International Refereed Journal of Engineering and Science (IRJES)
ISSN (Online) 2319-183X, (Print) 2319-1821
Volume 2, Issue 12(December 2013), PP.119-123

Privacy-Preserving Multi-keyword Ranked Search over Encrypted Cloud Data

[1] B.Kumar Swamy,[2] K. Mahesh Kumar,[3 B.] Hari Kumar
[1,2,3] *Computer Science Engineering Department, Sree Dattha Institute Of Engineering & Science*

ABSTRACT: The advent of cloud computing, data owners are motivated to outsource their complex data management systems from local sites to commercial public cloud for great flexibility and economic savings. But for protecting data privacy, sensitive data has to be encrypted before outsourcing, which obsoletes traditional data utilization based on plaintext keyword search. Thus, enabling an encrypted cloud data search service is of paramount importance. Considering the large number of data users and documents in cloud, it is crucial for the search service to allow multi-keyword query and provide result similarity ranking to meet the effective data retrieval need. Related works on searchable encryption focus on single keyword search or Boolean keyword search, and rarely differentiate the search results. In this paper, for the first time, we define and solve the challenging problem of privacy-preserving multi-keyword ranked search over encrypted cloud data (MRSE), and establish a set of strict privacy requirements for such a secure cloud data utilization system to become a reality. Among various multi-keyword semantics, we choose the efficient principle of "coordinate matching", i.e., as many matches as possible, to capture the similarity between search query and data documents, and further use "inner product similarity" to quantitatively formalize such principle for similarity measurement. We first propose a basic MRSE scheme using secure inner product computation, and then significantly improve it to meet different privacy requirements in two levels of threat models. Thorough analysis investigating privacy and efficiency guarantees of proposed schemes is given, and experiments on the real-world dataset further show proposed schemes indeed introduce low overhead on computation and communication.

I. INTRODUCTION

Due to the rapid expansion of data, the data owners tend to store their data into the cloud to release the burden of data storage and maintenance [1]. However, as the cloud customers and the cloud server are not in the same trusted domain, our outsourced data may be under the exposure to the risk. Thus, before sent to the cloud, the sensitive data needs to be encrypted to protect for data privacy and combat unsolicited accesses. Unfortunately, the traditional plaintext search methods cannot be directly applied to the encrypted cloud data any more. The traditional information retrieval (IR) has already provided multi-keyword ranked search for the data user. In the same way, the cloud server needs provide the data user with the similar function, while protecting data and search privacy. It is meaningful storing it into the cloud server only when data can be easily searched and utilized.

In the literature, searchable encryption techniques [2-4] are able to provide secure search over encrypted data for users. They build a searchable inverted index that stores a list of mapping from keywords to the corresponding set of files which contain this keyword. When data users input a keyword, a trapdoor is generated for this keyword and then submitted to the cloud server

Some researchers study the problem on secure and ranked search over outsourced cloud data. Wang *et al.*, [5] propose a secure ranked keyword search scheme. Their solution combines inverted index with order-preserving symmetric encryption (OPSE). In terms of ranked search, the order of retrieved files is determined by numerical relevance scores, which can be calculated by $TF \times IDF$. The relevance score is encrypted by OPSE to ensure security. It enhances system usability and saves communication overhead. This solution only supports single keyword ranked search. Cao *et al.*, [6] propose a method that adopts similarity measure of "coordinate matching" to capture the relevance of files to the query. They use "inner product similarity" to measure the score of each file. This solution supports exact multi-keyword ranked search. It is practical, and the search is flexible. Sun *et al.*, [7] proposed a MDB-tree based scheme which supports ranked multi-keyword search. This scheme is very efficient, but the higher efficiency will lead to lower precision of the search results in this scheme.

