Volume 3, Issue IX, Sep 2015
ISSN 2320-6802

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

Improving Accuracy with Ontology based Secured Search in Encrypted Cloud Data

Bismi M[1], Sulphikar A[2]

[1]PG Scholar, LBSITW, University of Kerala,
Thiruvananthapuram 695012, India
bismim90@gmail.com
[2]Associate Professor, LBSITW, University of Kerala,
Thiruvananthapuram 695012, India
sulphis@gmail.com

***Abstract**: Cloud computing is the new and emerging technology after grid computing in which resources are provided online. This facility reduces the hardware and software costs and users are allowed to store their sensitive data in the mass cloud storage. While storing our data in cloud, security must be ensured. So the data had to be encrypted with searchable symmetric encryption methods before storing in cloud. Users can retrieve their files later through keyword search. For retrieving more number of accurate files, an ontology based keyword search mechanism is introduced in this paper. Ontology helps to find out the meaning of the given keywords. Coordinate matching is used to find out the similarity between cloud data and given query. Experimental evaluation gives the probabilistic values for precision and recall. Accuracy is seen to be increased than that of existing system.*

***Keywords:** searchable encryption, ontology, keyword search, encrypted cloud data.*

1. INTRODUCTION

Cloud computing is an emerging technology that alters the way we use our computer and Internet. Instead of running programs and data on an individual system, everything is hosted in the "cloud". It provides data storage mechanism to access the data from anywhere, everywhere and at any time. According to NIST definition, Cloud computing can be defined as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources which can be rapidly provisioned and released with minimal management effort or service provider interaction[1]. The mass storage and attractive advantages provided by cloud motivates more and more users to store their data in the cloud. But here the sensitive data of users are stored in the premises of a third party. So security will be an issue. The data should be secured enough with confidentiality, integrity and availability. Figure 1 shows how different users are connected to the cloud from their own personal computers. For providing more security, we can make use of some cryptographic techniques like encryption. Searchable encryption is a scheme that provides a way to encrypt a search index so that its contents are hidden except to a party that is given appropriate tokens [2]. An authorized user can later search over the encrypted data for effective data utilization. A virtual private storage service based on cryptographic techniques achieves both the security of a private cloud and functionality and cost savings of a public cloud. Other advantages of cryptographic cloud storage are the control of data is in the hands of customer and security properties are taken from cryptography. This paper aims to provide data security in cloud with the help of encryption and ontology based secure search mechanism by users for efficient and accurate retrieval of files. Ontology based search gives the meanings of keywords provided and gives more relevant and accurate results.

The paper is organized into following sections. Section 2 gives the background study which is a summary of works related to the paper. Then comes the problem definition, system architecture and proposed method. Section 4 provides the implementation details and test results. Final section consists of conclusion and future work followed by acknowledgement and references.

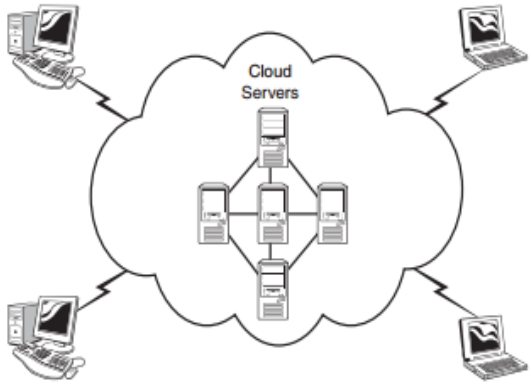

Figure 1: Users interaction with the cloud

2. LITERATURE REVIEW

Data can be stored either as public or private in the cloud. For ensuring security, the confidential data are stored in the cloud using encryption technique and only the authenticated members who know the key can access the data. Later the files can be accessed through keyword search. Traditional keyword search was based on plaintext. The traditional keywords are mainly of two types-single keyword or Boolean keyword. Boolean expression queries are composed of conjunction, disjunction and negation of keywords. But for protecting data privacy, certain cryptographic techniques are to be applied in

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

the original data. Here searchable encryption is used and so the sensitive data should be encrypted before outsourcing. Here comes the importance of encrypted cloud data search. Keyword based search helps to selectively retrieve the files of interest instead of retrieving all the encrypted files back.

