Searching of multi-keywords in the images using OCR Technique

Asst. Prof.R.Tamilarasi
Dept. of computer science and engineering
MVJCE,Bangalore, India

M.Tech.Scholar Ms. Bindu K.M
Dept. of computer science and engineering
MVJCE,Bangalore, India

Abstract – **Cloud accepts workload computations and resources of the storages, and also it constitutes the protest on verification for compute & seclusion of the documents. Verification is investigated for safety protect multi-keyword search over ciphered documents in this work. Based on the server, the cloud server may or may not be return the incorrect results due to faults of the system or it may reduce the incentive computation cost, so here it is difficult to accept verifiability for searching outputs based on keyword search and safety protection for outsourced documents at the same time. To succeed these achievements, Verifiable safety protect multi-keyword Search plan is produced, called Verification Privacy Search, and it selects the homomorphic Message Authentication Code manner, it is integrated with a privacy-preserving multi-keyword search plan. The input we are taking as Images which should contain text keywords, so that the keyword extraction can be done using OCR extraction. The scheme is proposed that enables the client for verifying the searching outputs efficiently without storing a local copy of the outsourced documents. And random challenge technology is proposed according to the order for verifying top multi-keyword searching outputs; this technique detects incorrect top-keyword outputs with probability close to 1. So this work provides details on securing performance, verification, safety, and efficiency of the proposed plan. At lastly, Verification Privacy Search is implemented using java and also it evaluates its process on three UCI bag-of-words data groups.**

Keywords: **vpsearch, OCRextraction,homomorpic MAC, cloud computing, privacy preserving.**

I. INTRODUCTION

Cloud computing provides lot of benefits to enterprises to offload their data and software services to cloud saving them lot of money that has to be spent on infrastructure setup cost. Enterprises wanted to offload their data to cloud and save on their infrastructure cost. But for offloading the data, security and privacy are the two most important concerns.

Nowadays there are lot of cloud service providers in the market. The two requirements of security and privacy are the important decision criteria for enterprises in choosing the cloud service providers. For security many encryption protocols were proposed which the enterprise documents are offloaded to cloud, so relating to other cloud users the data is difficult to decrypt and make sense. Cloud computing is one of the emerging model for an on-demand network in the IT organization which gives access to a clique of computing resources.

It also helps in storage of data and applications allowing it to be ubiquity to the users. Both the individuals and the enterprises are inspired by the features of this and are deploying the documents to the server, substitute of buying hardware and software requirements themselves to survive their data.

Although cloud has variety of merits, the utilization of the sensitive information such as email, societies data's, to the remote clouds that may cause safety issues. Cloud service providers may access data user's embarrassed instruction without any official permission. To overcome this problem i.e. data confidentiality, the data has to be encrypted before deployment even though the problem is solved the data usage costs more.

For example, the multi-keyword based instruction recover technique is generally applied on unencrypted data but it cannot be applied on unauthorized data because download all the documents from server and decrypting is illogical.

Fully-homomorphic encryption is used to overcome the above problem. Since the computation and overhead is more to both users and cloud server this is impractical. On the Wither ward, many practical answers such as specific information (SI) manner have produced in length of satisfactory, quality and safety. This specific information method franchises the client to store the cipher text data to the servers and launches multiple-keyword search on encryption technology. The multi-keyword ranked searching accomplishes more for its project applications over only one keyword information, same-passage information, logical-passage information, and ranked-keyword information. Lately, a new dynamic scheme has been proposed for adding and removing operations on cloud data objects. The documents need to be updated by the data owner. Unless, some of the developing plan accept the satisfactory multi-keyword ranked information.

II. RELATED WORK

Order preserving encryption (OPE) is proposed in the works on [1] by A.Swaminathan. Using this scheme secure top k retrieval was possible. But the problem in this solution is attacker is still able to infer document contents. Also difficult to add documents.

C.Wang et al, works on range of characters blurred set that construct their own digital tree through searching indepublic clod. During this search technique, if the changes extends between reversal passage and from the blurred groups is < than a expecting hash function, its judgement is as same and it returns the agreement documents. This expecting blurred set method encourages the tolerance of smaller category and piece of software self-contradictory, but it will not encourage formal logics blurred-search [2].

Julien Bringer works on the conjunctive keyword search technique that can be capable to improve the efficiency of searching outputs. This type of conjunctive technique will regain the more effort efficient and the applicable documents. This kind of combinatory passage searching independent points the ranked search outputs. So, for this purpose the search mechanism flexible and effective will be better for the mechanisms [5].

