An Efficient Privacy Preserving the Ranked Key Word Search Method

1.M.GOPIKRISHNA 2. M CHANDRASEKHAR 3. DR. K NAGESWARARAO

[1]Pg Scholar, Department of CSE, Mother Teresa Institute of Science and Technology, Sathupally
gopikrishna@gmail.com
[2]Assistant Professor , Department of CSE, Mother Teresa Institute of Science and Technology, Sathupally
chandu02508@gmail.com.

[3] Professor & HOD, Department of CSE, Mother Teresa Institute of Science and Technology, Sathupally
nageswararaokapu@yahoo.com.

ABSTRACT— Cloud data owners prefer to outsource documents in an encrypted form for the purpose of privacy preserving. Therefore it is essential to develop efficient and reliable ciphertext search techniques. One challenge is that the relationship between documents will be normally concealed in the process of encryption, which will lead to significant search accuracy performance degradation. Also the volume of data in data centers has experienced a dramatic growth. This will make it even more challenging to design ciphertext search schemes that can provide efficient and reliable online information retrieval on large volume of encrypted data. In this paper, a hierarchical clustering method is proposed to support more search semantics and also to meet the demand for fast ciphertext search within a big data environment. The proposed hierarchical approach clusters the documents based on the minimum relevance threshold, and then partitions the resulting clusters into sub-clusters until the constraint on the maximum size of cluster is reached. In the search phase, this approach can reach a linear computational complexity against an exponential size increase of document collection. In order to verify the authenticity of search results, a structure called minimum hash sub-tree is designed in this paper. Experiments have been conducted using the collection set built from the IEEE Xplore. The results show that with a sharp increase of documents in the dataset the search time

of the proposed method increases linearly whereas the search time of the traditional method increases exponentially. Furthermore, the proposed method has an advantage over the traditional method in the rank privacy and relevance of retrieved documents

INTRODUCTION

AS we step into the big data era, terabyte of data are produced world-wide per day. Enterprises and users who own a large amount of data usually choose to outsource their precious data to

cloud facility in order to reduce data management cost and storage facility spending. As a result, data volume in cloud storage facilities is experiencing a dramatic increase. Although cloud server providers (CSPs) claim that their cloud service is armed with strong security measures, security and privacy are major obstacles preventing the wider acceptance of cloud computing service[1].

A traditional way to reduce information leakage is data encryption. However, this will make server-side data utilization, such as searching on encrypted data, become a very challenging task. In the recent years, researchers have proposed many ciphertext search schemes by incorporating the cryptography techniques. These methods have been proven with provable security, but their methods need massive operations and have high time complexity. Therefore, former methods are not suitable for the big data scenario where data volume is very big and applications require online data processing. In addition, the relationship between documents is concealed in the above methods. The relationship between documents represents the properties of the documents and hence maintaining the relationship is vital to fully express a document. For example, the relationship can be used to express its category. If a document is independent of any other documents except those documents that are related to sports, then it is easy for us to assert this document belongs to the category of the sports. Due to the blind encryption, this important property has been concealed in the traditional methods. Therefore, proposing a method which can maintain and utilize this relationship to speed the search phase is desirable. On the other hand, due to software/hardware failure, and storage corruption, data search results returning to the users may contain damaged data or have been distorted by the malicious administrator or intruder. Thus, a verifiable mechanism should be provided for users to verify the correctness and completeness of the search results.

2 EXISTING SOLUTIONS

In recent years, searchable encryption which provides text search function based on encrypted data has been widely studied, especially in security definition, formalizations and efficiency improvement, the proposed method is compared with existing solutions and has the advantage in maintaining the relationship between documents.

2.1 Single Keyword Searchable Encryption

Song et al first introduced the notion of searchable encryption. They propose to encrypt each word in the document independently. This method has a high searching cost due to the scanning of the whole data collection word by word. Goh et al [9] formally defined a secure index structure and formulate a security model for index known as semantic security against adaptive chosen keyword attack (ind-cka). They also developed an efficient indcka secure index construction called z-idx by using pseudo-random functions and bloom filters. Cash et al recently

design and implement an efficient data structure.Due to the lack of rank mechanism, users have to take a long time to select what they want when massive documents contain the query keyword. Thus, the order-preserving techniques are utilized to realize the rank mechanism, use encrypted invert index to achieve secure ranked keyword search over the encrypted documents. In the search phase, the cloud server computes the relevance score between documents and the query. In this way, relevant documents are ranked according to their relevance score and users can get the top-k results. In the public key setting, Boneh et al designed the first searchable encryption construction, where anyone can use public key to write to the data stored on server but only authorized users owning private key can search. However, all the above mentioned techniques only support single keyword search.

