A Survey on Interest-Based Search Technique in a Cloud Environment

Vidhya Vishali.KV[1] Dr. V. Vijaya Chamundeeswari[2]
[1,2]Department of Computer Science & Engineering
[1,2]Velammal Engg. College. Chennai, India

Abstract—Cloud computing allows external storage capabilities. Due to hardware limitations in an organization, the organizational documents and data are sent to external scalable cloud storage. The organizational data are bound to reuse and hence the users may have to access the data in future. The organizational data might be considered confidential and hence they are secured in the cloud storage. Also it has to be ensured that the cloud service providers are not allowed to access the data being stored. This can be controlled by minimizing information leakage to the CSP. Since the organization comprises of users from different domains, the area of interest differs for each individual users. Practical challenges in existing search techniques are: Single-key word search provides only limited results, irrelevant search results, Less efficiency in terms of delay. A survey on various methodologies followed for secured search over encrypted environment is presented here. Also, the growth of event-based applications in cloud model handles a new challenges called publish/subscribe paradigm. Here, We have also taken into account the subscription based search paradigm to enhance the interaction between the user and the search system.

Key words: Cloud Computing, Multi-Keywords Text Search, Blind Storage, Publish/Subscribe Paradigm

I. INTRODUCTION

Cloud Computing is the concept that deals with outsourcing services through internet so that the service becomes available anytime anywhere. The hardware limitations in the organizations require scalable cloud storage. This utility based service allows the user to distribute and store huge data set at low cost and are secured by certain encryption standards. These data which need to be outsourced to extended clouds are accessed later by the users. These data includes User details, documents, policies, Organizational standards etc. Some of these data are considered to be confidential and are to be privacy-preserved. Thus they are encrypted before being sent to the external cloud system.

When the user wants to access their corresponding data, many issues may arise due to the search over encrypted data. This is one of the fields under research. The main privacy issue of any out sourced data arises from the cloud itself. Since the data are under control of cloud service providers, the communication that takes place between the user and the cloud storage may be illegally monitored. The primary privacy preserving step involved in cloud is encrypting the data into cipher text so that the data may not be deciphered by the cloud service providers.

But when the data is encrypted the main problem arises is performing search over the encrypted data to achieve relevant and efficient results. Numerous searchable encryption schemes have been proposed to allow search over the cipher text.

Some of the practical features expected from an efficient search technique over the encrypted system are:

- Searchable encryption that allows multiple keyword search, since single keyword search provides very limited and inaccurate results.
- The results retrieved are to be stored in a relevance-based order (Ranking the results).
- Efficient response with minimum delay.
- Reduced Information leakage.

II. RELATED WORKS

A. Dynamic Searchable Encryption with Small Leakage

Generally SSE was used majorly in static environment. Due to emergence of various dynamic versions of applications, Dynamic storage system has been in progress. It allows addition and deletion of documents in the data set which do not affect the search process over the storage. Leakage of information is nothing but unintentional information flow or knowledge about the data being transmitted to the storage system. This has to be maintained at the least for confidentiality purposes. Two main issues were faced in case of dynamic search which has a trade-off viz. Large amount of information leakage and inefficient search and/or update time.

E.Stefanov et al proposed a system [1] which tried to handle both the above stated problems by allowing minimum leakage including only search pattern, Access pattern and size pattern to be leaked out to cloud environment. This system also achieved Forward Privacy – ability to preserve the new keyword-document pair until the word is searched again. But Backward Privacy – id's of deleted documents should be kept private from storage structure, was not addressed. It is also stated that the Dynamic ORAM provides a stronger security capabilities but are not so efficient in practice due to its bandwidth overhead.

B. Multi-Client System with Searchable Encryption

So far, the search over cloud data primarily focused on enhancing single keyword based search only. The author came up with improvising conjunctive keyword search and also proposes methods to handle the complexity of the search for increase in the number of documents available in the storage.