K.Ren, C.Wang, works on “safety protest for the storage over server”. This kind of technique [6] is capable for the range of characters technique and for building blurred-keyword groups and its signs are involved on radix tree plan for pointing a multi-keyword way to store the blurred-passage groups that are to be pointed. The technique decreases the storage data of the process. And also mainly it is capable of changing extended idea to quantifier the passages as same.

Using the verification blurred-keyword searching, the data user points a signs based index tree with authorized documents and outsourced to the server. When the server receives the search request, the server will map the searching request to group of data’s or files. Each data is a task as a numbers and a group of passages. After searching, the server regains the searching petition and the fact for the result to the user. Using this fact, the data user will prove and rectify and perfection of the output search.

The searching is done based upon some following rules. (a)Firstly, The input given to search emphasizing conflict the pre-set passage. (b)Second, If the searching piece of software self-contradictory exist means, it will express the together specified conflicts are achieved. M.Belare, works on efficient searchable encryption method. A searchable re-encryption scheme is introduced in the index management scheme by Gentry in [7]. Using the re-encryption technique, the data user can contribute the documents with another’s securely by pointing searchable authorized index and re-authorized it.

The protecting specifications are set up and this security achieves 2 methods called authority factor and numbered authorization factor. The both kind of techniques provides efficiently. This searching technique is capable of multi-passages and then it provides the flexible performance.

M.Chuah in his work [8] proposed a multi-keyword search. Multi-keyword occurrence uses extent editing to construct blurred keyword groups. Spaces efficient are builted for every keyword. Then, it builds the index tree for all documents where each buds a hash function of a passage. But multi-keyword method is not suitable when the size of data increases, that offloaded to cloud. In work [9] by M. Van Dijik, first searchable encryption scheme using public and private key was proposed. In the scheme document owner encrypts the document using public key and uploads to cloud.

The users who knows private key can download and search the document. But the solution consumes lot of network bandwidth. In work [10] by N.Chao Coordinate matching and inner product similarity is used for keyword retrieval. But the concept is difficult to extend for multiple keywords. Boolean keyword retrieval is the technique proposed in works [11] by R.Curtmola. In this technique a search index is maintained with information of what terms found what document. It is a Boolean matrix.

When the search term is asked to search, the system searches in the Boolean matrix for the search term and finds the matching document. In this approach ranking of search result is not possible. Also the attackers can infer what the document contents are about even though it is encrypted. In work [12] by H.Hu, homomorphism solution was proposed for k nearest search queries. But the security is this solution can be easily compromised because of use of order preserving encryption (OPE).

III. SYSTEM DESIGN

A solid design is the fundamental for any good software. A good design always helps the structured planning, which foresees the problem area in advance and helps the project by avoiding redundant work. Software design involves creating and cleaning the equipment issues. The design of the project is given in this section.

1. Fundamental Design Concepts

Following design concepts are adhered in this project:

- Modular architecture.
- High Cohesion.
- Low Coupling.
- Reusability.
- Iterative development.

3.1. Input Design

The inputs for the project are:

- The document to upload.
- The number of search results.
- The multi keywords for searching.

3.2. Output Design

The project outputs are-

- The search results
- The decrypted document

3.3. The MVC Design Method

The project uses MVC design pattern.MVC pattern has three elements. Model, view and controller are three components.

3.4. Model-Model represents the state or low level behavior. In this project, the encryption, decryption and search component is represented as Model.

3.5. View-View represents the GUI screen, the interface for data uploading and searching are implemented in SWING.

3.6. Controller- The controller is component which manages the interaction between Model and View. The controller is realized using Network connection using UDP socket based abstraction.

2. System process methods

System process manners relevant to the idea are familiar to construct the process of the development. And it is a process in which a product will attain completion to meet the requirements. It is processed by the process technique in the plan is Water fall method. The System development method being followed for the software development is detailed in this process.

3.1. Model phases -The project uses traditional water fall model for the proposed software development. It consists for following steps and each step have defined milestone.

3.2. Requirement Analysis-Requirements are collected and documented in SRS.

3.3. System and software design-This process partitions into hardware or software systems.

3.4. Unit testing and implementation- According to the process, the equipment is realised as the group of programmers and the programming process.

3.5.Integrating and equipment testing-The individual programming process or programmers going to integrate and for the test as an entire equipment. To outcome this testing, the equipment specifications had met. Once the test is completed, the development equipment is delivering to data user in this process.