The basics for this come from the field of information retrieval [3]. The three most commonly used information retrieval (IR) models are vector space model (VSM), probabilistic model and inference network model. In VSM, text is represented by a vector of terms. VSM assigns a numeric score to a document for a query and the model finds out the similarity between query vector and the document vector by taking the inner product between two vectors. Probabilistic models make use of the probabilistic ranking principle (PRP). i.e., documents in a collection should be ranked by decreasing probability of their relevance to a query. Probability of presence/absence of a term in relevant/non-relevant documents can be found out. In inference network model, the document instantiates a term with certain strength, and the credit from multiple terms is accumulated to compute the numeric score for document. Term frequency, document frequency and document length are the three factors involved in final term weight calculation.

2.1. Basics of Searchable Encryption (SE)

Searchable encryption allows us to search in an index and thus the contents of original file are hidden and the contents can be viewed only by a person provided with necessary tokens. Index is also encrypted for security and necessary tokens are provided for each keyword. SE schemes are mainly of two types- symmetric searchable encryption (SSE) and asymmetric searchable encryption (ASE)[2]. In SSE, there will be a single private key for both encryption and decryption. It is applicable in those places where the person who searches over the data is also the same who generates it. Advantages of SSE includes efficiency and security where as the main disadvantage is functionality. In ASE, there will be two keys- the public and private keys. ASE schemes are applicable in places where the person searching over the data is different from the person that generates it. Advantages include functionality and disadvantage is security will be weak and low efficiency. Apart from SSE, other methods used are Property preserving encryption (PPE), functional encryption, fully - homomorphic encryption etc.:-[4]. PPE schemes encrypt messages using some of the properties of the message. The simplest form PPE is deterministic encryption which always encrypts the same message to give the same cipher text. Functional encryption uses a public-private key pair in which giving a secret key allows one to learn a function of what the cipher text is encrypting. A homomorphic encryption (HE) scheme encrypts data such that computations can be performed on the encrypted data without knowing the secret key. Operations like addition, multiplication, XOR can be used here. Thus, a fully-homomorphic encryption scheme allows any computation on encrypted data.

The various methods used in the previous systems for searching in encrypted data includes single keyword SE[5], Public Key Encryption with Keyword Search(PEKS)[6], Boolean Symmetric searchable Encryption[7], Fuzzy Keyword Search[8], Ranked searchable symmetric encryption(RSSE)[9], Multi keyword ranked search (MRSE)[11] etc.:-. The techniques used, pros and cons of the existing systems are given in table 1.

Table 1: Searching techniques in encrypted cloud data

Paper	Techniques used	Pros	Cons
Symmetric searchable encryption[5]	Non adaptive setting and adaptive adversary	Efficient storage and access of sparse tables	Does not provide accurate documents
PEKS[6]	Decision diffie-Hellman assumption and trapdoor permutation	Secure channel between sender and user	Supports only single query
Boolean SSE[7]	Gram Schmildt orthogonalization process	Randomized and linear search	Only for searching Boolean queries and longer computation phase
Ranked keyword search[9]	Ranking and order preserving mapping	Avoids network traffic and highly efficient	Do not supports multiple keywords and increased search time and cost
Multi-keyword ranked search[11]	Coordinate matching and inner product similarity	Low overhead in computation and improved search accuracy	Single Boolean keywords with ranking not possible

3. PROPOSED SYSTEM

3.1. Objective

By analyzing the existing systems, it is understood that none of the existing keyword searches support multiple keywords. So to improve the accuracy of search results and for obtaining more relevant files, it is worth sufficient to allow multiple keywords in the query request along with ranked search. Each keyword in the search request helps to narrow down the search results. For providing more accuracy, an ontology based search with multiple keywords is proposed here. Ontology

Volume 3, Issue IX, Sep 2015
ISSN 2320-6802

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

based search will give meaningful words related to the search query.

