IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

TF-IDF and KDC based Privacy Preserving Multi keyword Search over Distributed Encrypted Documents

Miss. Nilima Kasar[1], Prof. Vitthal S. Phad [2]

Department of Computer Engineering, STES's SMT. Kashibai Navale College of Engg,

Savitribai Phule Pune University, Pune, India [1, 2]

Abstract: In recent years, growth of private and semi-private information has grown up rapidly on information network; mechanisms to search such information have failed in privacy preservation. The privacy preserving searching is playing important role in the field of information networks to perform various data mining operations on encrypted data stored in various storage systems. It is also important and challenging task to protect the confidentiality of private data shared among service providers and data owners. Existing system provides one possible solution that is privacy preserving indexing (PPI). In this system, documents are stored in plain text form on private server that is privacy is compromised. So to enhance this system to make it more secure and efficient, first we store the documents on server in encrypted form and then use Key Distribution Center (KDC) for allowing decryption of data received from private server, at client side. We also implement TF-IDF, which provides the efficient ranking of results, to improve the user search experience. Finally we conduct the extensive experiments on dataset, to evaluate the performance of our proposed system. Experimental results will show that the proposed system is better than existing one, in terms of, privacy preserving, efficient and secure search on encrypted distributed documents.

Keywords: Privacy preserving indexing, Information networks, Encrypted data storage.

I. INTRODUCTION

Online social networks (OSNs) are most popular and it is the reason why millions of users accesses Internet. By this social networks users are share their information between the friends easily. On-demand computing is a kind of Internet- based computing that provides the data, resources as well as information to computers and other devices. Information is stored as well as processed by third-party data centers and the cloud computing as well as storages solutions provided to organizations and users with a number of capabilities. Base of cloud computing is a vast concept of converged infrastructure and shared services and it is based over distribution of resources to obtain consistency and economies of scale, same to a service on a network.

Within data networks, individual service suppliers store private personal information of distinct data owners. All the information sharing is accomplished in the observation of strong access control rules. These data networks have the following primary functions:

- Contributors together not trust each other in diverse domain
- Have responsibility of giving privacy to owners
- It is important to share information between providers from an application perspective.

In data Networks, data owners are free to store their files on number of distributed servers. It provides services to its users to store as well as access their information in and from number of server form anywhere and also by any device. Providing effective search over distributed files and additionally give the privacy to owners files it is a very difficult task. In existing system a technique used to solve this problem called privacy preserving indexing. The primary objective of PPI is to support a global search facility which is controlled by a third-party entity. The architecture of PPI is suitable for the providers such as entire access control over individual files and secures their privacy.

PPI [1] is a directory service available within a public cloud. Public cloud has control over different private servers. The information stored over number of private servers by distributed manner. This system permits different users for finding files over distributed data. For searching appropriate files user provide a query along with related keywords to the PPI [1] [3] server. After that, this public server return provides a list of private servers within the network.

Then user accesses the private server displayed in candidate list and after that user requesting for authentication before searching locally there. In this system, information stored in plain text manner on private server, thus user is able to search directly for needed files. But security of data is essential; so, in the proposed systems, data is stored in encrypted manner over private

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

servers. Therefore user has to authenticate and after authentication, user has access to encrypted files from private server. After obtaining encrypted files, decryption of files are done by utilizing the KDC. In cryptography, a key distribution center (KDC) is a portion of a cryptosystem expected to lessen the dangers inalienable in trading keys. KDCs regularly work in frameworks inside which a clients may have consent to utilize certain services at several times and not at others. KDC give a key to authorized users for decrypt the files. When original files are aggregated, then system implements TF-IDF ranking over files, to obtain top outcomes in ranking format.

This paper focus on different related work done by researches on the privacy preserving systems in section II, the implementation details in section III which focus on the system architecture and overview, mathematical model, algorithms and experimental setup. The section IV shows the expected results of a system. At last the conclusion is provided in section V.

II. RELATED WORK

An efficient and flexible technique for Multi keyword query search over encrypted data is proposed in [2]. System also performs ranking and verification of results. The multi keyword text search with similarity based ranking (BMTS) is enhanced with cipher text model and termed as Enhanced BMTS (EMTS). To provide efficient search, system uses search indexing with term frequency model and also uses vector space model with cosine similarity measure for highest accuracy results. System implements known cipher text model and known background model to provide privacy over search results. The system is evaluated with created document set. This document set contains recently published 3600 IEEE INFOCOM publication papers and extracts 9000 keywords. The performance of system is measured in terms of privacy, efficiency and effectiveness of proposed system.

