### A. Indexing

The steps invloved in indexing are described below:

**1.** Input Document: Distinct keywords are extracted from the document after removing stop words as shown in Table I.

TABLE I. INPUT DOCUMENT

| | |
|---|---|
| **Input** | Document |
| **Output** | Distinct keywords |
| **Results** | Array ( [1] => searchable, [2] => encryption) |

**2.** Generate Master Key and Split: Master key is generated from password and split into eight equal distinct keys as shown in Table II.

TABLE II. GENERATE MASTER KEY AND SPLIT

| | |
|---|---|
| **Input** | Password |
| **Output** | Eight split keys |
| **Results** | Array ( [0] => e6c83b282aeb2e02, [1] => 2844595721cc00bb [2] => da47cb24537c1779 , [3] => f9bb84f04039e167 [4] => 6e6ba8573e588da1, [5] => 052510e3aa0a32a9 [6] => e55879ae22b0c2d6, [7] => 2136fc0a3e85f8bb ) |

**3.** Generate Trapdoor: Concatinate each word of the document with all eight keys and apply hash (e.g. SHA1). Concatinate all eight hashed values by putting coma after each value as shown in Table III.

TABLE III. GENERATE TRAPDOOR

| | |
|---|---|
| **Input** | Words, eight splits keys, hash algorithm |
| **Output** | Trapdoor |
| **Results** | a104f3e24f622acbdb11b1480c21677b19eacf92,544967dc88058fe18c3e0ba135a6647216fadf82,958452b98248835b769cf0846dfd0b3790a943ad,3a08c7f1b1009ff3fe0ebdddd381c243875ef00a,296708c2666e5e1502403194885f8648719b4467,5e6f0bb55d70f70d0ae30e9b6ff1e0e745e0a408,794cc63d32fb86d6e20fd08f563a8756106b3ccf,394a339521ebc2dea91730501860e7ca02bbbe33 |

**4.** Generate Codeword: Add all trapdoors to bloom filters which will return five bit position for each trapdoor. Concatination of five bit positions yeilds codewords as shown in Table IV.

TABLE IV. GENERATE CODEWORD

| | |
|---|---|
| **Input** | Trapdoor, hash algorithm (crc32) |
| **Output** | Codeword |
| **Results** | 316322629265886458822148 |

**5.** Find positions: Find postions of the keywords of extracted form the document. Codewords of all keywords will have their position in document as shown in Table V.

TABLE V. FIND POSITIONS

| | |
|---|---|
| **Input** | Codeword |
| **Output** | Codeword with position |
| **Results** | Array ( [1] => 316322629265886458822148, [2 165543888265866790136263) |

**6.** Uplaod to Cloud: Upload encrypted document index, encrypted document and documents id‘s.

The document index generated form the above steps is shown in as shown in Table VI.

TABLE VI. DOCUMENT INDEX

| Codeword | Position | Did | enc_doc_name |
|---|---|---|---|
| 316322629265886458822148 | 1 | 3586c170e3e4262f0eb95a0cc24c5ebb3de14504 | hfyGhJXholF3gSTb4g== |
| 165543888265866790136263 | 2 | 3586c170e3e4262f0eb95a0cc24c5ebb3de14504 | hfyGhJXholF3gSTb4g== |

The graphical representation of indexing step is shown in Fig. 2.

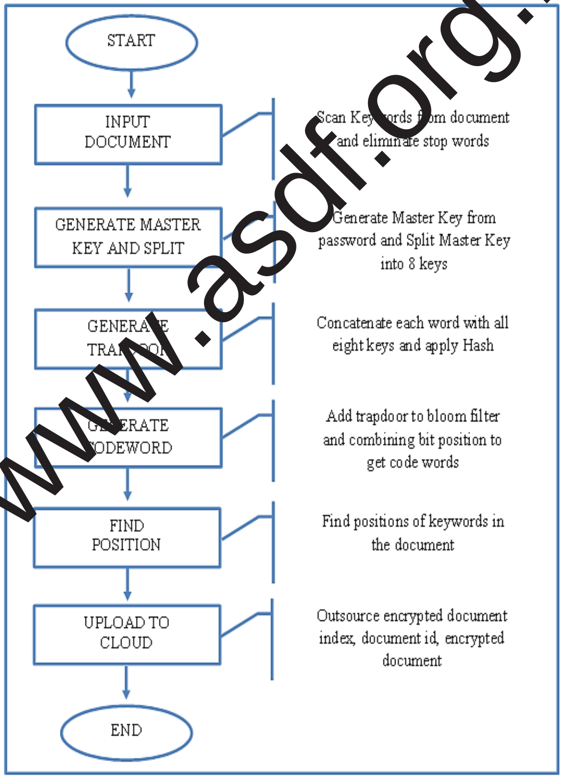
START
INPUT DOCUMENT
Scan Keywords from document and eliminate stop words
GENERATE MASTER KEY AND SPLIT
Generate Master Key from password and Split Master Key into 8 keys
GENERATE TRAPDOOR
Concatenate each word with all eight keys and apply Hash
GENERATE CODEWORD
Add trapdoor to bloom filter and combining bit position to get code words
FIND POSITION
Find positions of keywords in the document
UPLOAD TO CLOUD
Outsource encrypted document index, document id, encrypted document
END