In addition, fuzzy keyword search [8-10] have been developed. These methods employ a spell-check mechanism, such as, search for "wireless" instead of "wireiess", or the data format may not be the same *e.g.*, "data-mining" versus "datamining. Chuah *et al.*, [8] propose a privacy-aware bed-tree method to support fuzzy multi-keyword search. This approach uses edit distance to build fuzzy keyword sets. Bloom filters are

constructed for every keyword. Then, it constructs the index tree for all files where each leaf node a hash value of a keyword. Li *et al.,* [9] exploit edit distance to quantify keywords similarity and construct storage-efficient fuzzy keyword sets. Specially, the wildcard-based fuzzy set construction approach is designed to save storage overhead. Wang *et al.,* [10] employ wildcard-based fuzzy set to build a private trie-traverse searching index. In the searching phase, if the edit distance between retrieval keywords and ones from the fuzzy sets is less than a predetermined set value, it is considered similar and returns the corresponding files. These fuzzy search methods support tolerance of minor typos and format inconsistencies, but do not support semantic fuzzy search. Considering the existence of polysemy and synonymy [11], the model that supports multi-keyword ranked search and semantic search is more reasonable.

In this paper, we will solve the problem of multi-keyword latent semantic ranked search over encrypted cloud data and retrieve the most relevant files. We define a new scheme named Latent Semantic Analysis (LSA)-based multi-keyword ranked search which supports multi-keyword latent semantic ranked search. By using LSA, the proposed scheme could return not only the exact matching files, but also the files including the terms latent semantically associated to the query keyword. For example, when the user inputs the keyword "automobile" to search files, the proposed method returns not only the files containing "automobile", but also the files including the term "car" We take a large matrix of term-document association data and construct a semantic space wherein terms and documents are closely associated are placed near one another. To meet the challenge of supporting such multi-keyword semantic without privacy breaches, we propose the idea: the multi-keyword ranked search (MRSE) using "Latent Semantic Analysis

II. ARCHITECTURE

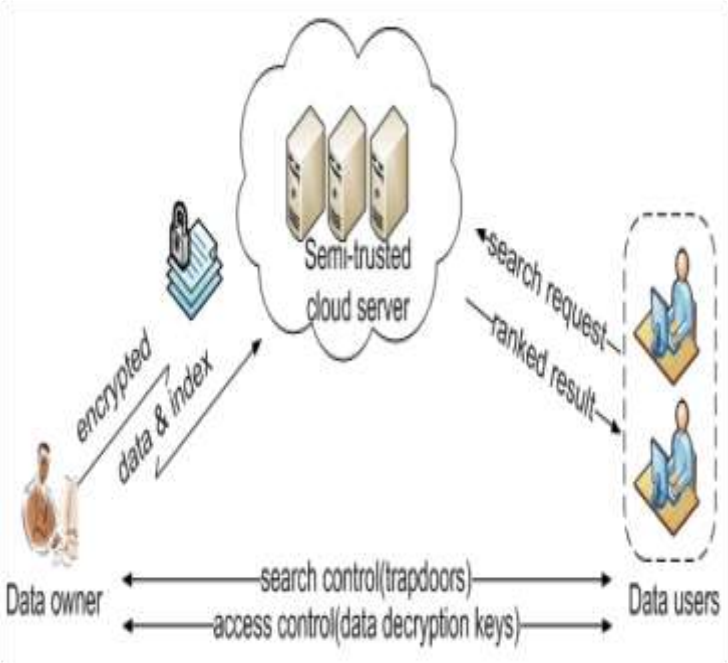

Fig1: architecture

Existing System:

The large number of data users and documents in cloud, it is crucial for the search service to allow multi-keyword query and provide result similarity ranking to meet the effective data retrieval need. The searchable encryption focuses on single keyword search or Boolean keyword search, and rarely differentiates the search results.

Disadvantage:

» Single-keyword search without ranking
» Boolean- keyword search without ranking
» Single-keyword search with ranking

Proposed System:

We define and solve the challenging problem of privacy-preserving multi-keyword ranked search over encrypted cloud data (MRSE), and establish a set of strict privacy requirements for such a secure cloud data

utilization system to become a reality. Among various multi-keyword semantics, we choose the efficient principle of "coordinate matching".

III. PROPOSED SCHEME

In this section, we give a detailed description of our scheme. We firstly propose to employ "Latent Semantic Analysis" to implement the latent semantic multi-keyword ranked search.

