International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 11 6371 - 6374

Survey Paper on Multi Keyword Similarity Search over Encrypted Cloud Data

Sumeet S. Pinjarkar , Tushar Saindane , Payal Rahatal , Pradip Shinde and Amol Kekan
Student, D. Y. Patil College of Engineering, Akurdi, Pune

Mrs. N. S. Patil, Mrs. M. Saravanapriya
Assistant Professor, D. Y. Patil College of Engineering, Akurdi, Pune

***Abstract*—**The tremendous amount of data outsourced every day by individuals or each enterprises . It is impossible to manage or to store this complex data at individual level, as the chances of crash the system is more, and the system becomes the single point of failure.When we feel the need of storing the data in such a way that it can be accessed uninterruptedly, then there the cloud comes into picture to store the data with better flexibility and cost saving. As the data might be confidential or sensitive. Considering the privacy of the data over the cloud, for that searchable encryption can be used. At the time of retrieval of data, consider the multi-keyword search over outsourced cloud text data only as it can handle the exact keywork matching. Multi-keyword similarity search overcomes the problem of not finding any related documents on searching. while encrypting the data before storing it to the cloud will help to preserve the privacy of the files. Searchable encryption also enables searching without revealing any additional information. Using multi-keyword similarity search cloud returns the files containing more number of matches with user input keywords and similar keyworks. Finding the similarities between input keyword or similar keyword is done by edit distance metric algorithm. Final design to achieve the user privacy, and to speedup the search task. At cloud side Bloom Filter's bit pattern is used to speedup and it is efficient in terms of the search time at the cloud side.
This paper presents a review on various existing Similarity searching techniques.

Index Terms*—*Software as a Service, Platform as a Service, Infrastructure as a Service, Bloom Filter.

*****

I. INTRODUCTION

Every individual is producing tremendous amount of date than ever before, and this rate is only going to increase day-by-day. Also more importantly the organizations have much higher rate of producing data which is in fact more sensitive too. Hence, organizations are often more concerned about the security of there data to store it on cloud storage, all of this leads to the increased authenticatinon demand. Considering the privacy of the data over the cloud, the searching techniques should be good enough to not to expose the data publicly, while searching.

As an approach to retaining control of data on cloud is to make use of the encryption of all cloud data. The problem is that encryption limits data. The encrypted data becomes prob-lematic in searching and indexing. Data stored in clear-text can be efficiently searched by specifying a keyword. This is not feasible to do with traditional encryption schemes. Enhanced and more sophesticated cryptography may offer new tools to make the data searchably encrypted. Encryption schemes like searchable encryption also known as "predicate encryption" that allow operation and computation on the ciphertext, allows the data owner to compute a capability from his secret key. A capability encodes a search query, and the cloud can use this capability to decide which documents match the search query, without the requirement of any additional information.

Other cryptographic techniques such as homomorphic en-cryption and Private Information Retrieval (PIR) perform computations on encrypted data without decrypting it.

Apart from introduction in section I, the paper is organized as follows- Cloud Computing System is explained in section II, a Multi keyword similarity search is explained in section III, Bloom Filter data structure is described in section IV using a block diagram. Section V concludes the paper.

II. CLOUD COMPUTING SYSTEM

The modules of a Cloud Computing System are discussed in this section. Fig 1 shows a typical speech recognition system.

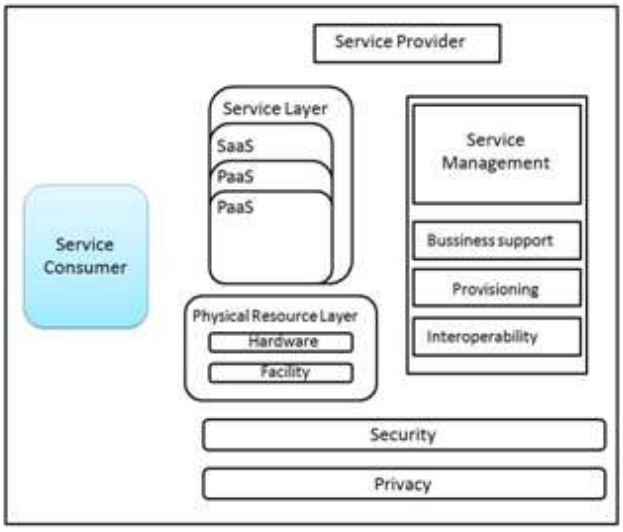

Fig. 1. Cloud Computing System[9].

The various modules of a Speech Recognition System are as follows-

A. Service Consumers

The entity that maintains bussiness relationships and user services.

B. Service Provider

It makes various cloud services available to the user.

