An Optimized and Secured Ranking Approach for Retrieving Cloud Data Using Keyword Search

Vineetha Viswan[1], Adarsh Sunil[2]
[1,2]*Dept. of Computer Science and Engineering, Sree Buddha College of Engineering Pattoor P.O, Alappuzha*

Abstract: *Cloud computing is a versatile technology that emerged as a solution to reduce costs in organizations by providing on-demand high quality applications and services from a centralized pool of configurable computing resources. With the advantage of storage as a service data owners are motivated to outsource their valuable and sensitive data from local sites to commercial public cloud since it costs less, easily scalable and can be accessed from anywhere anytime. Sensitive data that moves into and across cloud infrastructures increases the risk of data loss or compliance violations. However to protect data privacy, sensitive cloud data have to be encrypted before outsourcing. Data encryption makes effective data utilization a very challenging task. In this paper we define an optimized and secured semantic based ranked keyword search over encrypted cloud data. Existing search schemes are not much efficient as they entirely depend on the submitted query keyword and did not consider the keyword semantics. Ranking approach in the proposed scheme not only consider the keyword semantics but also deals with examining the different forms of a queried keyword for retrieving the most matching files relevant to the user request. This type of ranked search enhances system usability by computing relevance score and ensures security guarantee by using an order preserving symmetric encryption technique to protect those sensitive score information. Here user can retrieve more files from cloud server with less communication overhead.*

Keywords: *Cloud computing, keyword semantics, Ranking approach, Security*

I. Introduction

In the current information era, large amount of data being produced by various organizations and individuals which in turn increases the need to store and utilize those data for future purpose. The cost for building and maintaining such data store will be very high. Moreover the burden involved in storage management will be unpredictable. This increases the demand for shifting such complex data management systems from local machines to commercial public cloud. Cloud computing is a long dreamed vision of computing as a utility that allows individuals and business to work online rather than on a PC.

Cloud provides storage as a service that offers the cloud customers to outsource their valuable data into the cloud server and retrieve them whenever required. As we know that internet is a network of networks that provides hardware or software infrastructure in establishing and maintaining connectivity of the computers around the world. While cloud computing is a technology which delivers different types of resources over the internet. Therefore cloud computing could be identified as a technology that uses internet as the communication medium to deliver its services. This new class of internet technology reduces the burden involved in storage management, eliminates the expenditure on software and hardware maintenances, and provides universal data access irrespective of geographical locations.

As this type of countless benefits offered by the cloud computing model, the data owners force themselves to outsource their sensitive information such as personal data, finance data, confidential government data, and e-mails etc in to the cloud. But the fact that data owners and cloud server are no longer in the same trusted domain may leaves this outsourced unencrypted data at risk. There are many situations in which the cloud server can leak the sensitive information to unauthorized entities and may put this data vulnerable to threats. So it necessitates the need to encrypt the data before outsourcing in order to achieve data privacy. Unfortunately, data encryption makes effective data utilization a very challenging task as there could be a large number of outsourced data files.

For avoiding unauthorized user access the data owners may share their outsourced data with a large number of users, who might be interested in retrieving certain specific data files during a given session. One of the most popular approach to do so is to perform keyword based search. Keyword based searching technique has been widely used in plaintext search scenarios and which helps the users to selectively retrieve files of interest.

Data encryption restricts the user's ability to perform keyword search which in turn demands the keyword privacy, makes the traditional plaintext search methods fail for encrypted cloud data. Many of the existing searchable scheme will only support Boolean keyword search based on Boolean logic that combines words and phrases using logical operators AND, OR, NOT. This search scenario will suffer from the following drawbacks. The search result may contain too many items or files that may contain the word or phrase but are not relevant to our search. Therefore sending back all files based on presence or absence of the keyword further incurs unnecessary network traffic. Misspelling a keyword can often produce no results and the searching database may not provide an alternate spelling. On the other hand, users without having any knowledge about the encrypted cloud data have to go through every retrieved file for finding their most matching ones which increases the post-processing overhead.

To enhance system usability and search flexibility, some research has been done on fuzzy keyword search and the corresponding solution support tolerance of minor typos and format inconsistencies. Some further researches focus on search efficiency and some on secure dynamic updating. But they only support exact keyword search. Many of the existing schemes mainly takes the structure of terms into consideration and use edit distance to evaluate the similarity. Therefore existing search mechanisms are not much efficient as they entirely depend on the submitted query keyword and did not consider the keyword semantics. In short, lacking of effective mechanism to ensure file retrieval accuracy is a major drawback of existing systems.

In this paper, we propose an optimized and secured semantic based ranking approach for retrieving the most matching files over encrypted cloud. Ranking approach in the proposed scheme combines the advantages of both keyword and semantic based search for producing meaningful search results. This scheme not only consider the keyword semantics but also takes different forms of queried keyword into consideration for satisfying the users request.

