A Secure Ranked Search Model Over Encrypted Data in Hybrid Cloud Computing

Jiuling Zhang^(✉), Shijun Shen, and Daochao Huang

CNCERT/CC, Beijing 100029, China
zhangjl@cert.org.cn

Abstract. The security issue is becoming more and more prominent since user's private information being outsourced to the somewhat untrustworthy cloud. Encrypting the information before uploading them to the cloud is one of ultimate solutions. Secure searchable encryption schemes and secure ranking schemes have been proposed to help retrieving the most relevant documents over the cloud. However the present methods are encumbered by the huge computing and communicating occupation of the cipher text. In this paper, a fully homomorphic encryption based secure ranked search model over the hybrid cloud is proposed. By introducing hybrid cloud, which typically composed by private cloud and public cloud, the high cost of computing and communicating of the cipher text is transferred to the trustworthy private cloud, in which the decrypting are performed. The client does not need to perform any heavy computations, thence making the secure ranking practical from the client's point of view.

Keywords: Secure ranked search · Fully homomorphic encryption · Hybrid cloud computing

1 Introduction

With the unprecedented growing of information, as well as the limited storage and computation power of their terminal, or the limited capacity of battery, the users are outsourcing more and more individual information to remote servers or clouds. However, since the public cloud are not fully trustworthy, the security of the information stored on the cloud could not be guaranteed. The security issue has attracted a variety of attentions in both the area of engineering and research. One solution to meet the needs of data outsourcing while preserving privacy is to encrypt them before storing them on the cloud, and this is one of the generally accepted ultimate methods. After the sensitive data are encrypted with some scheme, the cipher text of sensitive private information may be deemed as secure and could be outsourced to public cloud. However, once the data are encrypted into cipher text, the processing and utilizing of the cipher text information will be the subsequent problem that needs to be taken into consideration. With the accumulation of the information outsourced over the cloud, the collection will be so large that the retrieving of the encrypted form of information is the subsequent conundrum.

Several kinds of schemes have been proposed since the pioneering work of Song et al. [1], in which a cryptography scheme for the problem of searching on encrypted data is proposed. The proposed scheme is practical and provably secure, as from the query the server cannot learn anything more about the plain text. Although the scheme performs well over the linear search, it is almost impractical over the huge information retrieval scenario. In another work, Goh introduced a Bloom filter based searchable scheme with a complexity of the number of documents in the collection over the cloud [2], which is also not applicable in huge information retrieval scenario. Other work includes a secure conjunctive keyword search over the encrypted information with a linear communication cost [6], privacy-preserving multi-keyword ranked searches over encrypted cloud data [3, 10], and fuzzy keyword search over encrypted cloud [4].

In the general huge plaintext retrieval scenario, different pieces of information are organized in the form of documents. The information size stored in the cloud is very large, and the acquisition of the requested information should be implemented with the help of retrieval methods. A number of documents may contain a given query and if the query is searched, many documents may be retrieved. After the more or less relevant documents are retrieved, the ranking of them over the cloud computing is necessary. Actually in the large information retrieval scenario, the retrieved information should be ranked by the relevance scores between the document and the queries. This is due to that the number of documents contains a keyword or a multiple of keywords is so large that it is hard to obtain the most relevant documents from the client's point of view. The most relevant documents should be retrieved and given to the users. In plaintext retrieval, a similarity based ranking schemes named the locality sensitive hashing [7] and an inner product similarity to value the relevance between the query and the document are separately presented [4].

In huge collection cipher text retrieval, a one-to-many order-preserving mapping technique is employed to rank sensitive score values [13]. A secure and efficient similarity search over outsourced cloud data is proposed in [14]. There are also work exploring semantic search based on conceptual graphs over encrypted outsourced data [8]. Though the one-to-many order-preserving mapping design facilitates efficient cloud side ranking without revealing keyword privacy and there is no cost on the client's terminal, the precision of this model is lower than that over the plaintext scenario. There is a counterbalance between the accuracy of the results and the security as the statistical information is provided. There are also efforts utilizing the fully homomorphic encryption [5] to calculate the relevance scores between the documents and queries. However, since the encryption and decryption are all performed on the client's terminal, and they are also resource consuming, the time cost is also intolerable.

In order to solve the problem that enormous computation and communication emerged in the fully homomorphic encryption based ranking, the hybrid cloud [12] is introduced to employ. Hybrid cloud generally consists public cloud and private cloud. The public cloud is provided by the entrepreneur, and not fully trust worthy, while the private cloud belongs to the organization, and thus trustable. Hybrid cloud also described the architecture and cooperation among different cloud vendors, and gave solution on the communication, storage, and computation among different cloud [11]. Here, we make the assumption that there is at least one secure private cloud in the hybrid cloud. The

plain text information is handled over the trustworthy or private cloud. With the cooperation with other public clouds on which encrypted information are stored and processed, the secure ranking model is proposed.

This paper is organized as follows, the related work is reviewed in Sect. 2, and then a secure ranked search model over the hybrid cloud is introduced in Sect. 3. Some experiments are carried out in Sect. 4. Finally, a conclusion is drawn in Sect. 5.