2.2 Multiple Keyword Searchable Encryption

To enrich search predicates, a variety of conjunctive keyword search methods have been proposed. These methods show large overhead, such as communication cost by sharing secret, e.g.computational cost by bilinear map, e.g.[7]. Pang et al propose a secure search scheme based on vector space model. Due to the lack of the security analysis for frequency information and practical search per- formance, it is unclear whether their scheme is secure and efficient or not. Cao et al present a novel architecture to solve the problem of multi-keyword ranked search over encrypted cloud data. But the search time of this method grows exponentially ac- companying with the exponentially increasing size of the document collections. Sun et al give a new architecture which achieves better search efficiency. However, at the stage of index building process, the relevance between documents is ignored. As a result, the relevance of plaintexts is concealed by the encrypttion, users expectation cannot be fulfilled well. For example: given a query containing Mobile and Phone, only the documents containing both of the keywords will be retrieved by traditional methods. But if taking the semantic relationship between the documents

into consideration, the documents containing Cell and Phone should also be retrieved. Obviously, the second result is better at meeting the users expectation.

2.3 Verifiable Search Based on Authenticated Index

The idea of data verification has been well studied in the area of databases. In a plaintext database scenario, a variety of methods have been produced . Most of these works are based on the original work by Merkle and refinements by Naor and Nissm for certificate revocation. Merkle hash tree and cryptographic signature techniques are used to construct authenticated tree structure upon which end users can verify the correctness and completeness of the query results.

Pang et al apply the Merkle hash tree based on authenticated structure to text search engines. However, they only focus on the verification-specific issues ignoring the search privacy preserving capabilities that will be addressed in this paper. The hash chain is used to construct a single keyword search result verification scheme by Wang et al. Sun et al use Merkle hash tree and cryptographic signature to create a verifiable MDB-tree.However, their work cannot be directly used in our architecture which is oriented for privacy-preserving multiple keyword search. Thus, a proper mechanism that can be used to verify the search results within big data scenario is essential to both the CSPs and end users.

3 OUR CONTRIBUTION

In this paper, we propose a multi-keyword ranked search over encrypted data based on hierarchical clustering index (MRSE-HCI) to maintain the close relationship between different plain documents over the encrypted domain in order to enhance the search efficiency. In the proposed architecture, the search time has a linear growth accompanying with an ex- ponential growing size of data collection. We derive this idea from the observation that users retrieval needs usually concentrate on a specific field. So we can speed up the searching process by computing relevance score between the query and documents which belong to the same specific field with the query. As a result, only documents which are classified to the field specified by users query will be evaluated to get their relevance score. Due to the irrelevant fields ignored, the search speed is enhanced. We investigate the problem of maintaining the close relationship between different plain documents over an encrypted domain and propose a clustering method to solve this problem. According to the proposed clustering method, every document will be dynamically classified into a specific cluster which has a constraint on the minimum relevance score between different documents in the dataset. The relevance score is a metric used to evaluate the relationship between different documents. Due to the new documents added to a cluster, the constraint on the cluster may be broken. If one of the new documents breaks the constraint, a new cluster center will be added and the current document will be chosen as a temporal cluster center. Then all the documents will be reassigned and all the cluster centers will be reelected. Therefore, the number of clusters depends on the number of documents in the dataset and the close relationship between different plain documents. In other words, the cluster centers are created dynamically and the number of clusters is decided by the property of the dataset.

4 ARCHITECTURE AND ALGORITHM

4.1 System Model

In this section, we will introduce the MRSE-HCI scheme. The vector space model adopted by the MRSE-HCI scheme is same as the MRSE while the process of building index is totally different. The hierarchical index structure is introduced into the MRSE-HCI instead of sequence index. In MRSE-HCI every document is indexed by a vector. Every dimension of the vector stands for a keyword and the value represents whether the keyword appears or not in the document. Similarly, the query is also represented by a vector. In the search phase, cloud server calculates the relevance score between the query and documents by computing the inner product of the query vector and document vectors and return the target documents to user according to the top k relevance score. Due to the fact that all the documents outsourced to the cloud server is encrypted, the semantic relationship between plain documents over the encrypted documents is lost. In order to maintain the semantic relationship between plain documents over the encrypted documents, a clustering method is used to cluster the documents by clustering their related index vectors. Every document vector is viewed as a point in the n-dimensional space. With the length of vectors being normalized, we know that the distance of points in the n-dimensional space reflect the relevance of corresponding documents. In other word, points of high relevant documents are very close to each other in the n-dimensional space. As a result, we can cluster the documents based on the distance measure. With the volume of data in the data center has experienced a dramatic growth, conventional sequence search approach will be very inefficient. To promote the search efficiency, a hierarchical clustering method is proposed. The proposed hierarchical approach clusters the documents based on the minimum relevance threshold at different levels, and then partitions the resulting clusters into sub-clusters until the constraint on the maximum size of cluster is reached. Upon receiving a legal request, cloud server will search the related indexes layer by layer instead of scanning all indexes.