II. Related Works

Privacy preserving documents that contains sensitive informations are generally kept inside data centers in an encrypted form. Encryption is performed to limit their access only to the authorized users. Providing the authorized users with efficient search capability over encrypted documents is a challenge for the data owners. For that a confidentiality preserving baseline model is developed in [1]. In [1], the main drawback is that the data center gains information about the retrieval log. Retrieval log maintains the information about which user searched for what encrypted queries, when and how often. Based on such information, the data center might perform statistical attacks.

In [2] a ranked searchable encryption scheme which helps the users to perform secure search over encrypted data using keywords is proposed. Here ranked search enhances system usability by enabling search result relevance ranking instead of sending undifferentiated results and improves file retrieval accuracy. But here efficiency of data retrieval is preferred rather than security. Also authorization of users is not mentioned.

In [3] a similarity search technique based on keyword similarity is proposed. Inorder to achieve effective system usability, this technique uses edit distance as a similarity metric. They didn't consider the terms semantically related to query keyword, thus many related files are omitted.

In [4], an eight-way orthogonal categorization is proposed which allows similarity measures to be specified as points in the eight-space and thereby permits the space of similarity measures to be explored in a systematic manner. The most striking feature of [4] is its poor performance. This is the main disadvantage of this paper. That is this approach is not successful in identifying the effective combination of similarity measures for retrieving the most matching files. All of these works discussed above have their own positives and negatives. The main objective of the proposed system is to develop a cloud server environment that enables file retrieval accuracy by ensuring as-strong-as-possible security guarantee along with less communication and computation overhead.

III. Proposed System

The proposed system aims at developing an optimized and secured ranking framework over encrypted cloud that supports efficient file retrieval using semantic based keyword search. Specifically the system has the following design goals:

1) Semantic based ranked keyword search: existing solutions entirely concentrated on the submitted query keyword without considering the keyword semantics, so inorder to make the search scheme more intelligent keyword semantics is considered and ranking is performed;

2) Security guarantee: to prevent the cloud server from learning additional information about the plaintext of either the data files or the keywords since the cloud server is curious to infer and analyze the search requests and results;

3) Efficiency: to optimize the overall performance of the system with minimum communication and computation overhead.

The proposed system mainly consists of three entities:

1) Data owner: Data owner is the owner of data files. He wants to outsource this data file collection into the cloud server. For protecting data privacy, files must be encrypted before outsourcing to cloud. First data owner performs semantic analysis which finds out the semantic relationship between extracted keywords and then performs stemming operation. A searchable index is built using the unencrypted data files for enabling the cloud

server to perform keyword search over the encrypted file collection. Finally the data owner outsources both encrypted files and index to the cloud server.

2) Data user: Data users are the users of the data. The data owner will share their outsourced data with a large number of users. Inorder to access the outsourced data the data user must be registered. Only after registration, the data user can send search request to cloud by entering a keyword.

3) Cloud server: After receiving a search request from an authorized user, the cloud server is responsible to search the index and returns the corresponding set of encrypted matched files in a ranked order to the user.

Following figure Fig. 1 shows the proposed system architecture that helps an authorized user to perform an optimized and secured semantic based ranked keyword search over encrypted cloud data for retrieving his interested file.

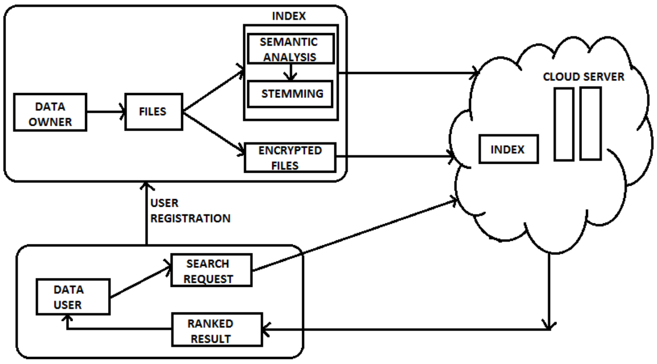

Fig. 1.System architecture

The proposed system is organized into four subsections which describes the various steps involved in retrieving the most matching files based on users interest. Following are the four subsections.

A. Semantic Analysis and Stemming

The data owner has a set of files. He first extracts certain distinct keywords from the file collection. Data owner then performs semantic analysis ie it is a process of finding semantically related words synonyms with the help of WordNet. Synonyms are the words or phrases with the same or similar meaning. Inorder to improve the search accuracy, keywords extracted from the file collection is to be extended by common synonyms as the cloud users searching input may be the synonyms of predefined keywords. Sometimes the cloud users are not able to use the exact keywords during search process due to lack of exact knowledge about the data. But with the help of synonyms the requested user will get satisfied with the exact file data. As a result of semantic analysis a synset is created with the help of WordNet. Set of synonyms are collectively called as synset.