3.6. **System Design**: The design is made based on SRS and UML models are documented.

3.7Coding: The coding is done in JAVA that based on the UML models.

3.8 Implementation: The code is linked with necessary libraries and final jar file is developed.

3.9.**Testing**: The software goes though unit testing, integration testing and finally system testing.

3. System Architecture

System architecture is the idea that is used to design and shows the reality of the construction behavior and most opinion of the equipments. An architecture statement is the constitional statement that represents of the equipments. And it is high level structure of the system software.

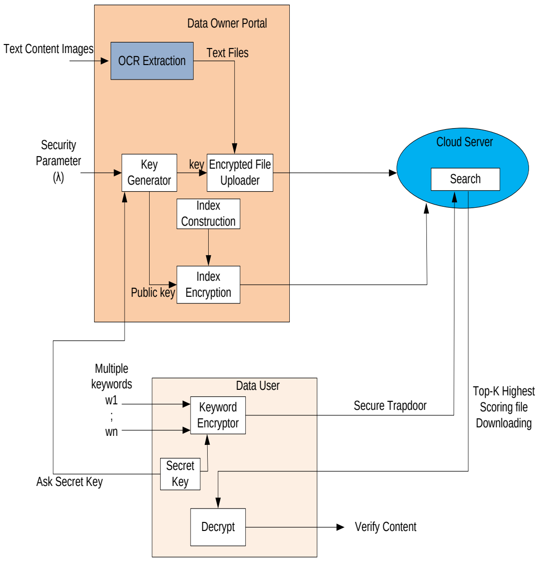

Fig.1.The System architecture

3.1. Module Description-

Three main modules are:

- Owner
- User
- Cloud Server

3.2. Data owner-The owner has the groups of the data, the owner wants to outsource to cloud in the cipher text. So here data owner uploads text content images to the cloud and owner will extract the text content images by using OCR method. After extracting it will save as a text files. Then the text files will be ciphered text then uploads to the cloud. So here to encrypt the files, the key is needed, for that purpose, the security parameter is used, using this parameter, the public key is generated. After that, the index construction should be done, called the construction of the index. In this index construction, first the keywords will be reading, second the stop words will be removed, and thirdly the vector file will be generated. All the keywords will be saved in the vector file. Then encrypt the index file and vector file by using the public key.

3.3.Data user- User is the one who searches for the documents by entering some keywords, only the authorized user's can access the documents of the data owner. User has to be searched for the multi-keywords in the extracted images. Randomly will enter some keywords regarding the images, then encrypted the keywords by asking the key from the owner, and then send it to the cloud server.

So with respect to this passages, the user can points to a trapdoor according to the search control mechanism technique to fetch the authorized data's from the server. And the server will display the matched wordlist from the databases and user can download the documents from the wordlist. The data user authorizes the list of the query and sends it to the cloud server, and that returns the relevant data documents to the user.

3.4.Cloud server-Server work is to store all the encrypted documents which is sent from the owner and the cloud will decrypt and send it to the user when the user requests for the documents

4. Implementation

3.1. Data Owner Algorithm

Data owner module is used for uploading the encrypted file to cloud and encrypted index to cloud. It is designed as swing standalone component with network socket based communication with other modules. By this way, it is possible to execute the data owner module in any different machine in the LAN network. Event based programming model is followed so that action code for every user action via button is realized.

The algorithms used by data owner consists of following steps

3.2. Input-The Image folder, and folder with both image and text files to upload

- Extract the Text info from image using OCR technique
- Extract all terms in the file
- Filter un important terms
- Homorphically encrypt the index.
- Generate RSA key.
- Encrypt the file using RSA public key.
- Upload the encrypted index to cloud.
- Upload the encrypted file to cloud Data user Algorithm.

5. Data user Algorithm

Data user is used for issuing multi keyword queries. The queries are encrypted and sent securely to the cloud server and the result of the query is given back to the user. Similar to data owner, event based programming model has been used to handle user actions via button.

The algorithm used by data user consists of following steps.

3.1. **Input-** The keywords to search, keyword index.

- Map the key words to index vector.
- Encrypt index vector homorphically.
- Send the encrypted index vector to Cloud server
- Wait for results.
- Display the results from cloud server Cloud server Algorithm.