A. Our Scheme

Data owner wants to outsource *m* data files D □ {*d* 1 , *d* 2 , ..., *d m* } that he prepares to outsource to the cloud server in encrypted form while still keeping the capability to search

definition about *LSA*, the data owner builds a term-document matrix **A** □ . Matrix**A** □ can be decomposed into the product of three other matrices. And then, we reduce the dimensions of the original matrix **A** □ to get a new matrix **A** which is calculated the best "reduced-dimension" approximation to the original term-document matrix [16]. With *t* keywords of interest in W =as input, one binary vector **Q** is generated

In this section, we show a thorough experimental evaluation of the proposed technique on a real dataset: the *MED* dataset [17] .The whole experiment is implemented by C++ language on a computer with Core 2.83GHz Processor, on Windows 7 system. For the proposed scheme, we will reduce to separate dimensions. The performance of our method is compared with the original MRSE scheme.

B. Efficiency

The proposed scheme is depicted in details in previous section, except the KeyGen algorithm. In our scheme, we adopt Gauss-Jordan to compute the inverse matrix. The time of generating key is decided by the scale of the matrix. Besides, the proposed scheme that processed by *SVD* algorithm will consume time. Other algorithms, such as index construction, trapdoor generation, query, which is put forward by us, are consistent with the original MRSE in time-consuming.

C. Measure

In this paper, we still use the measure of traditional information retrieval. Before the introduction of the F-measure's concept, we will firstly give the brief of the precision and recall. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance[18]. F-measure that combines precision and recall is the harmonic mean of precision and recall[19]. Here, we adopt F-measure to weigh the result of our experiments.

4. Performance Analysis

For a clear comparison, our proposed scheme attains score higher than the original *MRSE*in F-measure. Since the original scheme employs exact match, it must miss some similar words which is similar with the keywords. However, our scheme can make up for this disadvantage, and retrievethe most relevant files. Figure2 shows that our method achieves remarkable result.

50 100 150 200 0.4 0.5 0.6 0.7 0.8 0.91 # of documents in the dataset F-measure LSA-MRSE Original MRSE

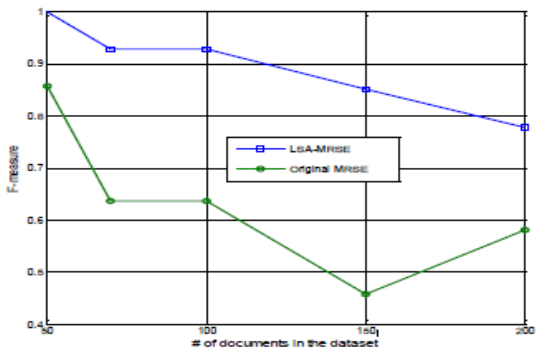

Figure 2. Comparison of Two Schemes

Figure2.Comparison of Two Schemes

Compared with the traditional vector space, the smaller the latent semantic spaceis, the more clearlythesemantic relationships. Yet, the fact is that the lower dimension will not bring the better result. For example, we will use the 100 documents of *MED*to do the test and reduce separate dimensionsrespectively.

Figure3 shows a recall-dimension curve. From the Figure3, the dimension reduces from 100 to 30, the recall has no change. It means that the relevant documents can be retrieved. Obviously,afterthe dimensions descendedto 30, the values of the recall go down .It means that some relevant documents can not be searched. Thus, when we conduct the experiments, we need to choose the appropriate dimension to achieve the best effect of experiment.
0 20 40 60 80 100 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.91 Dimensions Recall

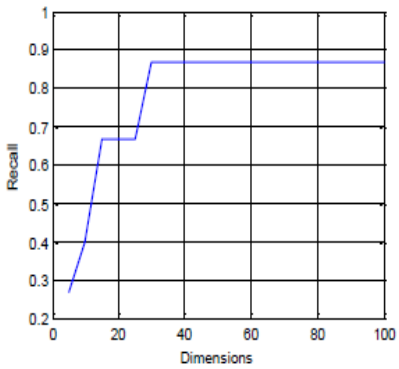

Figure 3. The Recall of Separate Dimensions

4.1 Security Analysis

We analyze our proposed scheme according to the predefined privacy requirements in design goals:International Journal of Security and Its Applications Vol.8, No.2 (2014)
330 Copyright ⓒ2014 SERSC
1) Index Confidentiality.In our proposed scheme,andare obfuscated vectors, which means the cloud server can not infer the original data vector and the query vector without the secret key *SK*. As is proven in [14], the cloud server cannot deduce *TF*values from the result relevance scores. Inother word, the index confidentiality is protected.i I w T

2) TrapdoorUnlinkability.The trapdoor of query vector is generated from a random splitting operation, whichmeans the same search requests are transformed into different query trapdoors. And thus, the query unlinkabilityis well preserved.

