IJDCST @October Issue- V-1, I-6, SW-50
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

Using State-of the-art Algorithm Fast and Efficient Search over Encrypted Data

P.Swapna[1], D.Hari Krishna[2]

[1]Student, Nova College of Engineering and Technology for Women, Ibrahimpatnam, Krishna Dist., Andhra Pradesh, India

[2]Assistant Professor, Nova College of Engineering and Technology for Women, Ibrahimpatnam, Krishna Dist., Andhra Pradesh, India

Abstract: The cloud computing plays a major role now a days. Due to the appealing features of the cloud computing large amount of data have been stored in cloud. For the protection of data privacy, this makes effective data utilization a very challenging task. Sensitive data has to be encrypted before outsourcing for protection of data privacy. Although traditional searchable encryption schemes allow users to securely search over encrypted data through keywords. Encrypted storage protects the data against illegal access, yet important functionality such as the search on the data. Considerable amount of searchable encryption schemes have been proposed in the literature to achieve search over encrypted data without compromising the privacy. Almost all of them handle exact query matching but not similarity matching; a crucial requirement for real world applications. We propose an efficient scheme for similarity search over encrypted data. We utilize a state-of-the-art algorithm for fast near neighbor search in high dimensional spaces called locality sensitive hashing. We provide a rigorous security definition and prove the security of the proposed scheme under the provided definition to ensure the confidentiality of the sensitive data. We provide a real world application of the proposed scheme and verify the theoretical results with empirical observations on a real dataset**.** To clarify the properties of the proposed scheme, we presented a real world application of it. This application enables keyword search, which is tolerant to the typographical errors in both the queries and the data sources. We illustrated the performance of the proposed scheme with empirical analysis on a real data.

Keywords: Confidentiality, Data Utilization, Security, Encryption, Protection.

I. INTRODUCTION

Cloud Computing enables cloud customers to remotely store their data into the cloud to enjoy the on-demand high quality applications and services from a shared pool of configurable computing resources. The benefits brought by this new computing model include but are not limited to: avoidance of capital expenditure on hardware, software, universal data access with independent geographical locations and relief of the burden for storage management. Cloud computing becomes prevalent because it removes the burden of large-scale data management in a cost effective manner. To mitigate the concerns sensitive data is usually outsourced in encrypted form, which prevents unauthorized access.

Cloud services should enable efficient search on the encrypted data to ensure the benefits of a full-fledged cloud-computing environment. Sizable amount of algorithms have been proposed to support the task which are called searchable encryption schemes. They enable selective retrieval of the data from the cloud according to the existence of a specified feature. It is more natural to perform retrieval according to the similarity with the specified feature instead of the existence of it. A query that specifies a value for a particular feature and a similarity metric to measure the relevance between the query and the data items. The goal is to retrieve the items whose similarity against the specified query is greater than a predetermined threshold under the utilized metric. Some sophisticated cryptographic

techniques enable similarity search over encrypted data.

Data encryption makes effective data utilization a very challenging task given that there could be a large amount of outsourced data files. Data owners in Cloud Computing may share their outsourced data with a large number of users. Cloud Computing might want to only retrieve certain specific data files they are interested in during a given session. Most popular ways to do so is through keyword-based search, and such allows users to selectively retrieve files of interest and has been widely applied in plaintext search scenarios.

Although traditional searchable encryption schemes allow a user to securely search over encrypted data through keywords without first decrypting it. Traditional encryption schemes support only conventional *Boolean* keyword search without capturing any relevance of the files in the search result. For each search request users without pre-knowledge of the encrypted cloud data have to go through every retrieved file in order to find ones most matching their interest. 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. According to a recent survey, secure edit distance computations require over two years to compute similarity between two datasets of 1000 strings each, on a commodity server. The basic building block of our secure index is the state-of the-art approximate near neighbor search algorithm in high dimensional spaces called locality sensitive hashing (LSH).

LSH is extensively used for fast similarity search on plain data in information retrieval community. It is critical to provide rigorous security analysis of the scheme to ensure the confidentiality of the sensitive data. We provide a strong security definition and prove the security of the proposed scheme under the provided definition.