2 Related Work

2.1 The Okapi BM25 Model Over Plain Text

In information retrieval, a document D is generally processed into a bag of words. The collection of documents is denoted by C . The documents and queries are generally preprocessed and stemmed, the index and inverted index are also built to facilitate further retrieval [9], the details are omitted here.

There are a variety of mature information retrieval models, which varies from the linear search to Boolean model to ranked vector space model (VSM) models. Different retrieval model applies in different scenarios. The ranking models are used most frequently for general purposes.

Okapi BM25 model [15] is one of the most popular ranking model for obtaining the relevance scores of documents and queries. In the Okapi BM25 model, the term frequency is defined by Eq. 1.

$$TF(q_i) = f(q_i, D) \quad (1)$$

While the inverse document frequency is given by Eq. 2.

$$IDF(q_i) = \log \frac{N}{n(q_i)} \quad (2)$$

In which $f(q_i, D)$ means the occurrence frequency of $n(q_i)$ in D . $n(q_i)$ means the number of documents which contain q_i .

The Okapi relevance scores between a query and a document is given by Eq. 3.

$$Score(D, Q) = \sum_{q_i \in Q} TF(q_i) \times IDF(q_i) \quad (3)$$

The relevance between a document and a query is quantified by the Okapi relevance scores.

2.2 Fully Homomorphic Encryption

Homomorphism is a very valuable property of encryption algorithms, which means that the computation results over cipher texts corresponds to that of the computation over plaintext. Fully homomorphic encryption (FHE) is both additive homomorphic and multiplicative homomorphic, satisfying both the Eqs. 4 and 5.

$$D(E(a) \oplus E(b)) = a + b \quad (4)$$

$$D(E(a) \otimes E(b)) = a \times b \quad (5)$$

Where $\oplus$ means the “addition” over the cipher text, while $\otimes$ denotes the “multiplication” over the cipher text.

In this work, the term frequency TF and inverse document frequency IDF values are encrypted by FHE separately. The documents which contain the terms are encrypted by some other encryption scheme, such as AES, only to protect the information stored on the public cloud.

All the information is thence uploaded to the public cloud after encrypted by a certain encryption scheme. The cipher text of Score (D, Q) could also be obtained.

2.3 The Applicability of Hybrid Cloud Computing

We assume that the hybrid cloud is simply constructed by one private cloud and one public cloud. The private cloud stores the client’s sensitive information and the public cloud performs computation over cipher text information.

A new scheme based on the private and the public cloud platform is proposed here. The public cloud in this hybrid cloud scenario is assumed to have the following characteristics: the computing resource is very enormous, and the resource allocated to a client can be elastically provided in order to meet the client’s computation demands.

The private cloud actually acts as an agent for the client in the scenario. Since the computation and storage resources are relatively abundant over private cloud, it has enough computing power to just encrypt a user’s plaintext information. The bandwidth between the private cloud and the public cloud is also large enough to transfer the cipher text of the relevance scores.

3 Secure Ranked Search Model Over Hybrid Cloud

A new encryption based secure and efficient retrieval scheme over hybrid cloud is proposed in this section.

3.1 The Architecture of the Secure Ranked Search Model Over Hybrid Cloud

There are three parties in this architecture, the client, the private cloud, and the public cloud. As shown in Fig. 1.

In the building process, the client uploads original sensitive information to the private cloud, as shown by step (1) in Fig. 1. The private cloud preprocesses the documents, and encrypts the TF, IDF values and the document itself. The encrypted information are then uploaded to the public cloud, as shown by step (2). Over the public cloud, an inverted index is built, and a variety of corresponding computations are performed.

In the retrieval process, the client gives a certain keyword to the private cloud, as shown by step (3). The private cloud encrypts the word, and search over the public cloud, as shown by step (4). On the public cloud, the calculation over the cipher text is carried out. The cipher text of evaluation scores are downloaded by the private cloud, as shown

Fig. 1. Schematic diagram of secure ranked search model

by step (5). After the decryption, the scores are ranked, thence the top N document IDs are sent to the public cloud, as shown by step (6). Then the private cloud downloads the encrypted document, as shown by step (7). After decryption, the plaintext documents are given back to the clients, as shown by step (8).

3.2 The Implementation of Fully Homomorphic Encryption Based Secure Ranking Scheme

In the inverted index building process, the computation of encryption of the plain text are performed over the private cloud.

The encrypted form of term frequency is expressed as Eq. 6.

$$v_{tf} = (FHE(tf_1), FHE(tf_2), \dots, FHE(tf_N)) \quad (6)$$

The encrypted form of inverse document frequency is given as Eq. 7.

$$v_{idf} = (FHE(idf_1), FHE(idf_2), \dots, FHE(idf_N)) \quad (7)$$

In the ranking process, the computation such as the addition and multiplication over the cipher text are performed over the public cloud.

The full process can be described as the following, Firstly the TF and in decimal form are transformed into binary, then each of them is encrypted, the relevance is obtained after addition and multiplication. The process is shown in Fig. 2.

