International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 02 | Feb 2019 www.irjet.net p-ISSN: 2395-0072

An Efficient Semantic Aware Search Method over Encrypted cloud data

JASNA. K.K[1], SHABNA. M[2]

[1]PG Scholar, Dept. of CSE, Cochin college of Engineering and Technology, Kerala, India
[2]Asst. Professor, Dept. of CSE, Cochin college of Engineering and Technology, Kerala, India

---

Abstract - *The data owners out source encrypted data to cloud server to preserve security. Searching over encrypted data is too difficult. In this paper, we propose ECSED, a novel semantic search scheme based on the concept hierarchy and the semantic relationship between concepts in the encrypted datasets. ECSED uses two cloud servers. One is used to store the outsourced datasets and return the ranked results to data users and the other one is used to compute the similarity scores between the documents and the query and send the scores to the first server. To further improve the search efficiency, we utilize a tree-based index structure to organize all the document index vectors. We employ the multi keyword ranked search over encrypted cloud data as our basic frame to propose two secure schemes. The experiment results based on the real world datasets show that the scheme is more efficient and accurate.*

***Key Words*: Searchable encryption, Cloud computing, Semantic search, Concept hierarchy, Extended Concept hierarchy**

1. INTRODUCTION

Cloud computing provides many services and applications. Handling of large quantity of data is very difficult. We seek help from the cloud server for data management. So the difficulty in handling the bulk data has addressed. However, the security of the sensitive data becomes a big concern. Cloud service providers can utilize the sensitive information for their own benefits. To address the security concern of the cloud data, we need to encrypt the data and then out source it to the cloud. Searching over encryption again becomes a concern. There are many methods proposed for searching over encrypted data over cloud. All the searching method is based on keyword based search, fuzzy keyword search, multi keyword search etc. The keyword based search will not consider the semantic relation between the keywords.

1.1 Objective

Here we propose an efficient and secure search scheme based on concept hierarchy. The concept hierarchy is a tree containing keyword with meaning. So that the user will get the document as per his wish. For the security purpose, here we are using 2 cloud servers. For each documents, two index vectors are created. One index vector is used to match the similarity between the nodes in the concept hierarchy and the other one is used to rank the result of the similarity calculation.

2. RELATED WORKS

There are many schemes that provide method for secure search over encrypted data. One of such method has proposed earlier [1].This technique has a variety of advantages. They are almost secure: those schemes provide high security for encryption, therefore the servers cannot understand the plain text they only get the cipher text; they provide uniqueness in searching, that means the server cannot guess more about the plain text; they provide authorized searching, so that the server cannot search for any words without user's permission; user can also search some hidden words without revealing it to the server. The algorithms used are simple and easy to implement.

Handling of multimedia data is very difficult [2]. One technique has proposed to manage the multimedia data. In that, media data is represented with content descriptions as data type. Like a picture is provided with a description, the multimedia data is represented. Prolog is used to process the queries. It is very convenient for users to understand. The main drawback is the database complexity.

The encryption method used in the cloud is becoming an important task. Choosing the best encryption method is very crucial. Since the data privacy is of great concern, the encryption method used should be very efficient. While investigating the well know primitive cryptographic method namely public key encryption with keyword search which is very useful in many applications of cloud storage. PEKS schemes suffer from an inherent insecurity regarding the trapdoor keyword privacy, namely inside Keyword Guessing Attack (KGA).To overcome this privacy concern, an new method has introduced Dual- Server Public Key Encryption with Keyword Search (DSPEKS) and for the generic construction of DSPEKS a linear and homomorphic SPHF has introduced [3]. An efficient instantiation of the new SPHF based on the Diffie-Hellman problem is also addressed, which gives an efficient DS-PEKS scheme without pairings.

We can reduce the overhead of using public key encryption by identity based encryption [4].Identity-Based Encryption (IBE) is a public key based encryption. It reduces overhead of traditional public key based encryption. One of the main drawbacks of IBE is the private key generation (PKG) during user revocation. It needs an efficient management for Private key management. There is a paper

presented earlier aiming at tackling the critical issue of identity Revocation. In that paper, explains outsourcing computation into IBE for the first time and propose a revocable IBE. This scheme reduces the overhead of managing key generation related operations held locally. Most of the key generation processes are outsourced to a cloud server. This aim achieved by using a novel hybrid private key for each user and AND gate is involved to connect identity component and time component.