D.Cash et al explored the drawbacks of conjunctive SSE techniques to overcome its search complexity and enormous leakage problem [2]. The proposed method reduced the linear search complexity into sublinear complexity besides support for all Boolean operators like disjunctions, negation etc. The author proposes that reducing the interactions between the user and the server for each search by some pre-computation performed already in the server, the complexity for each search can be made sub-linear. The proposal also ensures high efficiency by leaking only access patterns which are way minimum compared to native leakage pattern available in previous works.

The author considers a system that has multi-client systems hence the leakage of information is not only preserved from server but from other clients as well.

C. Multi-Owner System with Public Key Encryption

To handle the tough trust factors and increase owner capability over the data, more challenging multi-owner and multi-user model has been considered here. The necessity of authorization in search over cloud environment is addressed by M.Li et al. Here Public health record (PHR) system is considered as an instance. Since the system set up is multi-client hence to achieve scalability Public key encryption is preferred over Symmetric key encryption. They address two main issues viz. To reduce privacy exposure as a result of search operation using fine grained search authorization and revocation and in order to provide secured description technique they use Hierarchical Predicate encryption (HPE). The efficiency is achieved by high scalability (Low key management overhead).

It is a model in which the hierarchical levels of predicate vector are compared with plain text vector. The retrieved data is decrypted successfully only when all the conditions are satisfied [3].This method achieved reasonable search performance.

D. Multi-keyword ranked search (MRSE)

In order to overcome curious employee problems, encrypting data before outsourcing to cloud is necessary. In addition to the existing Search encryption (SE) techniques like single keyword search and Boolean search, search over huge data set to support both multi-keyword searches as well as providing similarity based ranked result was challenging and has been addressed in this paper. The similarity based ranking was achieved by "Co-ordinate matching" and used "inner product similarity" to quantify the degree of similarity between the retrieved document set [4].

Co-ordinate matching involved evaluating the matching between the provided query words and those words in the document. Over the retrieved data set, Boolean operations like UNION and INTERSECTION operations are performed based on users' requirement to obtain the final Dataset. This way, the most relevant documents are possible to be retrieved with specified rank order.

It also addressed privacy requirements like keyword privacy, trapdoor privacy and search pattern privacy in 2 levels of threat models Viz.

- Known cipher text model – Minimum Information leakage, Less efficient.
- Known background model; some background information are leaked as well. Statistical attacks on such model are possible using document frequency and keyword frequency.

E. State-Of-The-Art Matching Approach For Index

The author extended the previous work of MRSE where in order to overcome the linear complexity in that approach, W.Sun et al proposed a tree based index structure [5] in which the index is built based on vector space model to achieve higher search accuracy. This approach also considers many Multi-dimensional (MD) algorithms. Though the existing MRSE search attempts to perform privacy preserved multi-keyword search, the search complexity is linear to the number of documents available.

Also state-of-the-art matching approach provides more accurate results than co-ordinate matching system. It involves building the search index in terms of cosine measure between Term frequency * Inverse document frequency. The cosine measure helps us to reach precision in ranking the result set.

- TF → Frequency of a specific term within the document.
- IDF → Ratio of No. of documents holding the specific term to the total number of documents available in the storage medium.

The tree-based index scheme allows each index vector to be broken into sub-vectors that represents subset of keywords and maintained in hierarchical manner. Similarity score is calculated at each level and are summed up to yield the final result set.

They also introduced randomization approach in encryption to achieve index and query confidentiality and query unlink abilities to provide privacy. This way the method has proved to provide balance between accuracy of search results and security of information.

F. Blind Storage

So far all the methodologies dealt with increasing the efficiency and accuracy of the search. Leakage of information issue is primarily addressed by the system proposed by M.Naveed, where a storage approach called blind storage is used [6]. Blind storage is a scheme where in, the files are stored in a remote server in such a way that the server is not aware of the storage model. During retrieval, the server would just know the existence of some file while the file name and its contents are not revealed.