Fig.2. Graphical representation of Indexing

*B. Searching*

During search process the data user will send sentence to be searched from encrypted index. Stop words will be eliminated from the searched sentence. The extracted keywords will be then matched in the document index. If extracted keywords are present at consecutive positions in the document then it is returned. If extracted keywords are present at different positions then SD of all positions is calculated and TF of the extracted keywords is calculated in the documents. The documents are ranked by given formula:

## VI. Conclusion and Future Work

In this paper a technique for position based sentence searching has been proposed. The technique is implemented for encrypted unstructured data place in cloud environment. Encrypted indexes have been created for the distinct keywords extracted from the unstructured data. PBSS provides an efficient sentence search without revealing the documents' content identity. PBSS reduces the overhead of decryption before searches. Documents are decrypted only on retrieval by the data user. If exact sentence is not found and keywords of searched sentence are present in the documents then the spread of keywords is calculated as SD and TF of searched keywords is calculated to rank documents as most relevant. The prepressed index provides efficient and accurate search. The PBSS can be extended for more complex quires, sub match search and other pattern recognition algorithms.

## References

[1] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, ―Privacy- preserving multi-keyword ranked search over encrypted cloud data,‖ in Proc. of INFOCOM, 2011, on 10-15 April, pp 829-837

[2] Gurudatt Kulkarni, Ramesh Sutar, Jayant Gambhir, ―Cloud Computing-Storage as Service‖, International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 1,Jan-Feb 2012, pp.945-950

[3] P. Mell and T. Grance, ―Draft Nist Working Definition of Cloud Computing,‖ http://csrc.nist.gov/groups/SNS/cloudcomputing/ index.html, Jan. 2010.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] Youssef Gahi, Mouhcine Guennoun, Zouhair Guennoun, Khalil El-Khatib, ―Encrypted Processes for Oblivious Data Retrieval‖,6th International Conference on Internet Technology and Secured Transactions, 11-14 December 2011, Abu Dhabi, United Arab Emirates.

[5] C.Antognini.: Bloom Filters, http://antognini.ch/papers/BloomFilters20080620.pdf

[6] C. Wang, N. Cao, K. Ren, and W. Lou, "Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data", IEEE Transactions On Parallel And Distributed Systems, VOL. 23, NO. 8, AUGUST 2012

[7] R. Curtmola, J.A. Garay, S. Kamara, and R. Ostrovsky,―Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions,‖ Proc. ACM Conf. Computer and Comm. Security (CCS'06), 2006.

[8] Park et al.: PKIS: practical keyword index search on cloud datacenter. EURASIP Journal on Wireless Communications and Networking 2011 2011:64

[9] Sun-Ho Lee and Im-Yeong Lee, ―Secure Index Management Scheme on Cloud Storage Environment‖, International Journal of Security and Its Applications, Vol. 6, No. 3, Pages: 75-82, July, 2012

[10] Nikijahan Tamboli, Bharat Savani, Ruchita Choudhari, Nikesh Shah & Apeksha Gadkar, ―Fuzzy Keyword Search Over Encrypted Data‖, International Journal of Computer Science and Electrical Engineering (IJCSEE) ISSN No. 2315-4209, Vol-1 Iss-1, 2012

[11] Madiha Waris, Dr. Shoab Ahmad Khan, ―Indexing of unstructured data for searchable encryption in cloud environment‖, Accepted in IEEE Technically Co-Sponsored Science and Information Conference 2013, London (To be held on 7-9 October 2013)

[12] Dawn Xiaodong Song, David Wagner, and Adrian Perrig,―Practical Techniques for Searches on Encrypted Data‖, In proceedings of IEEE Symposium on Security and Privacy, May 2000.

[13] D. Boneh, G. Di Crescenzo, R. Ostrovsky, and G. Persiano, ―Public Key Encryption with Keyword Search‖. In C. Cachin and J. Camenisch, editors, Advances in Cryptology—EUROCRYPT 2004, volume 3027 of LNCS, pages 506–522. Springer, 2004

[14] E.-J. Goh, ―Secure Indexes,‖ Technical Report 2003/216, Cryptology Print Archive, http://eprint.iacr.org/, 2003.

[15] Y.-C. Chang and M. Mitzenmacher, ―Privacy Preserving Keyword Searches on Remote Encrypted Data,‖ Proc. Int'l Conf. Applied Cryptography and Network Security (ACNS '05), 2005.