III. Services Provided By Cloud

1) Software as a Service(SaaS): User gain access to ap-plication software and databases. Cloud providers manage the infrastructure and platform that rent the application. The SaaS is not suitable for applications that require real-time response or those for which data is not allowed to be hosted externally.

2) Platform as a Service(PaaS): In this model cloud providers delivers the computing platforms including oper-ating system, Programming language exicution environment, database and web server. Application developers can develop and run their software solutions on the cloud platform without the cost and complexity of buying and managing the under-lying hardware and software layer including network, servers, operating systems, or storage. The user has control over the deployed applications and, possibly, over the application host-ing environment configurations. Such services include session management, device integration, sandboxes, instrumentation and testing, contents management, knowledge management, and Universal Description, Discovery, and Integration (UDDI). PaaS is not particulary useful when the application must be portable, when proprietary programming languages are used, or when the underlaying hardware and software must be customized to improve the performance of the application.

3) Infrastructure as a Service(IaaS): Infrastructure-as-a-Service (IaaS) provides the physical or virtual machines and other resources. It is the capability to provision processing, storage, networks, and other computing resources. The con-sumer is able to deploy and run its own applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, stor-age, deployed applications, and limited control of networking components, such as host firewalls. It provides the services such as : Web servers, storage, computing hardware, operating systems, virtual instances, load balancing, Internet access.

IV. Multi Keyword Similarity Search

An important method of retrieving information which is in the form of the text data is the keyword search. Searching using vocabulary techniques can be effective and efficient. In many cases, it is preferable to search using controlled vocabu-lary terms because user often get higher quality,more specific results. Sometimess though controlled vocabulary searching is either not good or not even possible. In this situation user must rely on another set of techniques described as searching free text. Free text searching is also known as keyword searching. In addition to the advantages of keyword searching, making use of multiple keywords to search file is more benificial.

Usually when people do a search, they type in a keyword phrase instead of just a single keyword. Fifty-eight percent of search queries are three words or longer. So having keyword phrases on your site increases your chance of appearing higher on the page rank (because more keywords match the search query). The click-through rate (how many people click your listing to go to your site) also increases, due to more words matching the search query. Your conversion rate (how many visitors actually purchase something, sign up, or take whatever action is appropriate on your site) also increases because youre more likely to have what the user is looking for.

Computer index ”significant” words in databases in the title, summary, subject or even the text fields of a record or article. These words are then searchable. When you type these words into the database search window, this is called keyword searching. In Multi-Keyword searching all of the files containing at least threashold number of keywords specified by the user will be returned to the user[2].

For public-key encryption schemes where the encryption algorithm is deterministic. They use efficiently-searchable encryption schemes which permit more flexible privacy to search-time trade-offs via a technique called bucketization. This paper schemes only provide privacy for plaintexts that have high min-entropy. An important open question is to con-struct ESE or deterministic encryption schemes meeting our definition in the standard model. [3] Describes how Dynamic Searchable Symmetric Encryption allows a client to store a dynamic collection of encrypted documents with a server and later quickly carry out keyword searches on these encrypted documents, while revealing minimal information to the server.

A. Similarity Search-

Nearest neighbour search and Range queries are important subclasses of similarity search, and a number of solutions exist. Research in Similarity Search is dominated by the inherent problems of searching over complex objects.

B. Nearest neighbour search (NNS)

Also known as proximity search, similarity search, is an optimization problem for finding closest (or most similar) points. Closeness is typically expressed in terms of a dissimilarity function i.e. the less similar the objects, the larger the function values.

k-nearest neighbor search identifies the top k nearest neighbors to the query. This technique is commonly used in predictive analytics to estimate or classify a point based on the consensus of its neighbors. k-nearest neighbor graphs are graphs in which every point is connected to its k nearest neighbors.

C. A range query

It is a common database operation that retrieves all records where some value is between an upper and lower boundary. Range queries are unusual because it is not generally known in advance how many entries a range query will return, or if it will return any at all. Many other queries, such as the top

ten most senior employees, or the newest employee, can be done more efficiently because there is an upper bound to the number of results they will return. A query that returns exactly one result is sometimes called a singleton.[2] described the relevance of multi key-word search on encrypted data as, it is necessary to allow multiple keywords in the search request and return documents in the order of their relevance to these keywords. Related works on searchable encryption focus on single keyword search or Boolean keyword search, and rarely sort the search results a set of strict privacy requirements for such a secure cloud data utilization system.