Secure LSH Index: To utilize the appealing properties of LSH in the context of the encrypted data. We adapt the widely accepted adaptive semantic security definition for searchable symmetric encryption schemes and prove the security of the proposed scheme under the adapted definition.

Fault Tolerant Keyword Search: We provide an important application of the proposed scheme for fault tolerant keyword search over encrypted data. A fuzzy keyword set based scheme has been proposed to handle the typographical errors in the search queries and the data sources.

Separation of Leaked Information: Almost all the practical searchable encryption schemes leak some information such as the identifiers of encrypted items corresponding to the trapdoor of a search query. Multi-server setting also enables an alternative encryption scheme with lighter clients by transferring more computational burden to the servers.

II. RELATED WORK

Suppose Alice has a set of sensitive data that she wants to outsource to a cloud server owned by Bob. Data is stored in encrypted form due to the confidentiality, on the remote server in such a way that Bob cannot infer any useful information about the data except for the one Alice allows to leak. It permitted data users should be able to selectively retrieve items from the remote server. Server should be able to search over encrypted data and return the items that are most similar to the user's request in a reasonable amount of time.

Locality Sensitive Hashing

LSH is an approximation algorithm for near neighbor search in high dimensional spaces. The basic idea of LSH is to use a set of hash functions to map objects into several buckets such that similar objects share a bucket with high probability. The result should contain almost all the items that are close to the query point and it should not contain more than a reasonable amount of dissimilar items. We can easily construct such locality sensitive functions from a known family via simple constructions. LSH maps objects into several buckets such that similar objects collide in some buckets while dissimilar ones do not with high probability.

Security Definition

Many secure index based protocols have been proposed for searchable symmetric encryption (SSE). It is used for finding exact matches corresponding to a query and almost all practical SSE schemes leak some information such as the search and access patterns for efficiency. We try to achieve similarity SSE instead of a standard SSE. We extract sub features from each feature via LSH. Number of common components between distinct trapdoors may leak relative similarity between them. Leakage for multi-component trapdoors is captured by our new definition called similarity pattern. Security definitions that allow the leakage of only the search and access patterns have been proposed in the context of SSE. We adapt this definition for our similarity SSE such that similarity pattern is leaked instead of the search pattern.

III. EXISTING SYSTEM

In today's data intensive environment, cloud computing becomes prevalent because it removes the burden of large-scale data management in a cost effective manner. Therefore, large amount of data ranging from personal health records to e-mails are increasingly outsourced into the cloud. At the same time, transfer of sensitive data to un-trusted cloud servers leads to concerns about its privacy. For the protection of data privacy, this makes effective data utilization a very challenging task. Sensitive data has to be encrypted before outsourcing for protection of data privacy. Although traditional searchable encryption schemes allow users to securely search over encrypted data through keywords. Encrypted storage protects the data against illegal access, yet important functionality such as the search on the data. Almost all of them handle exact query matching but not similarity matching; a crucial requirement for real world applications. The data is stored in encrypted form due to confidentiality on the remote server in such a way that Bob cannot infer any useful information about the data except for the one Alice allows to leak. Server should be able to search over encrypted data and return the items that are most similar to the user's request in a reasonable amount of time. Similarity searchable symmetric encryption scheme can easily be achieved by utilizing available secure multiparty computation protocols, which enable distance calculation between encrypted objects. There are some disadvantages in the existing system that listed below:

- Transfer of sensitive data to un-trusted cloud servers leads to concerns about its privacy
- It does not provide more security and we get the files from only exact name required.