A group index structure for multi-keyword Search with ranking and efficient incremental update is proposed in [3]. This system provides guaranteed confidentiality of results by encrypting it. Document replication and partition with one-to- many order used for guaranteed privacy-preserving utilizes one too many order preserving replication and partition of documents. Group index structure is used to increase the efficiency. It has low communication overhead. The proposed approach is evaluated on real-world dataset. This dataset contains set of emails came from Cognitive Assistant that Learns and Organizes (CALO) Project.

The problem of privacy preservation and data retrieval from encrypted data over public clouds is proposed in [5]. To achieve this Double Layer Encryption (DLE) and Hierarchical Multi-Keyword Ranked Search schema (HMRS) are used by authors. With this technique, system provides confidentiality of the data and its ranking also provides privacy of users. This system is secure, efficient and requires minimum cost.

The problem of privacy preserving and ranking of fuzzy keyword search over cloud encrypted data is solving in [6]. It enhances the efficiency and usability of system. It ranks the results on the basis of certain criteria like keyword frequency. System edits the distance for quantifying keyword similarity and dictionary based fuzzy set construction for creating fuzzy keyword sets. It will reduce the size of index and cost of storage and communication. The system is evaluated with 10714 real word data files downloaded from http://www.ietf.org/rfc.html and prove that system works securely and efficiently.

The sensitive information leakage problem is solving in [7], by implementing core attribute aware techniques. It has ensured the outsourced data privacy. For attribute indexing system implements k-anonymity technique. It also protects the data from unauthorized entities.

A personalized privacy preserving index (ϵ-PPI) is implemented in [8]. System identifies a common identity attack to break previous PPI. It develops an identity mixing protocol to detect the attack. This is the first model which does not require any trusted third part. The protocol is implemented by using secure multi-party computation (MPC) technique. For network Simulation, system uses, $500 - 25,000$ small digital libraries included in distributed dataset and TREC- WT10g dataset.

Techniques based on well-known threshold-based visual secret sharing scheme are proposed in [9]. It solves the problem of privacy and trust in cloud databases and database-as-a- service offerings. System also implements indexing technique for the secret records of large database based on some important properties of secret sharing technique.

The RASP data perturbation Method is proposed in [11]. This system provides a secure and efficient range of query and kNN query services for data protection in cloud. This method combines data encryption, expansion of dimensionality, random noise injection, and random projection. This system is attack resistance and uses indexing process for increase the speed of query processing. The kNN-R algorithm is developed for kNN queries processing. The system is evaluated with three dataset namely, synthetic dataset, adult dataset and 2-dimensional North East location data from rtreeportal.org.
A scheme for privacy preserving outsourced mining is proposed in [12] based on background knowledge. This is the attack model based on conservative frequency. According to this model server having the idea of exact item sets in owner's data and support of item in original data. The main goal of this system is to develop an encryption scheme to provide the privacy over owner's data and evaluate this system on large- scale real-life transaction databases (TDB).

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

An efficient light-weight keyword search with ranking for cloud computing is proposed in [13]. First the polynomial function is implemented used for hiding the encrypted keywords and search patterns. To provide privacy search, a secure inner product method is used. Another, multi-keyword ranked search over encrypted cloud data (MRSE) is proposed in [14] and [15], based on coordinate matching and index search with inverted matrix.

III. IMPLEMENTATION DETAILS

A. System Overview

In PPI system and information network, each server having multiple documents. Every document has multiple keywords. System consists of public cloud server, multiple private servers and multiple users. The owners encrypted documents are store on private servers in distribute manner. AES algorithm is used for data encryption purpose. At each private server, index file of data is created. All index files created at various private servers are collected and merge at monitoring systems and from there send to public cloud. When client wants some document from server, it poses a query to public cloud server. In returns, public cloud provides the merged index. Now from this final merge index, client having the list of private server at which query related data is stored. Then to access the data at server, client sends the authentication request with user name and password. Private server verifies this details store in its database. After successful verification, private server generates the token and sends it to client and Key Distribution Center (KDC). After getting this token, user requesting to KDC for key. KDC verify this token with its token which is already getting from private server. After verification, KDC provides encryption key to the client. Then client send data request to private server in returns server provides all matching encrypted files. Using key, client can decrypt the data. And finally apply the TF-IDF ranking algorithm, to get all results in ranking format. The ranking of results provides better and efficient search experience to users.

