TOP K RESULT RETRIEVAL IN SEARCHING THE FILE OVER THE ENCRYPTED DATA IN CLOUD

Ms. T. Anusuyadevi, Mrs. A. Gokilavani B. Tech, ME.
PG Scholar, Assistant Professor/CSE,
Jay Shriram Group of Institutions, Tiruppur.
Email ID:anusivam.dev@gmail.com; Email ID: goke_surya@yahoo.co.in

Abstract- In cloud computing, data owners may share their outsourced data with a number of users, who might want to only retrieve the data files they are interested in. One of the most popular ways to do so is through keyword-based retrieval. We propose a new searchable encryption scheme, in which novel technologies in cryptography community and IR community are employed, including homomorphic encryption and the vector space model. In the proposed scheme, the data owner encrypts the searchable index with homomorphism based on cloud and then returns the encrypted scores of files to the data user. Next, the data user decrypts the scores and picks out the top-k highest-scoring files’ identifiers to request to the cloud server. The retrieval takes a two-round communication between the cloud server and the data user. The scheme the TRSE scheme, in which ranking is done at the user side while scoring calculation is done at the server side. To avoid the attacks, there is need to check whether the authorized user is logging or hackers is logging, so one new technique is used to avoid attacks by using color value technique, each user will have one color values and key color values. Also if same score is calculated for two queries, our technique finalizes one particular query by using the frequent calculation method at some particular time.

Index Terms—IR, Homomorph, Cloud, top-k

1.INTRODUCTION

Cloud computing has been envisioned as the next generation information technology (IT) architecture for enterprises, due to its long list of unprecedented advantages in the IT history: on-demand self-service, ubiquitous network access, location independent resource pooling, rapid resource elasticity, usage-based pricing and transference of risk. As a disruptive technology with profound implications, cloud computing is transforming the very nature of how businesses use information technology. One fundamental aspect of this paradigm shift is that data are being centralized or outsourced to the cloud. From users’ perspective, including both individuals and IT enterprises, storing data remotely to the cloud in a flexible on-demand manner brings appealing benefits: relief of the burden for storage management, universal data access with location independence, and avoidance of capital expenditure on hardware, software, and personnel maintenance, etc.

Cloud Computing

Cloud computing, or something being in the cloud, is an expression used to describe a variety of different types of computing concepts that involve a large number of computers connected through a real-time communication network such as the Internet. In science, cloud computing is a synonym for distributed computing over a network and means the ability to run a program on many connected computers at the same time. The phrase is also more commonly used to refer to network-based services which appear to be provided by real server hardware, which in fact are served up by virtual hardware, simulated by software running on one or more real machines. Such virtual servers do not physically exist and can therefore be moved around and scaled up (or down) on the fly without affecting the end user—arguably, rather like a cloud. The popularity of the term can be attributed to its use in marketing to sell hosted services in the sense of application service provisioning that run client server software on a remote location.

Advantages

Cloud computing relies on sharing of resources to achieve coherence and economies of scale similar to a utility (like the electricity grid) over a network. At the foundation of cloud computing is the broader concept of converged infrastructure and shared services.

The cloud also focuses on maximizing the effectiveness of the shared resources. Cloud resources are usually not only shared by multiple users but are also dynamically reallocated per demand. This can work for allocating resources to users. For example, a cloud computer facility, which serves European users during European business hours with a specific application (e.g. email) while the same resources are getting reallocated and serve North American users during North America's business hours with another application (e.g. web server). This approach should maximize the use of computing power, thus reducing environmental damage as well since less power, air conditioning, Rackspace, etc. is required for a variety of functions.

The term "moving to cloud" also refers to an organization moving away from a traditional CAPEX model (buy the dedicated hardware and depreciate it over a period of time) to the OPEX model (use a shared cloud infrastructure and pay as you use it).