Attribute-based encryption (ABE) [5] is one of useful cryptographic primitives to realize fine-grained access control, which has been widely adopted in cloud computing.

Even the method has many advantages, the main concern is high computational overhead and weak security. Due to this drawback, many applications needed to compromise their services. The main problem is simultaneously achieving fine-grandness, high-efficiency on the data owners side, and standard data confidentiality of cloud data sharing. A technique has been proposed to addresses this challenging issue by proposing a new attribute-based data sharing scheme suitable for resource-limited mobile users in cloud computing. The scheme removes majority of the computational task. Along with that it perform a public cipher test before decryption to avoid overhead of keeping the unnecessary cipher text. To address the data security, a Chameleon hash function is used to generate an immediate ciphertext, which will be updated by the offline ciphertexts to obtain the final online ciphertexts.

There exist a scheme for predicates corresponding to the evaluation of inner products over ZN [6] (for some large integer N). This, in turn, enables constructions in which predicates correspond to the evaluation of disjunctions, polynomials, CNF/DNF formulas, thresholds, and more. Even it has some drawbacks it performs excellent in some applications.

The challenging problem of privacy preserving multikeyword ranked search over encrypted cloud data (MRSE) had been introduced [7]. Here the data owner outsources the data and the encrypted index to the cloud server. To search the document collection for t given keywords, an authorized user need to get the corresponding search trapdoors through search control mechanisms, e.g. broadcast encryption. Once receiving the search trapdoors from the data user, the cloud server searches the index and returns the result to data users. One of the main concerns here is that the search request cannot be extensible.

To improve search accuracy, the search result should be ranked by the cloud server [8]. The ranked search will provides best matching among thousands of similar words. In addition to this, to reduce the communication cost, the data user may send an optional number along with the trapdoor, so that the cloud server only sends back documents that are most relevant to the search query. Finally, the access control mechanism also included to improve the security of data.

A privacy-preserving multi-keyword text search (MTS) scheme with similarity-based ranking has been introduced [9].To support multi-keyword search and search result ranking, a new scheme is introduced to build the search index based on term frequency and the vector space model with cosine similarity measure to get higher search result accuracy. To improve the search efficiency, a tree-based index structure and various adaption methods for multi-dimensional (MD) algorithm so that the practical search efficiency is much better than that of linear search. To further enhance the search privacy, there exists two secure index schemes to meet the highest privacy requirements under strong threat models, i.e., known ciphertext model and known background model.

Searching based on keyword is not an efficient technique. The search keyword cannot extend to get the information about what the data user exactly looking for. To overcome this situation, a new search scheme is developed. This search scheme is based on fuzzy keyword search [10]. The users need not to send the exact word for searching. The data owner keeps a file containing some keywords based on the document and an index. When the user requests for a document it first check the exact matching in the keyword set. If it fails, the server will then look for the closest matching keywords and send back to the user. The major task in this method is making of fuzzy keyword sets. Wildcard based Fuzzy Set Construction could generate efficient fuzzy keyword sets. Based on the storage efficient fuzzy keyword sets, an Efficient Fuzzy Keyword Search Scheme is used. The drawback here is that the search semantics is not considering, it only checks the similar words not the meaning and relation among keywords.

The keyword based searching schemes will not consider the relation between the keywords. A keyword weighting algorithm has been introduced to take the relation among query keywords [11]. A novel central keyword semantic extension ranked scheme also developed. To better express the relevance between queries and files, the TF-IDF rule when building trapdoors and the indexes have developed.

Algorithms and search methods for actual match, partial match, and vary queries square measure conferred and applied mathematics procedures square measure given to estimate the typical and worst case retrieval times. [12].

There is a new method [13] which offers more secure and efficient search method based on concept hierarchy. The advantage of using concept hierarchy is, it will get more precise search scheme. There is a cloud server to which the encrypted data and indexes are outsourced. The data owner creates concept hierarchy based on the domain knowledge of the document. Once receiving the search

request, the cloud server searches for the result based on the concept hierarchy provided by the data owner. The main disadvantage of this method is that, upon receiving the search request, the cloud server can guess the encrypted data.

Current information-retrieval techniques either rely on an encoding process-using a certain perspective or classification scheme to describe a given item, or perform a full-text analysis, searching for user-specified words. Neither case guarantees content matching, because an encoded description might reflect only part of the content, and the mere occurrence of a word (or even a sentence) does not necessarily reflect the documents content [14].