3.2. Architecture for the proposed method

Figure 2 shows the system architecture. This system has 3 entities: a data owner, cloud server and the data user. The data owner (individual or an enterprise) has a collection of documents DC to be outsourced to the cloud server in encrypted form C. Before uploading the data collection, data owner builds an encrypted searchable index I based on a set of distinct keywords W extracted from DC. Then both the encrypted index I and the encrypted document collection C are outsourced to the cloud server. When the data user wanted to search for an interested file, he gives keywords as request for searching the matching files. Each keyword is taken and finds out the synonyms of the given keyword. Now the files containing both the query words and their meanings are retrieved first and should be ranked with relevance score.

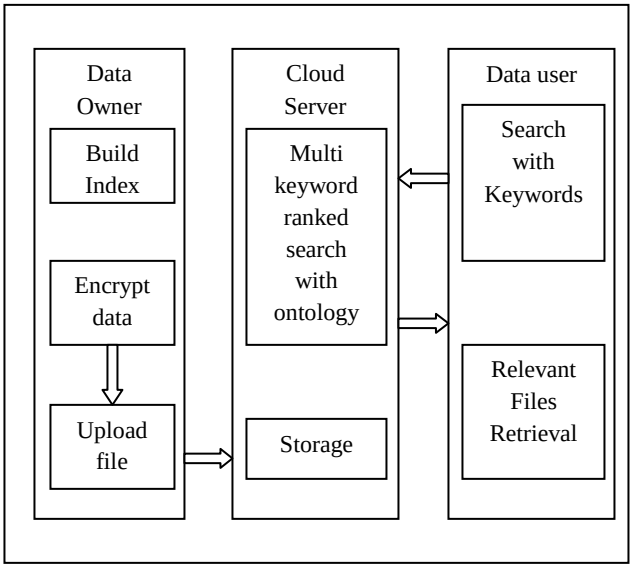

Figure 2: System Architecture

Indexing and synonym expansion are the main steps here. Index files are built for each document in the document collection. Indexing is a technique used to locate a term in a text document. This is done by using preprocessing and stemming techniques [10]. In stemming process, words are reduced to their morphological roots, i.e., suffixes and modifiers are removed. For example, compression, compressed and compressor, all these words are indexed as the word "compress", which is called as the common root. Preprocessing is the process of omission of stop words. Thus index files are created for each document and these index files are to be encrypted for index privacy. Index details are stored in the index table with unique id, index term in encrypted form term frequency and the corresponding file name for the index file created. This is shown in table 2. Term frequency gives the number of times a term appearing in a document.

Table 2: Index table for documents

Id	Index Term	Term Frequency	Filename
1	photo	2	allen.txt
2	network	1	neal.txt
3	machine	4	allen.txt
4	service	5	miller.txt
5	Liquid	3	arnold.txt

An authorized user can give multiple keywords as search query. The next process is to expand the keyword set with the meanings of each of the keywords given. This leads to retrieve more number of accurate files. Meanings or synonyms of the keywords are obtained with the help of Word Net Dictionary Class. The keywords and their obtained meanings are stored in a data structure known as data table.

3.3. Similarity Matching and Score Calculation

Coordinate matching technique is used to find out the similarity between query keywords and documents [11]. This technique counts the number of query keywords appearing in a document to quantify the relevance of that document to the query. The more query keywords that appear in a document, the more relevant the document to the query. Vector Space Model (VSM) is used to build document index.ie; each document is expressed as a vector where each dimension value will be the term frequency weight with the corresponding keyword. A new vector is generated in the query side with the dimension value as inverse document frequency (IDF) weight. Now similarity of one document to the search query can be found out by using inner product or cosine similarity measure. Scoring is a usual way to weight the relevance and ranking function is commonly used to evaluate relevant scores of matching files to a request. Here the TF - IDF rule is used [12]. The attributes are:

- Term Frequency (TF): gives the occurrence of a particular term in a document. It measures the importance of the term within the particular file.
- Inverse Document Frequency (IDF): gives the importance of a term in the whole document collection. It is calculated by dividing the total number of documents by the number of files containing the particular term.

This TF-IDF weighting rule is used to calculate the similarity between the query and document vectors by taking the relevant score between them.