4.2 MRSE-HCI Architecture

MRSE-HCI architecture is depicted by where the data owner builds the encrypted index depending on the dictionary, random numbers and secret key, the data user submits a query to the cloud server for getting desired documents, and the cloud server returns the target documents to the data user. This architecture mainly consists of following algorithms.

•Keygen (1l(n))→(sk,k):It is used to generate thesecret key to encrypt index and documents.

•Index(D,sk)→I:Encrypted index is generated in this phase by using the above mentioned secret key. At the same time, clustering process is also included current phase.

•Enc(D,k)→E:The document collection is encrypted by a symmetric encryption algorithm which achieves semantic security.

•Trapdoor(w,sk)→T w : It generates encryptedquery vector T W with users input keywords and

secret key.

•Search(Tw,I,k top) → (I w ,E w) : In this phase, cloud server compares trapdoor with index to get the top- k retrieval results.

• Dec(Ew,k)→F w

:The returned encrypted documents are decrypted by the key generated in the first step.

CONCLUSION

In this paper, we investigated ciphertext search in the scenario of cloud storage. We explore the problem of maintaining the semantic relationship between different plain documents over the related encrypted documents and give the design method to enhance the performance of the semantic search. We also propose the MRSE-HCI architecture to adapt to the requirements of data explosion, online information retrieval and semantic search. At the same time, a verifiable mechanism is also proposed to guarantee the correctness and completeness of search results. In addition, we analyze the search efficiency and security under two popular threat models. An experimental platform is built to evaluate the search efficiency, accuracy, and rank security. The experiment result proves that the proposed architecture not only prop- erly solves the multi-keyword ranked search problem, but also brings an improvement in search efficiency, rank security, and the relevance between retrieved documents.

R

EFERENCES

[1] S. Grzonkowski, P. M. Corcoran, and T. Coughlin, ”Security analysis of authentication protocols for next-generation mobile and CE cloud services,” in Proc. ICCE, Berlin, Germany, 2011, pp. 83-87.

[2] D. X. D. Song, D. Wagner, and A. Perrig, ”Practical techniques for searches on encrypted data,” in Proc. S & P, BERKELEY, CA, 2000, pp. 44-55.

[3] D. Boneh, G. Di Crescenzo, R. Ostrovsky, and G. Persiano, ”Public key encryption with keyword search,” in Proc. EURO- CRYPT, Interlaken, SWITZERLAND, 2004, pp. 506-522.

[4] Y. C. Chang, and M. Mitzenmacher, ”Privacy preserving key- word searches on remote encrypted data,” in Proc. ACNS, Columbia Univ, New York, NY, 2005, pp. 442-455.

[5] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, ”Search- able symmetric encryption: improved definitions and efficient constructions,” in Proc. ACM CCS, Alexandria, Virginia, USA, 2006, pp. 79-88.

[6] M. Bellare, A. Boldyreva, and A. O’Neill, ”Deterministic and efficiently searchable encryption,” in Proc. CRYPTO, Santa Bar- bara, CA, 2007, pp. 535-552.

[7] D. Boneh, and B. Waters, ”Conjunctive, subset, and range queries on encrypted data,” in Proc. TCC, Amsterdam, NETHERLANDS, 2007, pp. 535-554.

[8] D. X. D. Song, D. Wagner, and A. Perrig, ”Practical techniques for searches on encrypted data,” in Proc. S & P 2000, BERKE- LEY, CA, 2000, pp. 44-55.

[9] E.-J. Goh, Secure Indexes, IACR Cryptology ePrint Archive, vol. 2003, pp. 216. 2003.

[10] C. Wang, N. Cao, K. Ren, and W. J. Lou, Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data, IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 8, pp. 1467-1479, Aug. 2012.

[11] A. Swaminathan, Y. Mao, G. M. Su, H. Gou, A. Varna, S. He, M. Wu, and D. Oard, ”Confidentiality-Preserving Rank-Ordered Search,” in Proc. ACM StorageSS, Alexandria, VA, 2007, pp. 7-12.

[12] S. Zerr, D. Olmedilla, W. Nejdl, and W. Siberski, ”Zerber+R: top-k retrieval from a confidential index,” in Proc. EDBT, Saint Petersburg, Russia, 2009, pp. 439-449.

[13] C. Wang, N. Cao, J. Li, K. Ren, and W. J. Lou, ”Secure Ranked Keyword Search over Encrypted Cloud Data,” in Proc. ICDCS, Genova, ITALY, 2010.

[14] P. Golle, J. Staddon, and B. Waters, ”Secure conjunctive key- word search over encrypted data,” in Proc. ACNS, Yellow Mt, China, 2004, pp. 31-45.

[15] L. Ballard, S. Kamara, and F. Monrose, ”Achieving efficient conjunctive keyword searches over encrypted data,” in Proc. ICICS, Beijing, China, 2005, pp. 414-426.

[16] R. Brinkman, Searching in encrypted data: University of Twente, 2007