To get the users interest while searching, earlier introduced a method; there exists an user interest model stored in the user side [15]. The user interest model is built upon the users long-term search history. It records access frequency of both query keywords and their related keywords with the help of WordNet. Different access frequency of keywords, as keyword priority, can reflect their different importance in viewpoint of the data user. To search for files of interest, the data user should firstly produce a search request. And then query reformulation that achieves keyword priority of query terms will be carried out through the user interest model.

Due to the absence of interrelation of keywords in the keyword-based representation model, some researchers make use of a set of concepts derived from predefined ontology or reference ontology to express the user profile model.

A searchable encryption means it allows data owner to outsource his data in an encrypted manner while maintaining the selectively search capability over the encrypted data[16].

Data owner has a collection of n data that he wants to outsource to the cloud server in encrypted form while still keeping the capability to search through them for effective data utilization reasons. To achieve this, data owner should encrypt the data before outsourcing it to the cloud server along with encrypted index. Make sure that the authorization between the data owner and users is appropriately done. To search the file collection by entering a keyword, an authorized user generates and submits a search request in a secret format to the cloud server. When the cloud server receives the search request, the cloud server is responsible to search the index and return the corresponding set of files to the user.

The searching is predicated on multi-keywords [17] to access the dataset, the entities are: Data owner, data user and cloud service provider. The data owner encrypts the data and outsources it into the cloud. The data owner also generates an encrypted searchable index based on the set of distinct keyword extracted from the document. In the search stage, system will generate an encrypted search query and a parameter k .The cloud server will search the index I and return top k most relevant documents to the user based on similarity values. Here, the key distribution is out of scope of this paper.

A methodology[13] for constructing an inspiration hierarchy that represents this content of assortment of user mere documents that distinguishes them from the remainder of the documents in an exceedingly complete collection. The development takes place in 3 completely different phases that regressively tackle the 3 term dependence dimensions. In document redundancy term coefficient is applied to spot those out of the distinctive terms within the user mere documents that area unit most specific to those documents.

The extracted terms area unit used, because the idea hierarchies building blocks. A novel term coefficient methodology known as Relative Document Frequency (RelDF) that measures the relative importance of terms at intervals the user mere documents and a general assortment of documents.

Some document categorization algorithms might be adopted for information categorization [18]. Algorithms that take into thought the special characteristics of databases could also be simpler. We have a tendency to gift 3 ways for assignment databases to ideas within the construct hierarchy.

Wordnet [19] is lexical database. It teams words along supported their meanings. WordNet will so be seen as a mixture of lexicon and synonym finder. Words that area unit found in shut proximity to one another within the network area unit semantically disambiguated. It labels the linguistics relations among words, whereas the groupings of words in an exceedingly synonym finder don't follow any express pattern aside from that means similarity. Synonyms–words that denote identical construct and area unit interchangeable in several contexts–are sorted into unordered sets (synsets). Every of WordNets 117000 synsets is joined to different synsets by suggests that of a little range of conceptual relations. To boot, a set contains a quick definition (gloss) and, in most cases, one or additional short sentences illustrating the employment of the set members. Word forms with many distinct meanings area unit delineated in as several distinct synsets. Thus, every form-meaning try in WordNet is exclusive. While it's accessible to human users via an online browser, its primary use is in automatic text analysis and computing applications. The info and software package tools are free underneath a BSD vogue license and area unit freely on the market for transfer from the WordNet website.

To measure the semantic distance in wordNet, there are two algorithms, HS (Hierarchy Spread) and BDOS (Bi-Direction One Step), to search the relation with shortest semantic distance in WordNet [20]. Algorithms in searching Directed Graphs have been applied in searching semantic relations in WordNet. Rada and Resnik measure similarity

with hierarchy in WordNet. The HS algorithm is not accurate enough. In this algorithm; we could just find a path with simple patterns or follows a pattern of part of then is-a then part-of. In another word, if the path is follow additional part of or is-a or more complicated, it will be ignored and result that the path we found is not the shortest one, or even fail to find. The algorithm BDOS does not pay attention on type of relation. Connections between two nodes are all considered as one step, no matter it is is-a, or part of. But one disadvantage now is that, if the path between two start nodes is a long enough, memory resources will be consumed constantly till used up. So step control is necessary.