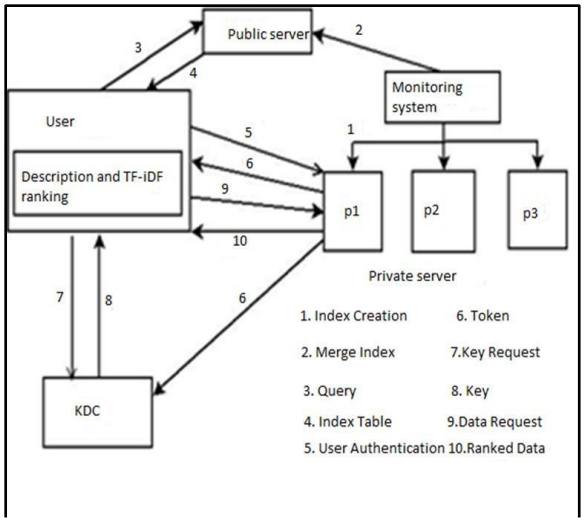


Fig. 1: System Architecture

System consisting of following modules:

- System Deployment

This module contains user registration as well as login with Database, Client and Server with socket programming and encrypted data by AES Encryption and transferred to client side. In Client side it decrypted along with GUI.

- MPPI Index creation algorithm

MPPI algorithm is implemented to generating index of every private server. Index indicates the detail explanation of data stored at private server.

- Index combining and Upload on Cloud

Observing system has responsibility for index mixture of every private server and uploads this last merge index file over public cloud.

- Input Query and response from public Cloud

User uploads a query to cloud server for obtaining specific information from private server and in response public cloud gives merged index.

- User Authentication and token generation

After more obtaining index, user has to connect to private server to access the outcomes. First, user will login to the server and after accomplishing successful authentication, private server create and disperse the token to user as well as KDC.

- Key distribution and file decryption

On the basis of authentication of tokens, KDC give the key to user and this key is used for decryption of outcomes that is collected from private server.

- TF IDF ranking results

After authentication, user will collect the outcomes from private server is in encrypted manner. These encrypted outcomes are decrypted by implementing key received from KDC. At the end, the ranking of outcomes are created by implementing TF IDF.

B. Algorithms

Advanced Encryption Standard (AES) Algorithm:

The algorithms used in AES are so easy that they can be easily implemented using cheap processors and a minimum amount of memory. The key factor of AES is its efficient Implementation. AES is based on a design principle known as a substitution-permutation network, combination of both substitution and permutation, and is fast in both software and hardware. Unlike its predecessor DES, AES does not use a Feistel network. AES is a variant of Rijndael which has a fixed block size of 128 bits, and a key size of 128, 192, or 256 bits. By contrast, the Rijndael specification with block and key sizes that may be any multiple of 32 bits, both with a minimum of 128 and a maximum of 256 bits.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

AES operates on a 4×4 column-major order matrix of bytes, termed the state, although some versions of Rijndael have a larger block size and have additional columns in the state. Most AES calculations are done in a special finite field.

1. Key Expansions
Round keys are derived from the cipher key using key schedule. AES requires a separate 128-bit round key block for each round plus one more.

2. Initial Round
AddRoundKey: Each byte of the state is combined with a block of the round key using bitwise xor.

3. Rounds
a. Sub Bytes—a non-linear substitution step where each byte is replaced with another according to a lookup table.
b. Shift Rows—a transposition step where the last three rows of the state are shifted cyclically a certain number of steps.
c. Mix Columns—a mixing operation which operates on the columns of the state, combining the four bytes in each column.
d. Add Round Key

4. Final Round (no Mix Columns)
a. Sub Bytes
b. Shift Rows
c. AddRoundKey.

Iterative publish Algorithm:

Algorithm: Iterative-Publish (Owner Pi; set $\{\beta'(rk\ )\}$)
1. For all$\in$ [0, $l-1$]
1. Do $\triangleright$ $'(rk)$is topologically sorted.
2. if match (cur_memvec, getStartingState(rk))
3. then$\triangleright$cur_memvec is the current membershipvector
4. cur_memvec $\leftarrow$publish(cur_memvec, $\beta'(rk\ )$)
5. end if
6. end for

TF_IDF:
TF_IDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

• Term Frequency:
In the case of the term frequency tf (t,d), the simplest choice is to use the raw frequency of a term in a document

$$tf\,t,d = 0.5 + 0.5 \quad \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

•Inverse document frequency:
The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient

$$idf\quad t,\ D\quad = \log \frac{N}{|\{d \in D : t \in d\}|}$$

C. Mathematical Model
Let S be a System. S= {I, P, O}

Where,
- Input I: The input for the system is multi word query from the user.
- Output O: Ranking results.
- Process P:

1. Single-term publication

$$\varepsilon = \frac{1-\sigma_j\ .\beta(t)}{1-\sigma_{i,}\beta_i+\sigma_i} \Rightarrow \qquad \beta_i = [(\sigma_l^{-1}\ -1)(\varepsilon_l^{-1}-1)]^{-1}$$

Where, β_j is number of probability values produces by source analytical computation for term.