Fig. 2. Implementation of the FHE based ranking

The process of calculating relevance scores between the query and the document is given as Eq. 8.

$$\text{FHE}(\text{score}) = \sum_{q_i} \text{FHE}(\text{tf}_i) \times \text{FHE}(\text{idf}_i) \quad (8)$$

Thence the relevance scores in FHE form are obtained over the hybrid cloud. By decrypting them, the documents could be subsequently ranked.

4 Experiment Result and Future Work

4.1 Preliminary Experimental Result

Based on the proposed retrieval and ranking model over hybrid cloud, some preliminary experiments are carried out. The experiment utilized a small-sized Cranfield collection. The experimental result is compared with the order preserving scheme (OPE), which is employed in [13].

The precision of top N retrieved documents and the MAP [9] are used to evaluate different ranking schemes. The experimental result is shown in the following table (Table 1).

Table 1. The comparison result of different methods.

Metric	OPE	Okapi BM25
Map	0.283	0.416
P@5	0.310	0.440
P@10	0.223	0.310
P@20	0.150	0.199
P@30	0.117	0.150
P@50	0.082	0.103
P@100	0.050	0.059

The tentative experimental result demonstrates that the order preserving encryption based retrieval result is dramatically lower than that of the Okapi BM25 ranking models for the crucial P@N criteria.

4.2 Future Work

While retrieving, the proposed scheme needs the private cloud to download all cipher text of the relevance scores of possibly relevant documents, which also would be enormous. In order to make it more practicable, the future work may incorporate both the OPE and the FHE over the hybrid cloud. By OPE, a pre-rank could be performed over the public cloud, and give a top M relevance scores to private cloud. Here, M should be a

large enough number, say 10000. Then the private cloud then decrypts the top M scores and ranks them. By this way, both the computation and communication cost over the private cloud would be limited, the efficiency of retrieving and ranking will be greatly enhanced.

5 Conclusion

A fully homomorphic encryption based secure ranked search model over the hybrid cloud is proposed, the implementation of the retrieval and ranking process are described in detail. Experimental result shows its precedence over the existing purely OPE based ranking. In the future, we would incorporate both OPE and the FHE to implement industrial model while preserving user's privacy over the hybrid cloud.

Acknowledgments. This work is supported by The National Key Research and Development Program of China (2016YFB1000105).

References

1. Song, D.X., Wagner, D., Perrig, A.: Practical techniques for searches on encrypted data. In: Proceeding 2000 IEEE Symposium on Security and Privacy, pp. 44–55 (2000)
2. Goh, E.-J.: Secure indexes. IACR Cryptology ePrint Archive, p. 216 (2003)
3. Curtmola, R., Garay, J., Kamara, S., Ostrovsky, R.: Searchable symmetric encryption: improved definitions and efficient constructions. *J. Comput. Secur.* **19**, 895–934 (2011)
4. Li, J., Wang, Q., Wang, C., Cao, N., Ren, K., Lou, W.: Fuzzy keyword search over encrypted data in cloud computing. In: 2010 Proceedings IEEE INFOCOM, pp. 1–5 (2010)
5. Gentry, C.: Fully homomorphic encryption using ideal lattices. In: STOC 2009
6. Golle, P., Staddon, J., Waters, B.: Secure conjunctive keyword search over encrypted data. In: ACNS (2004)
7. Kuzu, M., Islam, M.S., Kantarcioglu, M.: Efficient similarity search over encrypted data. In: 2012 IEEE 28th International Conference on Data Engineering, pp. 1156–1167 (2012)
8. Fu, Z., Huang, F., Sun, X., Vasilakos, A., Yang, C.: Enabling semantic search based on conceptual graphs over encrypted outsourced data. *IEEE Trans. Serv. Comput.* **12**, 813–823 (2019)
9. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, 1st edn. Cambridge University Press, Cambridge (2005)
10. Vishvapathi, P., Reddy, M.J.: Privacy-preserving multi-keyword ranked search over encrypted cloud data (2016)
11. Wang, H., Ding, B.: Growing construction and adaptive evolution of complex software systems. *Sci. China Inf. Sci.* **59**, 1–3 (2016)
12. Wang, H., Shi, P., Zhang, Y.: JointCloud: a cross-cloud cooperation architecture for integrated internet service customization. In: IEEE 37th International Conference on Distributed Computing Systems, pp. 1846–1855 (2017)
13. Wang, C., Cao, N., Ren, K., Lou, W.: Enabling secure and efficient ranked keyword search over outsourced cloud data. *IEEE Trans. Parallel Distrib. Syst.* **23**, 1467–1479 (2012)
14. Wang, C., Ren, K., Yu, S., Urs, K.M.: Achieving usable and privacy-assured similarity search over outsourced cloud data. In: Proceedings IEEE INFOCOM, pp. 451–459 (2012)
15. Whissell, J.S., Clarke, C.L.: Improving document clustering using Okapi BM25 feature weighting. *Inf. Retr.* **14**, 466–487 (2011)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.