The terms needed for similarity calculation:

- fd,j gives the TF of a keyword wj in a document d
- fj gives the total number of documents with the keyword wj
- M gives the total number of files in document collection
- N gives total number of keywords used for searching
- Wd,j gives the TF calculated from fd,j
- Wq,j gives the IDF calculated from N and fj

Volume 3, Issue IX, Sep 2015
ISSN 2320-6802

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

The formula for score calculation is:

$$Score(Q, Dd) = \frac{\sum_{j=1}^{N} wq.j.wd.j}{\sqrt{\sum_{j=1}^{N} W(wq, j)2.}} \cdot \frac{1}{\sqrt{\sum_{j=1}^{N} W(wd, j)2}} \quad \ldots\ldots(1)$$

where $w_{q,j} = 1+ln\, fd.j$ and $w_{q,j} = ln\,(1 + N/fj)$

Term frequency and inverse document frequency can be calculated as:

$$TF = \frac{Wd.j}{\sqrt{\sum_{j=1}^{N} (wd.j)2}} \quad \ldots\ldots\ldots\ (2)$$

$$IDF = \frac{wq.j}{\sqrt{\sum_{j=1}^{N} (wq.j)2}} \quad \ldots\ldots\ldots\ (3)$$

The resulting files are arranged in the decreasing order of their scores obtained. The files with higher score value is the most relevant file to be retrieved.

4. EXPERIMENTAL RESULTS

Here, the dataset taken is Enron email dataset which contains documents of different authors in email related forms. The input will be multiple keywords given by the users. The users have to receive matching files with given keywords and their meanings from the data owner. This is the output to be obtained.

4.1. System Modules

The system can be explained as five modules. First module consists of details about encryption or decryption. Here the search has to be in encrypted index with encrypted keyword trapdoor. So a searchable symmetric encryption like homomorphic encryption is used here [4].

Second module gives the functions of data owner and file upload module. Owner key generation, build index and uploading encrypted documents to cloud are done in this module. Third module gives the details of user and expansion of keyword set. The synonyms of keywords are found out and thus the keyword set is expanded. These keywords are further encrypted using private keys and sent as trapdoors to the cloud server.

In fourth module, cloud server checks for similarity match between keywords in trapdoor and those in the index files. Resulting files are ranked according to the relevance score.

The final module allows the user to download and view the files. Only authorized users have the permission to download and decrypt owner files.

4.2. Test Results

The system is allowed to run with existing and proposed system cases. Existing system allows only multiple keywords in the search request. Proposed system considers the meanings of the multiple keywords also. The system is allowed to run on different number of files with various keywords and the performance is measured.

Precision and recall values are calculated with the help of true positive, true negative, false positive and false negative values. Precision gives the fraction of retrieved documents that are relevant whereas recall gives the fraction of relevant documents that are retrieved [13].

The formula for precision:

$$Precision = \frac{tp}{(tp + fp)} \quad \ldots\ldots\ldots\ (4)$$

The formula for recall:

$$Recall = \frac{tp}{(tp + fn)} \quad \ldots\ldots\ldots\ (5)$$

Accuracy can be calculated as:

$$Accuracy = \frac{(tp + tn)}{(tp + fp + fn + tn)} \quad \ldots\ldots\ldots\ (6)$$

The experiment is conducted in different set of input data with 50 to 200 files. The values obtained from the comparative study can be tabulated and corresponding graph can be plotted. P indicates precision values and R indicates recall values.

Table 3: Accuracy Calculation for Search with Ontology

Files	tp	tn	fp	fn	P	R	Accuracy
50	31	13	4	2	0.88	0.93	0.88
100	60	29	7	4	0.89	0.94	0.89
150	111	28	6	5	0.94	0.95	0.92
200	145	41	10	4	0.94	0.97	0.93

Table 4: Accuracy Calculation for Search without Ontology

Files	tp	tn	fp	fn	P	R	Accuracy
50	28	15	4	3	0.87	0.90	0.86
100	57	28	8	7	0.88	0.89	0.85
150	102	33	9	6	0.92	0.94	0.90
200	128	51	13	8	0.91	0.94	0.89