WordNet is a large lexical database for the English language. WordNet groups English words into synset. It also provides short definitions and usage examples and keeps track of a number of relations among the synonym sets or their members. WordNet can be viewed as a combination of dictionary and thesaurus. WordNet is freely and publicly available for download from the WordNet website. WordNet can be used as a useful tool for natural language processing and for computational linguistics.

The proposed system not only deals with keyword semantics but also consider the different forms of keywords for better performance. So for obtaining different forms of a keyword the system uses stemming operation. Stemming is a technique used in information retrieval for finding the morphological variants of search terms. Stemming deals with the process of finding the word stem, base or root form. The proposed system uses Porter Stemmer algorithm. The synset and the keyword is stemmed using this Porter Stemmer algorithm.

Porter Stemmer algorithm is one of the most popular stemming method proposed in 1980. This algorithm follows the idea that suffixes in English language are mostly made up of a combination of smaller and simpler suffixes. Porter Stemmer provides an excellent trade-off between speed, readability and accuracy. Here the stem uses a set of rules or transformations applied in a succession of steps. It is mainly consists of six steps and within each step rules are applied until one of them passes the conditions. About sixty rules are used in six steps. Following are the steps used in Porter Stemmer.

1) Gets rid of plurals and -ed or -ing suffixes.
2) Turns terminal y to i when there is another vowel in the stem.
3) Maps double suffixes to single ones: -ization, -ational, -iveness etc.

4) Deals with suffixes, -full, -ness etc.
5) Takes off -ant, -ence, -ive etc.
6) Removes a final –e .

Finally the system generates a set of semantically related and a set of different forms of keyword along with the predefined keywords.

B. Index Creation

A secure and searchable index is created using the unencrypted file collection before outsourcing to the cloud server. The index structure generally deals with a list of mappings from keywords to the corresponding set of files that contain this keyword. We can use some notations to represent the index creation section. Let C = (F1,F2,...,Fn) be the set of n files that is to be outsourced and W= (w1,w2,...,wm) be the set of m words. F(w_i) represents the set of identifiers of files in C that contain keyword w_i. id(Fj) denotes the identifier of file Fj which helps to locate the actual file. Let I represents the index built from the set of files, including a set of posting lists I(w_i). N_i is the number of files containing the keyword w_i where N_i = | F(w_i)| and let v represents the maximum number of files containing the keyword w_i.

The index creation scheme is as follows:

1) Generate random keys x, y and z.
2) For each w_i belongs to the keyword set W, build F(w_i).
3) For each w_i Є W and for 1≤ j ≤ | F(w_i) |, first calculate the ranking score Sij for file Fij using the TF X IDF rule explained later and then compute Ɛz (Sij) and store it with Fij ’s identifier ie (id(Fij) || Ɛz (Sij)) in the index posting list I(w_i). ‘Ɛ’ is an order preserving symmetric encryption scheme.
4) For each I(w_i) where 1 ≤ i ≤ m, first encrypt all N_i entries with l ’padding 0’s using key fy(w_i) where 1 ≤ j ≤ v ie ($0l'$ || id(Fij) ||Ɛz (Sij)). Then set remaining v - N_i entries, if any, to random values of the same size as the existing N_i entries of I(w_i). As a final step replace w_i with π_χ (w_i). The l ’padding 0’s is used to indicate the valid posting entry. Here f be a pseudorandom function and π be a hash function like SHA-1.
5) Output I.

A ranking function is used to calculate the rank of the files. We are using TF X IDF rule for finding the ranking score where TF is the term frequency and IDF is the inverse document frequency. Term frequency is the number of times a given term or keyword appears within a file whereas inverse document frequency is obtained by dividing the total number of files by the number of files containing the term. Following equation is used to calculate the score. Score (Q ,F_d) $=\sum_{t \in Q} \frac{1}{|F_d|} . \left(1 + \ln f_{d,t}\right) . \ln \left(1 + \right.$ Nft .

Here Q denotes the keywords searched; | Fd | is the length of the file Fd; $f_{d,t}$ denotes the TF of term t in file F_d; f_t denotes the number of files that contain term t; N denotes the total number of files in the collection. Finally inorder to protect data privacy, the data owner encrypts the whole file collection using AES symmetric encryption. Then both the index and encrypted files are outsourced to the cloud.

C. User Registration

This module provides registration facility to a new user. Data owner can add users. Only a registered user can access the encrypted cloud data. For the registration purpose user has to enter a valid email id. When a new user registers to the system a registration mail containing registration details are sent to the user specified email id. Registration details includes user id, password and a registration code.

User can activate their account using the registration code. For that the user must possess an android device. User can use an android based mobile device so that the system will work like a mobile application. After successful registration an authorized user can access the cloud data whenever he required.

