International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 5, Volume 2 (May 2015) www.ijirae.com

A PERUSAL STUDY OF METHODS USED FOR EFFICIENT SEARCH OVER ENCRYPTED DATA

Jayanand A. Kamble *
PG Scholar, Dept. of CSE,
GCOE, Aurangabad.

Vaishali P. Barde
Lecturer, Dept. of E&TC,
MIT (P), Aurangabad.

Abstract— With the increase in business large amount of data is getting generated continuously. Storing and accessing this data becoming a huge concern for the business owner. Also storing this huge data requires large storage devices which incurs a considerable cost for purchasing these devices. So a business owner may outsource this data on storage devices owned by someone else. The storage devices owned by someone is a cloud acting as infrastructure as a service. Here cloud service provider will provide storage space to a business owner, as per his requirement, with some charges. The business owner will store his or her data on space provided by cloud service provider. Now what about security of this data? Here security and privacy of the sensitive data is a huge concern. At the cloud service provider end, there is lack of internal security due to this cloud data storage becomes insecure. So, sensitive data needs to be encrypted before outsourcing it on cloud for protection of data privacy. Here securing data largely depends on the cryptographic techniques. Once data is encrypted using any encryption algorithm, it can be stored on cloud without fear of leaking of sensitive information. Now fast and efficient accessing of this encrypted data becomes a huge problem. Even though conventional searchable encryption methods permit users to securely search over encrypted data through keywords, they are computationally exhaustive and do not scale for large data sources. Encrypted storage protects the data against illegal access, but complicate yet important functionality such as the efficient search on the data. Significant amount of searchable encryption schemes have been proposed in the literature to achieve search over encrypted data with no compromising the privacy. Basically all of them handle strict query matching but not similarity matching; a fundamental requirement for real world applications. Many searching methods are existed based on fuzzy logic, keyword matching, similarity and many more ideas as conferred in this paper. However, almost all of them are focusing on bringing as many as more documents for the given query. These techniques are in fact provides accurate results however they increases the time complexity of the searching techniques.

Keywords— Cryptography, Encryption, Security, Protection, fuzzy logic.

I. INTRODUCTION

In recent years, due to increase in business data, the burden of storing large scale data is also increased. Due to strong data storage and management ability of the cloud server it allows cloud customers to remotely store their data into the cloud server. Hence, large amount of data, ranging from personal health records to emails, are increasingly outsourced into the cloud. Here the privacy of data becomes the huge concern as cloud servers are untrusted. Also many users continuously keep accessing data from cloud server, so, enterprise should always keep concern of its data privacy from the prying eyes over a network. The sensitive data stored on cloud could be secured by intrusion detection system, firewalls etc but privacy breaches is likely to occur in the paradigm. Consequently, some technique should be authorized to shield the user data and user queries from illegal person in the cloud server. To mitigate the fear, sensitive data is usually outsourced in encrypted form which prevents illegal access.

Though many user think that encryption of data before outsourcing on cloud offers a strong assurance that the data privacy would be preserve against the cloud service providers. For instance, the user may encrypt a text document that may be an email or image by using a public key before outsourcing it to the cloud service. As, only the user knows the public key, the service provider could not violate the privacy of text document or email. Though encryption provides privacy, it makes data utilization a challenging task such that it complicates the computation on the data such as the fundamental search operation being carried out on cloud. Without keyword search function the cloud will become a remote storage which provides limited value to all the enterprises that store its data on the cloud.

To guarantee the benefits of cloud computing environment cloud services should allow efficient search on the encrypted data. To support the searching task considerable amount of algorithms have been proposed which are called searchable encryption schemes. They facilitate selective retrieval of the data from the cloud according to the existence of a specified attribute. It is more natural to carry out retrieval according to the similarity with the specific feature instead of the existence of it. Lacking of effective mechanisms to ensure the file retrieval accuracy is a significant drawback of existing searchable encryption schemes in the context of cloud computing.

II. RELATED WORK

1. Practical Techniques for Searches on Encrypted Data
   Storing data on data storage servers such as mail servers and file servers in encrypted form to reduce security and privacy risks is desirable. However this typically implies that one has to give up functionality for security.