IV. PROPOSED SYSTEM

Considerable amount of searchable encryption schemes have been proposed in the literature to achieve search over encrypted data without compromising the privacy. Almost all of them handle exact query matching but not similarity matching; a crucial requirement for real world applications. We propose an efficient scheme for similarity search over encrypted data. We provide a rigorous security definition and prove the security of the proposed scheme under the provided definition to ensure the confidentiality of the sensitive data. We provide a real world application of the proposed scheme and verify the theoretical results with empirical observations on a real dataset**.** To clarify the properties of the proposed scheme, we presented a real world application of it. We provide an overview of our solution. To enable efficient similarity search, Alice builds a secure index and outsources it to the cloud server along with the encrypted data items. Server performs search on the index according to the queries of the data users without learning anything about the data other than what Alice allows an adversary to learn. Traditional encryption schemes support only conventional *Boolean* keyword search without capturing any relevance of the files in the search result. For each search request users without pre-knowledge of the encrypted cloud data have to go through every retrieved file in order to find ones most matching their interest. 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. This application enables keyword search that is tolerant to the typographical errors in both the queries and the data sources. Finally, we illustrated the performance of the proposed scheme with empirical analysis on a real data.

V. STATE-OF-THE-ART TECHNOLOGIES

The state-of-the-art implementations of cloud computing is presented. Technologies used for cloud computing are describes here.

Architectural Design of Data Centre:

A **data center** is a facility used to house computer systems and associated components like telecommunications and storage systems. It includes redundant data communications connections, security devices, redundant or backup power supplies and environmental controls.

Key Design Areas

a. Resilience - ensuring maximum uptime without compromising on performance
b. Availability - business continuity is especially important
c. Performance - the faster the better; in a predictable manner
d. Security - ensuring data separation and controlled access to all Data centre resources
e. Effective architecture for data separation – a common infrastructure that provides facilities for network-based backup and efficient back-end network access
f. Predictable Failover - for maximum service availability

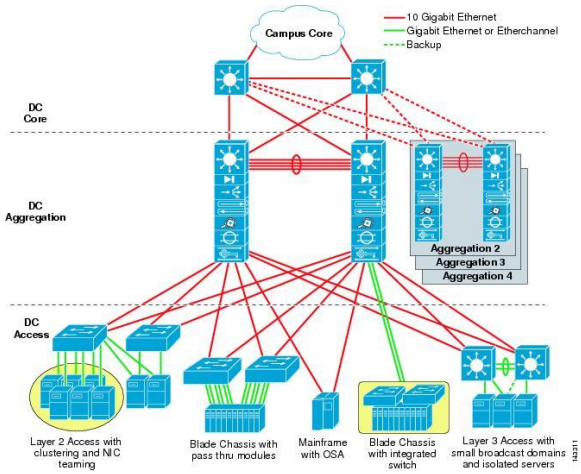

Figure 1: Data Centre Design

Distributed File System over Cloud:

We focused on Google File System that is a proprietary distributed file system developed by Google and specially designed to provide efficient, reliable access to data using large clusters of commodity servers. GFS is designed and optimized to run on data centers to provide extremely high data throughputs, survive individual server failures and low latency. The open source Hadoop Distributed File System stores large files across multiple machines.

Distributed Application Framework over Clouds:

MapReduce is a software framework introduced by Google to support distributed computing on large data sets on clusters of computers. MapReduce consists of one Master to which client applications submit MapReduce jobs. Master pushes work out to available task nodes in the data centre striving to keep the tasks as close to the data as possible. Open source Hadoop MapReduce project is inspired by Google's work. Currently, many organizations are using Hadoop MapReduce to run large data intensive computations.

VI. EXPERIMENTAL ANALYSIS

We investigated the success of our scheme in the context of error aware keyword search, although we would like to stress that the scheme can

be applied to distinct similarity search contexts over encrypted data. We used a publicly available Email dataset, namely the Enron dataset. We constructed a sample corpus of 5000 e-mails via random selection from Enron. We need to determine distance thresholds for our Fuzzy Search scheme according to requirements of our application. We perturbed the keywords by introducing typographical errors and measured the Jaccard distance between the encodings of the original and perturbed versions. We used a publicly available typo generator that produces a variety of spelling errors. We evaluated the performance of both basic and one round multi-server search schemes. We measured the average search time and transferred protocol data for 1000 queries with distinct settings in our local network. Protocol data is the amount of the data transferred between the client and server for the protocol. It does not include the final transfer of the requested items that highly depends on the size of the data items but not the protocol. Both protocol data and search time decreases with increasing *k* as shown in the fig.2. Because of less number of data items are identified by the search due to the more restricted distance range of larger k. Decrease in k and increase in λ have similar effects in terms of final search results. Increase in λ has additional cost of larger trapdoors.