D. Search and Result Management

First the user must login to the system to avail cloud services. After activating the account, the user can login to the system from his android device with the help of username and password sent via email. For retrieving his files of interest, an authorized user can send a request to the cloud by specifying the keyword. User also has the right to specify the number of files for retrieval ie top k relevant files at the time of sending search request. This will reduce the bandwidth.

After receiving the user request cloud server first searches the index and locates the matching list of index using π_χ(w) and then uses fy(w) to decrypt the entries. Cloud server now sees a list of matched file ID’s and their corresponding encrypted scores. Based on the relevance score topmost k matched files that are encrypted are sent back to the user. Files can be downloaded and decrypted at the user end using AES decryption algorithm.

IV. Advantages

Cloud computing becomes a fascinating topic now a days. Many enterprises are interested in it due to the countless benefits offered by this platform. In business point of view, it is smarter to rent than to buy. As a result everything now resides under the cloud. So the next question is how it is possible to retrieve those data's from the cloud effectively without losing its privacy. This is the main challenge faced by the data owners. Our proposed system stands with a better solution to this. The system provides keyword privacy and data confidentiality by using secured encryption schemes. Less number of communication rounds is needed between the users and cloud server for each search and retrieval process. Here the authorization of data users is properly done ie only after getting the data owner permission, an authorized user can access the cloud data. The system offers top k retrieval facility which reduces the bandwidth. Retrieval accuracy is improved in such a way that an authorized user without any preknowledge about the encrypted cloud data can also retrieve the relevant data which in turn reduces the post processing overhead.

V. Application

With the advent of cloud computing and due to the rapid increase in the number of mobile users, the advantage of mobility can be achieved by integrating cloud computing to mobile world. By doing this an authorized user can get access to real time data whenever and wherever he wants. Moreover the data can be accessed by multiple users simultaneously. By taking these advantages in mind, several applications like mobile health care system, mobile gaming, mobile commerce etc can be designed. For achieving stronger client relationships, financial institutions like banking industry can also take the advantages offered by the system.

VI. Conclusion

Usually due to the lack of exact knowledge about the data, cloud users searching input may be the synonyms of predefined keywords. So by combining the benefits of both keyword and semantic based search meaningful search results can be achieved. Due to the security strength of the encryption schemes file content is well protected. Semantic based ranking approach requires less communication and computation overhead which improves search efficiency.

Acknowledgment

I am grateful to the almighty for the good health and wellbeing that were necessary to complete this work. I wish to express my sincere thanks to Mr. Adarsh Sunil for the valuable guidance and constant supervision as well as for the support offered to me in completing the work. I am also grateful to Mr. Gopu Darsan for sharing expertise, and sincere guidance and encouragement extended to me. I also express my sincere gratitude to all the Department faculty members for their help and support. I also thanks my parents for their encouragement, support and attention. I am also grateful to my partner who supported me through out this venture.

References

[1] A. Swaminathan, Y. Mao, G.-M. Su, H. Gou, A.L. Varna, S. He, M.Wu, and D.W.Oard, "Confidentiality-Preserving Rank-Ordered Search," Proc. Workshop Storage Security and Survivability, 2007.

[2] Cong Wang, Ning Cao, Kui Ren and Wenjing Lou,"Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data", IEEE Transactions on Parallel and Distributes Systems, Vol.23, NO.8, AUGUST 2012).

[3] C. Wang, K. Ren, S. Yu, K. Mahendra, and R. Urs, "Achieving Usable and Privacy-Assured Similarity Search over Outsourced Cloud Data," Proc. IEEE INFOCOM, 2012.

[4] J. Zobel and A. Moffat, "Exploring the Similarity Space," SIGIR Forum, vol. 32, no. 1, pp. 18-34, 1998.

[5] Jaspreet Kaur and Manish Mahajan, "A Review on Semantic Multi-Keyword Ranked Search Techniques over Encrypted Cloud Data".

[6] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure Ranked Keyword Search over Encrypted Cloud Data," Proc. IEEE 30th Int'l Conf. Distributed Computing Systems (ICDCS '10), 2010.

[7] P. Mell and T. Grance, "Draft Nist Working Definition of Cloud Computing,"http://csrc.nist.gov/groups/SNS/cloud computing/index.html, Jan. 2010.

[8] Zhihua Xia, Yanling Zhu, Xingming Sun and Lihong Chen"Secure semantic expansion based search over encrypted cloud data supporting similarity Ranking, " Journal of Cloud Computing: Advances, Systems and Applications 2014.

[9] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report UCB-EECS-2009-28, Univ. of California,Berkeley, Feb. 2009.

[10] Cloud Security Alliance "Security Guidance for Critical Areas of Focus in Cloud Computing," http://www.cloudsecurityalliance. org, 2009.