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

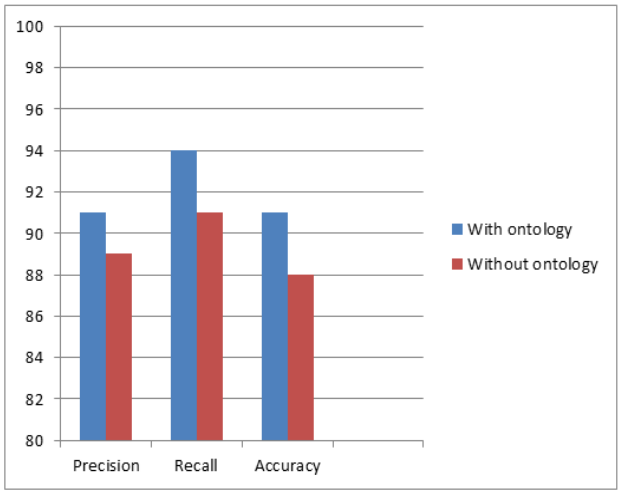

Figure 3: Graph showing percentage of precision, recall and accuracy with searching in two cases.

It is understood from the graph that recall rate is higher than that of precision. i.e., the probability of relevant data retrieved is high which implies that relevant data are retrieved through the proposed system. Also, the proposed system is found to be more accurate than the existing method.

5. CONCLUSION AND FUTURE WORK

With the attractive advantages of cloud storage, more customers are ready to outsource their personal data in the cloud. Cryptographic techniques like encryption are used to provide data security in cloud. Keyword search is used to retrieve files of interest by users. To overcome the disadvantages of existing system and for obtaining more accurate results, ontology based multiple keyword search is proposed in this paper. The system architecture, proposed method and experimental results are discussed. Test results shows that precision, recall and accuracy values of proposed system are higher than that of existing method. Thus the proposed system is more accurate in retrieving the most relevant files.

For future work, semantic-based search method can be taken into consideration while searching over encrypted cloud data. The semantic- based search method consists of syntactic transformation and anaphora resolutions to find out the most relevant results.

ACKNOWLEDGEMENT

I am extremely happy to acknowledge the service and co-operation rendered by Mr. Sulphikar A., Associate Professor, Department of Computer Science and Engineering, LBSITW for his valuable suggestions and guidance throughout the completion of this work.

I also acknowledge with grateful thanks the authors of the references I have referred for doing this paper.

References

[1] **PENG Yong et al. / "Secure cloud storage based on cryptographic techniques"/ science direct / 2012.**

[2] **S. Kamara and K. Lauter, "Cryptographic cloud storage", in RLCPS, January 2010, LNCS, Springer, Heidelberg.**

[3] **A.Singhal, "Modern information retrieval: A brief overview", IEEE Data Engineering Bulletin.**

[4] **S.Kamara,"http://outsourcedbits.org/2013/10/06/how-to-search-on-encrypteddata/".**

[5] **R. Curtmola, J. A. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: improved definitions and efficient constructions", in Proc. of ACM CCS, 2006.**

[6] **D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search", in Proc. of EUROCRYPT, 2004.**

[7] **Tarik Moataz, Abdullatif Shikfa, "Boolean Symmetric Searchable Encryption", ASIA CCS '13 Proceedings of the 8th ACM SIGSAC, pp. 265-276, NY, USA ,2013.**

[8] **J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy keyword search over encrypted data in cloud computing", in Proc. of IEEE INFOCOM10 Mini - Conference, San Diego, CA, USA, March 2012.**

[9] **N. Cao, C. Wang, M. Li, W. Lou and K. Ren, "Enabling Secure and Efficient Ranked Keyword Search Over Outsourced Cloud Data", IEEE Trans. Parallel and Distributed Systems, vol 23, no 8,pp. 1467-1479, Aug 2012.**

[10] **I.H. Witten, A. Moffat, and T.C. Bell, "Managing Gigabytes: Compressing and Indexing Documents and Images". Morgan Kaufmann Publishing May 1999.94.**

[11] **Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data", Proceedings of IEEE INFOCOM 2014, pp. 829-837, 2014.**

[12] **Zhangjie Fu, Xingming Sun, Nigel Linge and Lu Zhou, "Achieving Effective Cloud Search Services: Multi- keyword Ranked Search over Encrypted Cloud Data Supporting Synonym Query", IEEE Transactions on Consumer Electronics, Vol.60, No. 1, February 2014.**

[13] **Powers, DavidMW"Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies2 (1): 3763, 2011.**