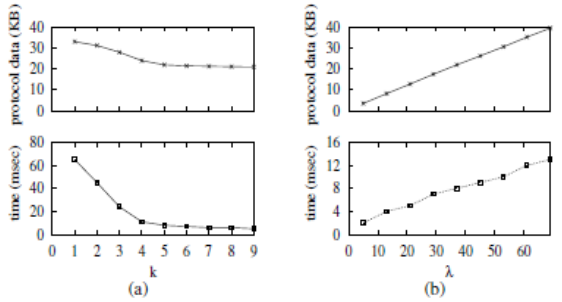

Fig. 2: Basic Scheme Search Performance

Where as in the fig.3 shows the one round scheme search performance related to the state of art algorithm.

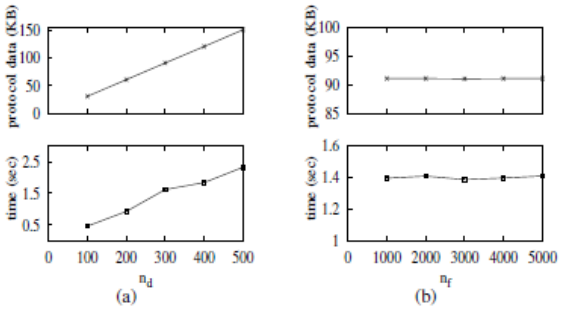

Fig. 3: One Round Scheme Search Performance

VII. CONCLUSION

The advancement of cloud computing is dramatically changing the horizon of information technology and ultimately turns the utility computing into a reality. Opportunities are enough in this arena for some groundbreaking contribution and bring significant development in the industry. We propose an efficient scheme for similarity search over encrypted data. We utilize a state-of-the-art algorithm for fast near neighbor search in high dimensional spaces called locality sensitive hashing. We provide a rigorous security definition and prove the security of the proposed scheme under the provided definition to ensure the confidentiality of the sensitive data. We provide a real world application of the proposed scheme and verify the theoretical results with empirical observations on a real dataset**.** To clarify the properties of the proposed scheme, we presented a real world application of it. We believe our paper will provide a better understanding of the cloud computing and different research issues.

VIII. REFERENCES

[1] 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.

[2] M. Bellare, A. Boldyreva, and A. ONeill, "Deterministic and efficiently searchable encryption," in Proc. of Crypto'07, 2007, pp. 535–552.

[3] Y. Chang and M. Mitzenmacher, “Privacy preserving keyword searches on remote encrypted data,” in Proc. of ACNS’05, 2005, pp. 442–455.
[4] D. Boneh and B. Waters, “Conjunctive, subset, and range queries on encrypted data,” in Theory of Cryptography, ser. LNCS, 2007, vol. 4392, pp. 535–554.
[5] Qi Zhang,Lu Cheng, Raouf Boutaba."Cloud Computing: State-of the- art and research challenges". The Brazilian Computer Society 2010.
[6] S. Kamara and K. Lauter, “Cryptographic cloud storage,” in Proceedings of Financial Cryptography: Workshop on Real-Life Cryptographic Protocols and Standardization 2010, January 2010.
[7] D. Song, D. Wagner, and A. Perrig, “Practical techniques for searches on encrypted data,” in Proc. of IEEE Symposium on Security and Privacy’00, 2000.
[8] 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.
[10] M. Atallah, F. Kerschbaum, and W. Du, “Secure and private sequence comparisons,” in *Proc. of the WPES’03*, 2003, pp. 39–44.
[14] A. Gionis, P. Indyk, and R. Motwani, “Similarity search in high dimensions via hashing,” in *Proc. of VLDB’99*, 1999, pp. 518–529.