2. False Positive Rate:

$$F(0,1) = \frac{F(0,1)}{F\ 0,1 + \sigma 0\sigma 1}$$

Where, FP (0, 1) is the false positive values, $\beta 0$, $\beta 1$ are the probability at which a non-positive owner publishes data as a positive owner.

3. Index Generation I= {I1, I2, …, In}
Where I is the set of all index of all private servers.

4. Merge and upload index at private cloud.
MI= {MI1, MI2, MIn}
Where MI is the set of all merge indexes collected from monitoring system.

5. User Query to public cloud
Q= {Q1, Q2, …, Qn}
Where, Q is the set of all queries poses to public cloud.

6. User Authentication at private server
U= {U1, U2, …, Un]
Where U is the set of all authenticated users of private server.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

7. Token Generation and distribution
T= {T1, T2, …, Tn}
Where T is the set of all tokens generated by private server for its authenticated users.

8. Key Generation at KDC
K= {K1, K2, …, Kn}
Where K is the set of all keys stored at KDC, used for decryption of data at user side.

9. Data decryption and TF_IDF ranking
R= {R1, R2, …, Rn}
Where R is the set of all ranked results for particular input query.

IV. RESULT AND DISCUSSION

A. Experimental Setup
The system is built using Java framework version jdk 1.8 on Windows platform. The Netbeans version 8.2 is used as a development tool. The system doesn't require any specific hardware to run; any standard machine is capable of running the application. The system analysis is carried out on datasets consisting of files.

B. Dataset
Dataset for peer to peer network with 2500 file names and their sizes are used for proposed system. Dataset consists of 1.6 million queries and data derived from NIST's available TREC WT10g.We can create index table by uploading files on private server.

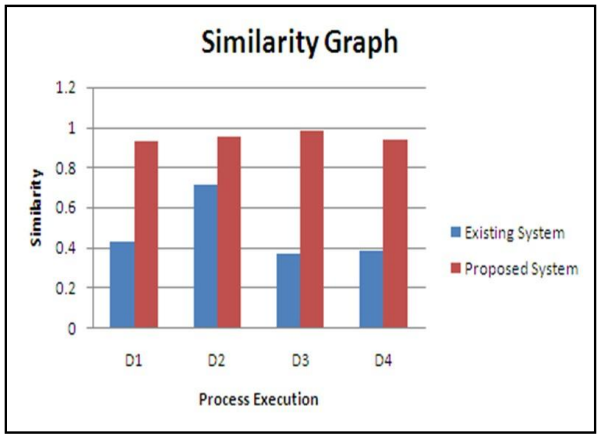
C. Results
By using non-grouping based approach of PPI the proposed system will going to provide better preservation of user's privacy in terms of data confidentiality through encryption and better quality of results i.e. relevant results to the queried using ranking techniques.

Similarity Measurement:
In the table 1 evaluate the similarity for the both existing and proposed system. Run the project four times and obtained the result which contain different similarity values. From the values it shows that the similarity values for the existing system is less than the proposed system.

Table 1: Similarity Table

	Existing System	Proposed System
D1	0.43	0.93
D2	0.71	0.95
D3	0.37	0.98
D4	0.38	0.94

Graph in Fig. 2 shows that the proposed system performs better than existing system in terms of similarity measures for text file.



Time Measurement:
In the table 2 measure the time for different process like uploading the file, query searching, encryption time, token generation, and ranking time. Run the project twice and plot the graph.

Table 2. Time Measurement Table

	File Upload	Query Search	Encryption Time	Token Generation	Ranking
D1	3.08	0.97	2.03	0.23	0.84
D2	5.93	0.38	3.87	0.47	0.72

In the graph Fig. 3 shows the time graph for the proposed system. Fetch the value from the above table and plot the graph.