6. Cloud Server Algorithm

Cloud server is written using background process model. A thread is started in background and waits for network messages from the data owner and server and handles the message. Cloud server use native API's of the cloud web services to invoke upload and download on the cloud. The API usage part is abstracted enough so that it is portable across any other cloud services with minor changes.

The algorithm at the cloud server is given below:

3.1. Input-The encrypted index vector from cloud user

- Download the encrypted index from cloud.
- Search the encrypted index vector on the encrypted index.
- The matching documents based on no of times of hit.
- Provide search result to the user.

The RSA algorithm steps are below:

- Choose p and q which are top quality numbers.
- Compute n = p*q .
- Compute r = (p-1)*(q-1).
- Choose e such that 1<e<r.
- Choose d such that (d*e) mod r = 1.
- Public Key (e,n).
- Private Key (d,n).
- Encrypt a message m, m^e mod n = C .
- Decrypt a Message, C^d mod n = m (Original message) .

IV.CONCLUSION

Searching on encrypted documents in cloud is an important research area and proposed effective solutions which consumes less bandwidth and also search time is lower in proposed solution. In future, the project will extend the solution for searching images and other contents which are encrypted and stored in cloud. The project shows the effectiveness of search by comparing with MRSE scheme and showed that the time to search is reduced in proposed approach and the accuracy of the search result is increased in this approach. The project also shows the size of the index constructed is efficient in this approach and simple compared to complex tree structures. The search time is also shown to polynomial bound. MRSE scheme computation time is non polynomial bound. The complexity of polynomial bound is most preferred because the search time increase linearly with the size of files uploaded to cloud. With the non-polynomial bound the time increases exponentially with the size of files uploaded to cloud, so the non-polynomial bound algorithms like MRSE is not suitable for bigger data sizes and not scalable.

V. FUTURE WORK

One of important problem in existing work is that integrity of search result cannot be ensured. Any attacker in between the cloud and user network side can corrupt the result or reorder the search result. Say if the list of hotels in an area is kept and user searching for hotel based on preference like food, the corrupt hotels can use attacker’s service to reorder the search result for search gains.To avoid this problem this project has integrity preserving and search result recovery mechanism. Integrity preserving scheme must be able to identify at the client side, if the search result returned by the cloud server is integral without any middle man modification attacks. Search result recovery mechanism must be able to restore original result from the cloud server when the integrity testing detects attack.Through the use of search result recovery mechanism, the user can avoid invoking the search again and will help to reduce the load on server and network communication overhead.

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,” in Proceedings of the ACM Workshop on Storage, Security, and Survivability, 2007, pp. 7–12.

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

[3]. E. Shi, J. Bettencourt, H. Chan, D. Song, and A. Per rig, “Multidimensional Range Query over Encrypted Data,”Proc. IEEE Symp. Security and Privacy,2007.

[4]. JianfengWang,XiaofengChen,HuaMa,Qiang Tang and Jin Li, “A Verifiable Fuzzy Keyword Search Scheme Over Encrypted Data”, Journal of Internet Services and Information Security (JISIS), volume: 2, number: 1/2, pp. 49-58.

[5]. Julien Bringer and HervéChabanne,” Embedding edit distance to enable private keyword search”Springeropen journal 2012.

[6]. K. Ren, C. Wang, and Q. Wang, “Security Challenges for the Public Cloud,”IEEE Internet Computing, vol. 16, no. 1, pp. 69-73, 2012.

[7]. M.Belare, A.Boldyreva, and A.O’Neil, “Deterministic and efficiently searchable encryption,” in Proceedings of rypto 2007, volume 4622 of LNCS. Springer- Verlag, 2007.

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

[9]. M. van Dijk, C. Gentry, S. Halevi, and V. Vaikuntanathan, “Fully Homomorphic Encryption over the Integers,” Proc. 29th Ann. Intl Conf. Theory and Applications of Cryptographic Techniques, H. Gilbert, pp. 24-43, 2010

[10]. N.Chao, Cong Wang, Ming Li, KuiRen, and Wenjing Lou“Privacy-Preserving Multi-keyword Ranked Search over Encrypted Cloud Data” Department of ECE, Illinois Institute of Technology.

[11]. R. Curtmola, J.A. Garay, S. Kamara, and R. Ostrovsky, “Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions,” Proc. ACM 13th Conf. 0Computer and Comm. Security (CCS), 2006.

[12]. T. M Nisha and V. P Lijo ,“Improving the Efficiency of Data Retrieval in Secure Cloud by Introducing Conjunction of Keywords”,