3. CONCLUSION

In this paper, to address the problem of semantic retrieval, a new efficient and secure scheme is proposed. This scheme uses two cloud servers for reducing the time complexity. To make the system more accurate, it uses extended concept hierarchy and two index vectors are created for each document. The security analysis shows that the proposed scheme is secure in the threat models.

REFERENCES

1. 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 on. IEEE, 2000, pp. 44–55.

2. V. Y. Lum and K. Meyer-Wegener, "An architecture for a multimedia database management system supporting content search," in International Conference on Computing and Information. Springer, 1990,pp. 304 - 313.

3. R. Chen, Y. Mu, G. Yang, F. Guo, and X. Wang, "Dual-server publickey encryption with keyword search for secure cloud storage," IEEE transactions on information forensics and security, vol. 11, no. 4, pp. 789–798, 2016.

4. J. Li, J. Li, X. Chen, C. Jia, and W. Lou, "Identity-based encryption with outsourced revocation in cloud computing," Ieee Transactions on computers, vol. 64, no. 2, pp. 425–437, 2015.

5. J. Li, Y. Zhang, X. Chen, and Y. Xiang, "Secure attribute-based data sharing for resource-limited users in cloud computing," Computers & Security, vol. 72, pp. 1–12, 2018.

6. J. Katz, A. Sahai, and B. Waters, "Predicate encryption supporting disjunctions, polynomial equations, and inner products," in Annual International Conference on the Theory and Applications of Cryptographic Techniques. Springer, 2008, pp. 146–162.

7. Z. Fu, X. Sun, Q. Liu, L. Zhou, and J. Shu, "Achieving efficient cloud search services: multi-keyword ranked search over encrypted cloud data supporting parallel computing," IEICE Transactions on Communications, vol. 98, no. 1, pp. 190–200, 2015.

8. N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," IEEE Transactions on parallel and distributed systems, vol. 25, no. 1, pp. 222– 233, 2014.

9. W. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li, "Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking," in Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security. ACM, 2013, pp. 71–82.

10. J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy keyword search over encrypted data in cloud computing," in Infocom, 2010 proceedings ieee. IEEE, 2010, pp. 1–5.

11. Z. Fu, X. Wu, Q. Wang, and K. Ren, "Enabling central keyword based semantic extension search over encrypted outsourced data," IEEE Transactions on Information Forensics and Security, vol. 12, no. 12, pp. 2986–2997, 2017.

12. P. Scheuermann and M. Ouksel, "Multidimensional b-trees for associative searching in database systems," Information systems, vol. 7, no. 2, pp. 123–137, 1982.

13. N. Nanas, V. Uren, and A. De Roeck, "Building and applying a concept hierarchy representation of a user profile," in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003, pp. 198–204.

14. N. Guarino, C. Masolo, and G. Vetere, "Ontoseek: Content-based access to the web," IEEE Intelligent Systems and their Applications, vol. 14, no. 3, pp. 70–80, 1999.

15. Z. Fu, K. Ren, J. Shu, X. Sun, and F. Huang, "Enabling personalized search over encrypted outsourced data with efficiency improvement," IEEE transactions on parallel and distributed systems, vol. 27, no. 9, pp. 2546–2559, 2016.

16. C. Wang, N. Cao, K. Ren, and W. Lou, "Enabling secure and efficient ranked keyword search over outsourced cloud data," IEEE Transactions on parallel and distributed systems, vol. 23, no. 8, pp. 1467–1479, 2012.

17. Z. Fu, X. Sun, S. Ji, and G. Xie, "Towards efficient content-aware search over encrypted outsourced data in cloud," in Computer communications, IEEE INFOCOM 2016-the 35th annual IEEE international conference on. IEEE, 2016, pp. 1–9.

18. W. Wang, W. Meng, and C. Yu, "Concept hierarchy based text database categorization in a meta search engine environment," in Web Information Systems Engineering, 2000. Proceedings of the First International Conference on, vol. 1. IEEE, 2000, pp. 283–290.

19. G. A. Miller, "Wordnet: a lexical database for english," Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

20. D. Chen, Y. Jianzhuo, F. Liying, and S. Bin, "Measure semantic distance in wordnet based on directed graph search," in E-Learning, EBusiness, Enterprise Information Systems, and E-Government, 2009. EEEE'09. International Conference on. IEEE, 2009, pp. 57–60.R.