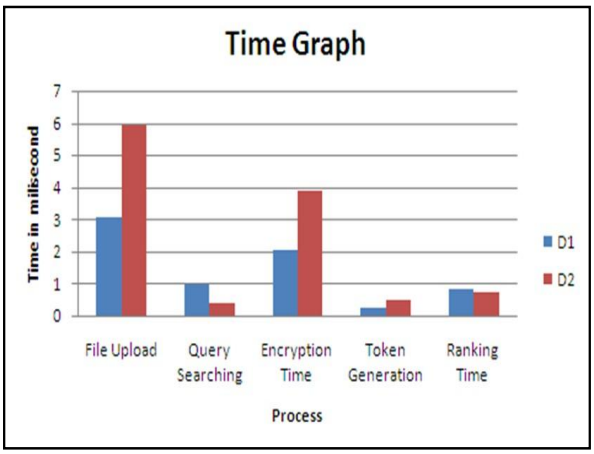


Fig.3 Time Graph

V. CONCLUSION

The proposed system is about linking between local server and cloud server for data sharing among the users. Some authentication is required to access specific data or information. This authentication is handled through encryption system. For sensible performance of secure computations, it proposes Associate in Nursing MPC [9] reduction technique supported the economical use of

secret sharing schemes. So, through the proposed system user can get an access to required data in ranked order using PPI and encryption technique.

REFERENCES

[1] Yuzhe Tang and Ling Liu, Fellow , "Privacy-Preserving Multi-Keyword Search in Information Networks", IEEE transactions on knowledge and data engineering ,vol. 27, no. 9, Sept 2015
[2] Sun, Wenhai, et al. "Verifiable privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking." Parallel and Distributed Systems, IEEE Transactions on 25.11 (2014): 3025-3035.
[3] Tseng, Ching-Yang, Chang Chun Lu, and Cheng-Fu Chou. "Efficient privacy-preserving multi-keyword ranked search utilizing document replication and partition." Consumer Communications and Networking Conference (CCNC), 2015 12th Annual IEEE. IEEE, 2015.
[4] Hu, Haibo, et al. "Private search on key-value stores with hierarchical indexes." Data Engineering (ICDE), 2014 IEEE 30th International Conference on. IEEE, 2014.
[5] Ajai, Ajni K., and R. S. Rajesh. "Hierarchical Multi-Keyword Ranked search for secured document retrieval in public clouds." Communication and Network Technologies (ICCNT), 2014 International Conference on. IEEE, 2014.
[6] Xu, Qunqun, et al. "Privacy-Preserving Ranked Fuzzy Keyword Search over Encrypted Cloud Data." Parallel and Distributed Computing, Applications and Technologies (PDCAT), 2013 International Conference on. IEEE, 2013.
[7] Zhang, Hongli, et al. "Practical and privacy-assured data indexes for outsourced cloud data." Global Communications Conference (GLOBECOM), 2013 IEEE. IEEE, 2013
[8] Tang, Yuzhe, et al. "e-ppi: Locator service in information networks with personalized privacy preservation." Distributed Computing Systems (ICDCS), 2014 IEEE 34th International Conference on. IEEE, 2014.
[9] Dutta, Ritaban, and B. Annappa. "Privacy and trust in cloud database using threshold-based secret sharing." Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on. IEEE, 2013.
[10] Kim, Hyeong-Il, Hyeong-Jin Kim, and Jae-Woo Chang. "A kNN query processing algorithm using a tree index structure on the encrypted database." 2016 International Conference on Big Data and Smart Computing (BigComp). IEEE, 2016.
[11] Xu, Huiqi, Shumin Guo, and Keke Chen. "Building confidential and efficient query services in the cloud with rasp data perturbation." Knowledge and Data Engineering, IEEE Transactions on 26.2 (2014): 322-335.
[12] Giannotti, Fosca, et al. "Privacy-preserving mining of association rules from outsourced transaction databases." Systems Journal, IEEE 7.3 (2013): 385-395.
[13] Ren, Yanzhi, et al. "Privacy-preserving ranked multi-keyword search leveraging polynomial function in cloud computing." Global Communications Conference (GLOBECOM), 2014 IEEE. IEEE, 2014.
[14] Karapakula, Anjaneyulu, M. Puramchand, and G. Mohammad Rafi. "Coordinate matching for effective capturing the similarity between query keywords and outsourced documents." Sustainable Energy and Intelligent Systems (SEISCON 2012), IET Chennai 3rd International on. IET, 2012.
[15] X. Jiang, -A Novel Privacy Preserving Keyword Search Scheme over Encrypted Cloud Data, 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), IEEE, Nov-2015.