For instance, if a client desires to retrieve only documents containing certain words, it was not formerly known how to let the data storage server complete the search and answer the query without loss of data privacy. In this paper, Song et. al. explained cryptographic schemes for the problem of searching on encrypted data and provided proofs of security for the resulting crypto systems [1]. Their techniques have a number of vital advantages. They are provably protected: they provide provable secrecy for encryption, in the manner that the untrusted server cannot learn anything regarding the plaintext when simply given the cipher-text; they present query separation for searches, in the sense that the untrusted server cannot learn anything more about the plaintext than the search outcome; they provide controlled searching, so that the untrusted server without the user's authorization cannot search for an random word; they also support hidden queries, so that without revealing the word to the server the user might ask the untrusted server to search for a secret word. The algorithms Song et. al. presented were simple, fast (for a document of length , the encryption and search algorithms only need stream cipher and block cipher operations), and introduced almost no room and communication overhead, and hence are realistic to use today.

2. Secure Ranked Keyword Search over Encrypted Cloud Data

As Cloud Computing becoming common now a days, sensitive data are being increasingly centralized into the cloud. To protect the data privacy, sensitive data needs to be encrypted before outsourcing, which makes efficient data utilization a very difficult task. Although conventional searchable encryption schemes permit users to securely search over encrypted data through keywords, these techniques uses only Boolean search, without capturing any significance of data files. When directly applied in the context of Cloud Computing this approach suffers from two main disadvantages. Users, on the one hand, who do not essentially have pre-knowledge of the encrypted cloud data, have to post process each retrieved file in order to discover ones most matching their interest; On the other hand, consistently retrieving all files containing the queried keyword additional incurs unnecessary network traffic, which is totally uninvited in today's pay-as-you-use cloud model.

For the first time Wang et. al. defined and solved the problem of efficient yet secure ranked keyword search over encrypted cloud data [2]. Ranked search deeply improves system usability by returning the identical files in a ranked order concerning to certain relevance criteria (e.g., keyword frequency), thus making one step nearer towards practical deployment of privacy-preserving data hosting services in Cloud Computing. Author first give a straightforward yet perfect building of ranked keyword search under the state-of-the-art searchable symmetric encryption (SSE) security definition, and demonstrate its incompetence. Wang et. al. then propose a definition for ranked searchable symmetric encryption to achieve more practical performance, and give an efficient blueprint by appropriately exploiting the existing cryptographic primitive, order-preserving symmetric encryption (OPSE). Systematic analysis shows that their proposed solution compared to previous SSE schemes, enjoys "as-strong-as-possible" security guarantee, while correctly understanding the goal of ranked keyword search. The efficiency of the proposed solution is demonstrated through extensive experimental results.

3. Secure and Private Sequence Comparisons

Atallah et. al. gave an efficient protocol for sequence comparisons of the edit-distance kind, such that neither party discloses anything about their confidential sequence to the other party [3] (other than what can be concluded from the edit distance between their two sequences – which is inevitable because computing that distance is the purpose of the protocol). The time complexity of the best-known algorithm for performing the sequence comparison is proportional to the amount of communication done by protocol. In a large number of applications, in particular in bioinformatics, the problem of determining the similarity between two sequences arises. In these application areas, the edit distance is one of the most extensively used ideas of sequence similarity: It is the least cost set of deletions, insertions, and substitutions essential to transform one string into the other string. The generalizations of edit distance that are resolved by the similar kind of dynamic programming recurrence relation as the one for edit distance, cover an even wider area of applications.

4. Multi-Probe LSH: Efficient Indexing for High-Dimensional Similarity Search

For feature-rich data such as images, audio, videos, and other sensor data similarity indices for high-dimensional data are very desirable for building content-based search systems. In recent times, locality sensitive hashing (LSH) and its variations have been suggested as indexing techniques for approximate similarity search. A considerable drawback of these approaches in order to achieve good search excellence is the necessity for a huge number of hash tables. This paper offers a new indexing method called multi-probe LSH that defeats this drawback. Multi-probe LSH is constructed on the well-known LSH technique, but it cleverly probes several buckets that are possible to have query results in a hash table [4]. This method is encouraged by and improves upon current theoretical work on entropy-based LSH intended to decrease the space necessity of the basic LSH technique. The multi-probe LSH method is implemented by Josephson et. al. and evaluated the execution with two dissimilar high-dimensional datasets. Their assessment shows the multi-probe LSH method significantly improves upon formerly proposed methods in both space and time efficiency. For achieving the similar search quality, multi-probe LSH has a same time-efficiency as the basic LSH method while reducing the number of hash tables by an order of magnitude. In contrast with the entropy-based LSH method, multi-probe LSH uses less query time and 5 to 8 times fewer number of hash tables to achieve the same search quality.