The author has considered that the server is able to provide only storage facilities and no other requirements can be obtained from it. This system let the client keep all the information regarding the file secret from the storage until they are accessed. Search index is also maintained as a part of blind storage. Eg. Scatterstore. – The files are arranged in pseudo-random location. Thus each file is associated with set of locations rather than a specific block. This way it makes the prediction file location tough for the curious employees. The author has performed dynamic searchable encryption over the stored data and has evaluated better performance than existing dynamic searchable encryption using random oracle model.

G. Efficient Fuzzy search

Various techniques have already been discussed on search over encrypted environment while this paper focuses on search with spelling errors rather than supporting only exact matches. This technique is termed as multi-keyword fuzzy search. Achieving this over a multi-keyword is a challenging task without affecting the size and complexity of the index. The author proposes an optimal technique that eliminates the requirement of pre-defined keyword dictionary using Local-sensitive hashing and Bloom filters. It involved index per file, which contains all the keywords available in the document. The index is an m-bit bloom filter. LSH functions are performed to insert the keywords into the Index. They

also used Euclidean distance to evaluate the similarity between the keywords.

H. Publish and Subscribe Paradigm

Any application is expected to be user friendly and hence publish/subscribe paradigm is introduced to enhance personized response. Flow of information – from sender to receiver – is determined by receiver's specific interest. With the publish/subscribe interaction paradigm, the users are allowed to express their interest in particular event rather than all events. When such information is published, the corresponding subscribers are delivered to the subscribers. This act is called notification. This is similar to producer-consumer model. This paper [8] describes the decoupling models available between publisher and subscriber. Space decoupling – the publishers and subscribers do not need to know each other. Time decoupling - Asynchronous interaction is allowed. Synchronization decoupling – Publisher is not blocked while the subscriber is consuming the event. In case of search model, this publish and subscribe can be adopted with time decoupling.

I. Secure KNN computation

Other works also includes some of the querying computations. W.K.Wong, B.Kao et al came up with kNN nearest neighbor concept as their case study for computation. KNN query processing can be considered as one of the core module in data mining tasks [7]. The author proposes a secure K-nearest neighbor scheme which confidentially encrypts 2 vectors and calculate the Euclidean distance between them. Distance preserving transformation is used in the Encrypted database to inherit the distance relationship from the plain text. But when the distance relationship in E(DB) is known and some knowledge about plain text is known.. Entire DB can be recovered. To overcome this Asymmetric scalar product preserving encryption (ASPE).

ASPE – It does not involve direct distance calculation. Instead it compares the distance of 2 points p1 and p2 with candidate point q.

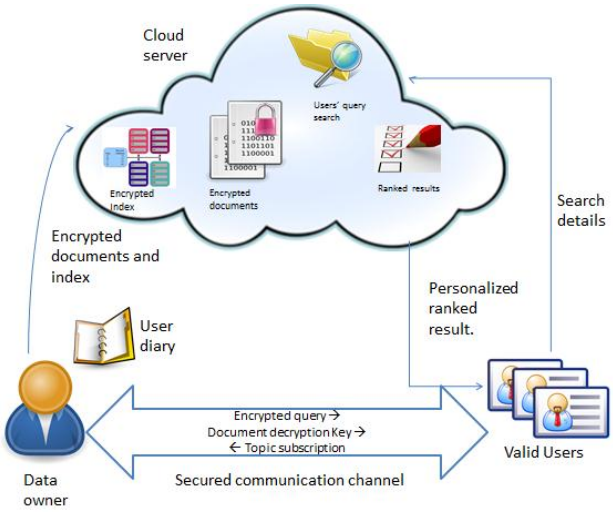
III. PRIMARY CHALLENGES AND THEIR EFFECTIVE SOLUTION

S. No	Challenges	Issues
1	Single-key word search provides only limited results	EMRS
2	Efficiency in terms of minimum delay	Index- Construction using TD & IDF approach
3	Information leakage for increased confidentiality	Blind storage and randomization
4	Interest Based results	Incorporating publish/subscribe paradigm

IV. SYSTEM MODEL

The system consists of 3 components viz., Data owner, Registered Users and the cloud storage system. The system assumes that there exist an authorized communication channel between data owner and the user. The documents in the cloud may be uploaded by the user via the data owner. The data owner acts as an intermediate between the user and the cloud system where the data to be out sourced is encrypted before being sent to the cloud.