3) Keyword Privacy. In the known backgroundscheme, the cloud server is supposed to have more knowledge, such as the distribution *TF* values of keywords in the dataset. The cloud server is able to identify keywords by analyzing these specific distributions.In our scheme, the *TF*distributionsof keywords will be leaked directly when there is only one query keyword. Thus, our proposed scheme is designed to obscure the *TF*distributionsof keywords with the dummy values. Thatis to say, the keyword privacy is protected.

IV. CONCLUSION

In this paper, amulti-keyword ranked searchschemeover encrypted cloud datais proposed, which meanwhile supports latent semantic search. We use the vectors consisting of *TF* values as indexes to documents. These vectors constitutea matrix, from which we analyze the latent semanticassociation between terms and documents by LSA. Taking security and privacy into consideration,we employ a secure splitting*k-NN*technique to encryptthe index and the queried vector, so that we can obtain the accurate ranked results and protect the confidence of the data well. The proposed scheme could returnnot onlythe exact matchingfiles, but alsothe files including the terms latent semantically associatedto the query keyword.
As our future work, we will concentration the encrypted data of semantic keyword search in order that we can confrontwiththemore sophisticated search.

REFERENCES

[1]. M. Armbrust, "A view of cloud computing", Communications of the ACM, vol. 53, no. 4, **(2010)**, pp. 50-58.
[2]. D. Boneh, "Public key encryption with keyword search", Advances in Cryptology-Eurocrypt 2004, Springer, **(2004)**.
[3]. R. Curtmola, "Searchable symmetric encryption: improved definitions and efficient constructions", Proceedings of the 13th ACM conference on Computer and communications security, ACM, **(2006)**.
[4]. D. X. Song, D. Wagner and A. Perrig, "Practical techniques for searches on encrypted data. in Security and Privacy", 2000. S&P 2000, Proceedings 2000 IEEE Symposium, IEEE, **(2000)**.

[5]. C. Wang, "Secure ranked keyword search over encrypted cloud data", Distributed Computing Systems (ICDCS), 2010 IEEE 30th International Conference, IEEE, **(2010)**.

[6]. N. Cao, "Privacy-preserving multi-keyword ranked search over encrypted cloud data", INFOCOM, 2011 Proceedings IEEE, IEEE, **(2011)**.

[7]. W. Sun, "Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking", Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security, ACM, **(2013)**.

[8]. M. Chuah and W. Hu, "Privacy-aware bedtree based solution for fuzzy multi-keyword search over encrypted data", Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference, IEEE, **(2011)**.

[9]. S. Deshpande, "Fuzzy keyword search over encrypted data in cloud computing", World Journal of Science and Technology, vol. 2, no. 10, **(2013)**.

[10]. C. Wang, "Achieving usable and privacy-assured similarity search over outsourced cloud data", INFOCOM, 2012 Proceedings IEEE, IEEE, **(2012)**.

[11]. S. C. Deerwester, "Indexing by latent semantic analysis", JASIS, vol. 41, no. 6, **(1990)**, pp. 391-407.

[12]. S. Zerr, "Zerber+ r: Top-k retrieval from a confidential index", Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, ACM, **(2009)**.

[13]. G. W. Furnas, "Information retrieval using a singular value decomposition model of latent semantic structure", Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, **(1988)**.

[14]. W. K. Wong, "Secure kNN computation on encrypted databases", Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, ACM, **(2009)**.

[15]. C. Yang, "A Fast Privacy-Preserving Multi-keyword Search Scheme on Cloud Data", Cloud and Service Computing (CSC), 2012 International Conference, IEEE, **(2012)**.