5. Software Protection and Simulation on Oblivious RAMs
   One of the most important issues concerning computer practice is software protection. For protection there exist many heuristics and ad-hoc methods, but the dilemma as a whole has not received the theoretical treatment it deserves. Goldreich et. al. provided theoretical treatment of software protection [5] here they reduced the problem of software protection to the problem of efficient simulation on oblivious RAM. If the sequence in which machine accesses memory locations is equivalent for any two inputs with the same running time then a machine is said to be oblivious. For example, an oblivious Turing Machine is one for which the movement of the heads on the tapes is indistinguishable for each calculation. (Thus, it is independent of the real input). For a machine to be oblivious, what is the slowdown in the running time? In 1979 Pippenger and Fischer demonstrated how a two-tape oblivious Turing Machine can replicate, online, with a logarithmic slowdown in the running time, a one tape Turing Machine. An analogous result for the random-access machine (RAM) model of computation was shown by Goldreich et. al. In particular, they illustrated how to do an online simulation of an random RAM input by a probabilistic oblivious RAM with a polylogarithmic slowdown in the running time. On the other hand, they also show that a logarithmic slowdown is a lower bound.
6. Enabling Search over Encrypted Multimedia Databases
   Performing information recovery tasks while protecting data privacy is an enviable ability when a database is stored on a server maintained by a third-party service provider. The problem of allowing content-based retrieval over encrypted multimedia databases is addressed in [6]. Search indexes are first encrypted by the content owner, along with multimedia documents and then store onto the server. Through equally applying cryptographic techniques, such as randomized hash functions and order preserving encryption, with image processing and information retrieval techniques, secure indexing schemes are planned to offer both privacy safety and rank-ordered search ability. Safety investigation of the secure indexing schemes under different attack models and retrieval results on an encrypted colour image database show that data confidentiality can be conserved while retaining very good retrieval performance. This work has hopeful applications in secure multimedia management.
7. Secure Similarity Search
   Encryption is one of the most substantial ways to protect users' sensitive information. This paper is regarding the keyword index search system on encrypted documents. It has been considered that the search with errors over encrypted data is impossible because 1 bit dissimilarity over plaintexts may reduce to enormous bits difference over cipher texts. Park et. al. suggested a novel idea to deal with the search with errors over encrypted data [9]. They developed two similarity search schemes, implemented the prototypes and provided extensive analysis and also defined security necessities for the similarity search over encrypted data. The first scheme can accomplish great privacy in similarity search but the second scheme is more competent.
8. Enabling Efficient Fuzzy Keyword Search over Encrypted Data in Cloud Computing
   As Cloud Computing becomes ubiquitous, more and more sensitive information are being centralized into the cloud. For the protection of data privacy, before outsourcing sensitive data usually have to be encrypted, which makes efficient data utilization a very difficult task. Although conventional searchable encryption schemes allow a user to securely search over encrypted data through keywords and selectively retrieve files of interest, these methods support only accurate keyword search. That is, there is no tolerance of slight typos and format inconsistencies which, on the other hand, are characteristic user searching activities and occur very commonly. This major drawback makes existing techniques inappropriate in Cloud Computing as it greatly influences system usability, rendering user searching practices very frustrating and system effectiveness very low. In this paper [10], for the first time Li, Wang et. al. formalized and resolved the problem of efficient fuzzy keyword search over encrypted cloud data while preserving keyword privacy. Fuzzy keyword search to a huge extent enhances system usability by returning the identical files when users' searching inputs exactly match the predefined keywords or the closest probable matching files based on keyword similarity semantics, when precise match fails. In their solution, they utilize edit distance to quantify keywords similarity and develop two superior methods on constructing fuzzy keyword sets, which attain optimized storage and representation expenses. They further proposed a brand new symbol-based tree-traverse searching method, where a multi-way tree structure is built up using symbols transformed from the resulted fuzzy keyword sets. They also show that their proposed solution is secure and privacy-preserving, through rigorous security analysis, while correctly realizing the goal of fuzzy keyword search. Extensive experimental results show the effectiveness of the proposed solution.
9. Efficient Similarity Search over Encrypted Data
   The encrypted storage protects the data against illegitimate access, but it makes difficult some basic, yet important functionality such as the search on the data. To accomplish search over encrypted data without compromising the privacy, substantial amount of searchable encryption schemes have been proposed in the literature. However, almost all of them handle strict query matching but not similarity matching; a vital requirement for real world applications. In this paper, Kuzu et. al. proposed an efficient scheme for similarity search over encrypted data [11]. To do so, they exploit a state-of-the art algorithm for fast near neighbour search in high dimensional spaces called locality sensitive

hashing. To guarantee the confidentiality of the sensitive data, they offered a rigorous security definition and proved the security of the proposed scheme under the provided definition.