For every document uploaded an index is maintained and updated where in the keywords are attributed with list of documents containing the keyword. The index is maintained in the cloud storage in encrypted from.



The data owner also maintains a dictionary that holds the keys of encryption of the records in the cloud. The registers users are allowed to access only specific data from the cloud. The user who wants to access the records from the cloud requests the data owner to obtain the authorization key. The user then built a query vector Q and is sent to the cloud for searching along with optimal value k. The data owner builds an encrypted query vector which is understandable by the cloud to perform the search. Each accessed document is accounted for rank analysis.

The relevant documents are retrieved on which ranking is done to obtain k document result set. The ordering of the document also considers the subscription details of the user.

The retrieved results are decrypted at the user end using the secret key obtained from the owner. The interest of each user is registered as subscription in the user diary maintained by the data owner. Topic based subscription is followed where in the valid users are notified of specific document addition.

V. PROBLEM DESCRIPTION

We have reviewed the papers for an encrypted cloud environment. The main problem addressed here are efficiency and information leakage minimization. Here, an idea is proposed, which uses multi-keyword search with encrypted index stored in the cloud to increase the search efficiency while maintaining a user-dairy to provide interest based search results. This approach also limits the number of results being retrieved based on users choice. The relevance is checked by k-nearest neighbor technique backed up by a dictionary maintained. Thus the proposed system aims to provide

- Multi keyword search via searchable encryption technique.
- Efficiency through index
- Interest based results through User diary

This way results are searched and retrieved based on user's interest with high relevance and efficiency. For index confidentiality, the indexes may be encrypted or can be stored in external environment. The documents being stored in blind storage which relatively conceals the access pattern of the users, i.e., this stops the cloud server from knowing about what documents the user search result set may require. Our system not only addresses the stated issues but also allows results retrieved based on subscription using matching and dispatching. This idea is yet to be implemented and evaluated.

VI. CONCLUSION

Cloud storage is known for storage services in order to overcome the hardware limitations available in organizations. Maintaining privacy of those data from curious employees becomes necessary. Security in cloud environment is still under research. In this paper, we reviewed all the techniques which have its application in effective search over an encrypted cloud environment along with issues which are addressed and those are not. Of these techniques, multi-keyword search encryption along with ranking over a blind storage allows more accurate and efficient search mechanism. This way the document search is more enhanced with increased security constraints. The proposed idea also considers interest-based notification concept and hence publish/subscribe paradigm is adapted to achieve so. We aim to provide a moderate system to address both efficiency and privacy of the data.

REFERENCES

[1] E.Stefanov, C.Papamanthou, E.Shi, "Practical dynamic searchable encryption with small leakage", in Proc. NDSS, Feb. 2014.
[2] D.Cash, S.Jarecki, C.Jutla, "Highly-scalable searchable symmetric encryption with support for Boolean queries", in Proc. CRYPTO, 2013, pp. 353-373.
[3] M.Li, S.Yu, N.Cao, W.Lou," Authorized Private Keyword Search Over encrypted data in cloud computing", 31st International Conference on Distributed Computing Systems (2011)
[4] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, ``Privacy-preserving multi-keyword ranked search over encrypted cloud data," IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 1, pp. 222-233, Jan. 2014.
[5] 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. Communication . Security, 2013, pp. 71-82.
[6] 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.
[7] W. K. Wong, D. W. Cheung, B. Kao, and N. Mamoulis, ``Secure kNN computation on encrypted databases," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2009, pp. 139152.
[8] Vinay Jeyavarma shetty, Ph.D thesis on " Publish/subscribe for large-scale social interaction : Design analysis and Resource provisioning. Jan 2015.
[9] 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. 21122120.