Proponents claim that cloud computing allows companies to avoid upfront infrastructure costs, and focus on projects that differentiate their businesses instead of infrastructure. Proponents also claim that cloud computing allows enterprises to get their applications up and running faster, with improved manageability and less...
maintenance, and enables IT to more rapidly adjust resources to meet fluctuating and unpredictable business
demand.

**Services in Cloud**

Cloud computing is mostly used to sell hosted services in the sense of application service provisioning that run client server software at a remote location. Such services are given popular acronyms like 'SaaS' (Software as a Service), 'PaaS' (Platform as a Service), 'IaaS' (Infrastructure as a Service), and 'HaaS' (Hardware as a Service) and finally 'EaaS' (Everything as a Service). End users access cloud-based applications through a web browser, thin client or mobile app while the business software and user's data are stored on servers at a remote location.

Cloud management challenges

Cloud computing presents a number of management challenges. Companies using public clouds do not have ownership of the equipment hosting the cloud environment, and because the environment is not contained within their own networks, public cloud customers don’t have full visibility or control. Users of public cloud services must also integrate with an architecture defined by the cloud provider, using its specific parameters for working with cloud components. Integration includes tying into the cloud APIs for configuring IP addresses, subnets, firewalls and data service functions for storage. Because control of these functions is based on the cloud provider’s infrastructure and services, public cloud users must integrate with the cloud infrastructure management.

Deployment models

**Private cloud**

Private cloud is cloud infrastructure operated solely for a single organization, whether managed internally or by a third-party and hosted internally or externally. Undertaking a private cloud project requires a significant level and degree of engagement to virtualize the business environment, and requires the organization to reevaluate decisions about existing resources. When done right, it can improve business, but every step in the project raises security issues that must be addressed to prevent serious vulnerabilities. They have attracted criticism because users "still have to buy, build, and manage them" and thus do not benefit from less hands-on management, essentially "[lacking] the economic model that makes cloud computing such an intriguing concept".

**Public cloud**

A cloud is called a 'Public cloud' when the services are rendered over a network that is open for public use. Technically there may be little or no difference between public and private cloud architecture, however, security consideration may be substantially different for services (applications, storage, and other resources) that are made available by a service provider for a public audience and when communication is effected over a non-trusted network. Generally, public cloud service providers like Amazon AWS, Microsoft and Google own and operate the infrastructure and offer access only via Internet (direct connectivity is not offered).

**Community cloud**

Community cloud shares infrastructure between several organizations from a specific community with common concerns (security, compliance, jurisdiction, etc.), whether managed internally or by a third-party and hosted internally or externally. The costs are spread over fewer users than a public cloud (but more than a private cloud), so only some of the cost savings potential of cloud computing are realized.

**Hybrid cloud**

Hybrid cloud is a composition of two or more clouds (private, community or public) that remain unique entities but are bound together, offering the benefits of multiple deployment models. Such composition expands deployment options for cloud services, allowing IT organizations to use public cloud computing resources to meet temporary needs. This capability enables hybrid clouds to employ cloud bursting for scaling across clouds.

Cloud bursting is an application deployment model in which an application runs in a private cloud or data center and "bursts" to a public cloud when the demand for computing capacity increases. A primary advantage of cloud bursting and a hybrid cloud model is that an organization only pays for extra compute resources when they are needed.

**III. OBJECTIVE**

To solve the challenging problem of privacy-preserving multi-keyword ranked search over encrypted data in cloud computing, to improve search experience of the data search service, we further extend these two schemes to support more search semantics.