III. COMPARISONS OF DATA SEARCHING METHODS IN CLOUD COMPUTING

TABLE I - COMPARISONS OF DATA SEARCHING METHODS

Sr. No	Paper	Method	Process Used	Advantage	Disadvantage
1	[1]	Practical techniques	Pseudo random, function, Pseudo-random generator, Sequential scan, Cryptographic scheme	Provably secure, Efficient and practical. Support controlled and hidden search and query isolation. Simple and fast.	Sequential scan is not efficient if data size is large. Storing and updating the index can be of substantial overhead
2	[2]	Ranked Keyword search	Ranking technique and Order Preserving Mapping Technique	Highly Efficient	Network Traffic occurs, Large amount of Post – Processing of encrypted files
3	[3]	Secure and private sequence compare-sons	protocol for sequence comparisons of the string-edit kind, Edit Distance, Homomorphic Encryption	Efficient Protocol	Time Complexity is same as best known algorithm and not improved.
4	[4]	Multi-Probe LSH	Locality Sensitive Hashing (LSH), Multi-probe LSH Indexing	More time efficient	No experiment for the multi-probe LSH indexing method with a 60-million image dataset.
5	[6]	Secure Indexing Schemes	Secure Inverted Index, Secure Min-Hash Algorithm	Good retrieval performance, Improved efficiency and security of search	No applicable for video.
6	[7]	Secure similarity search	Perfect Similarity Search Privacy, Approximate String Matching, SSS-I, SSS-II	Provides best security, SSS-II scheme is more efficient then SSS-I	SSS-I requires heavy Computational overhead SSS-II cannot provide 'Query Privacy'
7	[8]	Fuzzy keyword search	Symbol-Based tree-Traverse Search Scheme, ram-Based Technique, and Wild card-Based Technique	Edit Distance can be implemented. Highly efficient. Increase searching effectiveness	Large storage complexity. Support only Boolean keyword search. not support Ranked search problem

IV. CONCLUSIONS

In this paper we did rigorous study on different methods for searching encrypted data over cloud. Many searchable techniques have been analyzed which are based on single keyword, multiple keyword search, Ranking, Similarity search, Fuzzy tolerance etc. The majority of the methods described above has problem with them, that is, they are taking more time to search the data. So a significant research is essential to reduce the searching time over the encrypted data in cloud.

REFERENCES

[1] D. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in Proc. of the IEEE Symposium on Security and Privacy'00, 2000, pp. 44–55.
[2] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure ranked keyword search over encrypted cloud data," in Proc. of ICDCS'10, 2010, pp. 253–262.
[3] M. Atallah, F. Kerschbaum, and W. Du, "Secure and private sequence comparisons," in Proc. of the WPES'03, 2003, pp. 39–44.
[4] Q. Lv, W. Josephson, Z. Wang, M. Charikar, and K. Li, "Multi-probe lsh: Efficient indexing for high-dimensional similarity search," in Proc. of VLDB'07, 2007, pp. 253–262.
[5] O. Goldreich and R. Ostrovsky, "Software protection and simulation on oblivious rams," Journal of the ACM, vol. 43, pp. 431–473, 1996.
[6] W. Lu, A. Swaminathan, A. L. Varna, and M. Wu, "Enabling search over encrypted multimedia databases," in Proc.of SPIE Media Forensics and Security'09, 2009.
[7] H. Park, B. Kim, D. H. Lee, Y. Chung, and J. Zhan, "Secure similarity search," in Cryptology ePrint Archive, Report 2007/312, 2007.
[8] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Enabling efficient fuzzy keyword search over encrypted data in cloud computing," in Cryptology ePrint Archive, Report 2009/593, 2009.
[9] Mehmet Kuzu, Mohammad Saiful Islam, Murat Kantarcioglu, "Efficient Similarity Search over Encrypted Data," Department of Computer Science, The University of Texas at Dallas Richardson, TX 75080, USA.