**III. EXISTING SYSTEM**

Coordinate matching means, as many matches as possible, is an efficient similarity measure among such multi-keyword semantics to refine the result relevance,
and has been widely used in the plaintext information retrieval (IR) community. However, how to apply it in the encrypted cloud data search system remains a very challenging task because of inherent security and privacy obstacles, including various strict requirements like the data privacy, the index privacy, the keyword privacy, and many others. In the literature, searchable encryption is a helpful technique that treats encrypted data as documents and allows a user to securely search through a single keyword and retrieve documents of interest. However, direct application of these approaches to the secure large scale cloud data utilization system would not be necessarily suitable, as they are developed as crypto primitives and cannot accommodate such high service-level requirements like system usability, user searching experience, and easy information discovery. Although some recent designs have been proposed to support Boolean keyword search as an attempt to enrich the search flexibility, they are still not adequate to provide users with acceptable result ranking functionality. Our early works have been aware of this problem, and provide solutions to the secure ranked search over encrypted data problem but only for queries consisting of a single keyword. How to design an efficient encrypted data search mechanism that supports multi-keyword semantics without privacy breaches still remains a challenging open problem. In this project, for the first time, we define and solve the problem of multi-keyword ranked search over encrypted cloud data (MRSE) while preserving strict system wise privacy in the cloud computing paradigm. Among various multi-keyword semantics, we choose the efficient similarity measure of “coordinate matching,” i.e., as many matches as possible, to capture the relevance of data documents to the search query. Specifically, we use “inner product similarity”, to quantitatively evaluate such similarity measure of that document to the search query. During the index construction, each document is associated with a binary vector as a sub-index where each bit represents whether corresponding keyword is contained in the document.

MRSE Framework

For easy presentation, operations on the data documents are not shown in the framework since the data owner could easily employ the traditional symmetric key cryptography to encrypt and then outsource data. Considering a cloud data hosting service involving three different entities, the data owner, the data user, and the cloud server. The data owner has a collection of data documents F to be outsourced to the cloud server in the encrypted form C. To enable the searching capability over C for effective data utilization, the data owner, before outsourcing, will first build an encrypted searchable index I from F, and then outsource both the index I and the encrypted document collection C to the cloud server. To search the document collection for given keywords, an authorized user acquires a corresponding trapdoor T through search control mechanisms, for example, broadcast encryption

Privacy-Preserving and Efficient MRSE

To efficiently achieve multi-keyword ranked search, we propose to employ “inner product similarity” to quantitatively evaluate the efficient similarity measure “coordinate matching.” Specifically, Di is a binary data vector for document Fi where each bit Di represents the existence of the corresponding keyword Wj in that document, and Qi is a binary query vector indicating the keywords of interest where each bit; Ig represents the existence of the corresponding keyword Wj in the query f W. The similarity score of document Fi to query f W is therefore expressed as the inner product of their binary column vectors. For the purpose of ranking, the cloud server must be given the capability to compare the similarity of different documents to the query. But, to preserve strict system wise privacy, data vector Di, query vector Q and their inner product Di should not be exposed to the cloud server.

Privacy-Preserving Scheme in Known Cipher text Model

In our more advanced design, instead of simply removing the extended dimension in the query vector as we plan to do at the first glance, we preserve this dimension extending operation but assign a new random number t to the extended dimension in each query vector. Such a newly added randomness is expected to increase the difficulty for the cloud server to learn the relationship among the received trapdoors. In addition, as mentioned in the keyword privacy requirement, randomness should also be carefully calibrated in the search result to obfuscate the document frequency and diminish the chances for identification of keywords. Introducing some randomness in the final similarity score is an effective way toward what we expect here. More specifically, unlike the randomness involved in the query vector, we insert a dummy keyword into each data vector and assign a random value to it.

Privacy-Preserving Scheme in Known Background Model

When the cloud server has knowledge of some background information on the outsourced data set, for example, the correlation relationship of two given trapdoors, certain keyword privacy may not be guaranteed anymore by the MRSE I scheme. This is possible in the known background model because the cloud server can use scale analysis as follows to deduce the keyword specific information, for example, document frequency, which can be further combined with background information to identify the keyword in a query at high probability. After

IV.PROPOSED SYSTEM

The proposed system is for searching the data from the encrypted data. The data gets encrypted by data owner with the keyword and stored in cloud. The user search for the data, the system will search for the results from the encrypted data. The relevance scoring and ranking methods are used for providing the accurate top k results.
Data encryption protects data security to some extent, but at the cost of compromised efficiency. Searchable encryption scheme allows retrieval of encrypted data over cloud. In this project, we focus on addressing data privacy issues using Searchable encryption scheme. For the first time, we formulate the privacy issue from the aspect of similarity relevance and scheme robustness. To eliminate the leakage, we propose a two-round searchable encryption (TRSE) scheme that supports top-k multikeyword retrieval. In TRSE, we employ a vector space model and homomorphic encryption. The vector space model helps to provide sufficient search accuracy, and the homomorphic encryption enables users to involve in the ranking while the majority of computing work is done on the server side by operations only on cipher text. As a result, information leakage can be eliminated and data security is ensured.

System Architecture

![System Architecture Diagram]

Fig.1 diagram for System Architecture

Data Flow Diagram

![Data Flow Diagram]

Fig.2 Data Flow Diagram

V. MODULE DESCRIPTION

The following modules are implemented in this technique

Uploading data

The data owner has a collection of n files to outsource onto the cloud server in encrypted form and expects the cloud server to provide keyword retrieval service to data owner himself or other authorized users. To achieve this, the data owner needs to build a searchable index from a collection of keywords extracted out of files, and then outsources both the encrypted index and encrypted files onto the cloud server.

Encryption of data

Solve the insecurity problem by proposing a two-round searchable encryption (TRSE) scheme. We propose a new searchable encryption scheme, in which novel technologies in cryptography community are employed, including Homomorphic encryption and the vector space model. In the proposed scheme, the data owner encrypts the searchable index with Homomorphic encryption. When the cloud server receives a query consisting of multikeywords, it computes the scores from the encrypted index stored on cloud and then returns the encrypted scores of files to the data user. Next, the data user decrypts the scores and picks out the top-k highest-scoring files’ identifiers to request to the cloud server. The retrieval takes a two-round communication between the cloud server and the data user. We, thus, name the scheme the TRSE scheme, in which ranking is done at the user side while scoring calculation is done at the server side.

Data Search

The data user is authorized to process multikeyword retrieval over the outsourced data. The computing power on the user side is limited, which means that operations on the user side should be simplified. The authorized data user at first generates a query. For privacy consideration, which keywords the data user has searched must be concealed. Thus, the data user encrypts the query and sends it to the cloud server that returns the relevant files to the data user. Afterward, the data user can decrypt and make use of the files.

Top k result

Considering the large number of data users and documents in the cloud, it is necessary to allow multikeyword in the search query and return documents in the order of their relevancy with the queried keywords. Scoring is a natural way to weight the relevance. Based on the relevance score, files can then be ranked in either ascendingly or descendingly. Several models have been proposed to score and rank files in IR community. Among these schemes, we adopt the most widely used one tf-idf weighting, which involves two attributes-term frequency and inverse document frequency.
VII. CONCLUSION AND FUTURE WORK

In this paper, we motivate and solve the problem of secure multikeyword top-k retrieval over encrypted cloud data. We define similarity relevance and scheme robustness. Based on OPE invisibly leaking sensitive information, we devise a server-side ranking SSE scheme. We then propose a TRSE scheme employing the fully Homomorphic encryption, which fulfills the security requirements of multikeyword top-k retrieval over the encrypted cloud data. By security analysis, we show that the proposed scheme guarantees data privacy. According to the efficiency evaluation of the proposed scheme over a real data set, extensive results demonstrate that our scheme ensures practical efficiency.

VIII. REFERENCES


AUTHORS BIOGRAPHY

T. Anusuyadevi received his B.E degree in Vivekananda College of Engineering for Women, Tiruchendur, India and currently pursuing M.E degree in Jay Shriram Group of Institutions, Tiruppur, India. His research interests include Data Mining, Advanced Database and Operating System.

Mrs. A. Gokilavani received her B.Tech. degree in Anna University, Chennai, India and M.E. degree in Anna University. Currently she is working as an Assistant Professor in Jay Shriram Group of Institutions, Tirupur, India. Her research interests include Data mining, Cloud Computing and Big Data.