V. Bloom Filter

A Bloom filter is a simple space-efficient probabilistic data structure for representing a set in order to support membership testing queries. The space efficiency is achieved at the cost of a small probability of false positives, but often this is convenient. Four types of network-related applications of Bloom filters are: Collaborating in overlay and peer-to-peer networks, Bloom filters can be used for summarizing content to aid collaborations in overlay and peer-to-peer networks.

Bloom Filter data structure has not been standardized yet for the Internet, because protocols implemented using it, do have some vulnerabilities .There are some proposed security mechanisms, that can be used to reduce the vulnerabilities and make the protocols that uses this data structure, more secure. Bloom Filter is probabilistic in the sence that the membership testing may return false positive results, that means Bloom filter may return true result for some element which is not the member of the set. The probability of false positive grows with the number of elements.

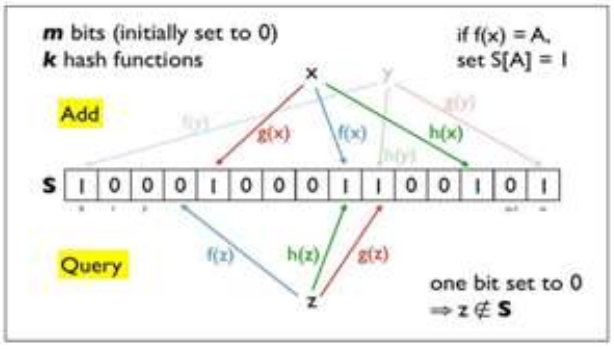

Fig. 2. Bloom FIlter Bit Representation

VI. Searching Over Encrypted Cloud Data

Searching architecture for encrypted cloud data includes following components: Data Owner, Data User, Cloud Storage or Server.

A. Data Owner

Data owner extraxt keyword from data collection. He also construct searchable encrypted index from the data collection, then he encrypt all files and send both encrypted index and file to cloud server.

B. Data User

User requests in the form of keywords to the cloud server.

C. Cloud Server

Recieves the request from user and then send the corre-sponding encrypted files to the user as response.

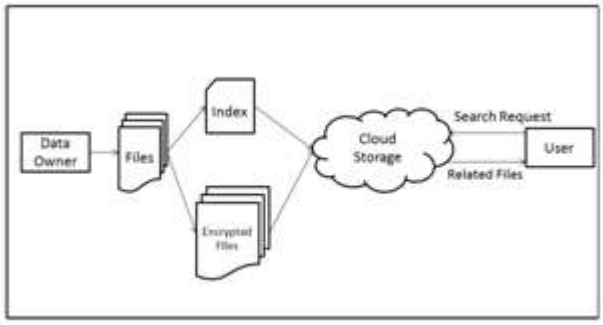

Fig. 3. Searching Over Encrypted Cloud Data

VII. Conclusion

In this paper we have described the importance of Multi key-word similarity search varientv,viz. Nearest Neighbour search, Ranked search, Range query. We mentioned the benifits of multi keyword over single keyword searching. We have also enlighten the importance of encrypting cloud data. Also the effect of using Bloom Filter for membership testing. We have discussed the theory of cloud computing services viz. SaaS, PaaS, IaaS.

VIII. ACKNOWLEDGMENT

We express our deepest gratitude to the guides for technical guidance and infrastructure. Lastly we wish to thank the researchers for their contributions because of which we could complete this work.

IX. References

[1] Hanhua Chen , Hai Jin , Lei Chen, Yunhao Liu, and Li-onel M. Ni., Optimizing Bloom Filter Settings in Peer-to-Peer Multikeyword Searching. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 4, APRIL 2012

[2] Ning Cao, Cong Wang, Ming Li, Kui Ren, and Wenjing Lou., Privacy-Preserving Multi-Keyword Ranked Search

over Encrypted Cloud Data. IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 1, JANUARY 2014

[3] Muhammad Naveed, Manoj Prabhakaran, Carl A. Gunter University of Illinois at Urbana-Champaign . Dynamic Searchable Encryption via Blind Storage. 2014 IEEE Symposium on Security and Privacy

[4] Zhihua Xia, Member, IEEE, Xinhui Wang, Xing-ming Sun, Senior Member, IEEE,and Qian Wang, Mem-ber, IEEE. A Secure and Dynamic Multi-keyword Ranked Search Scheme over Encrypted Cloud Data.

[5] IEEE TRANSACTIONS ON PARALLEL AND DIS-TRIBUTED SYSTEMS, VOL. NO 1, FEBRUARY 2012 [5] D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, Public key encryption with keyword search, in Proc. Int. Conf. Theory Appl. EUROCRYPT, 2004, pp.506522.

[6] L. Ballard, S. Kamara, and F. Monrose, Achieving efficient conjunctive keyword searches over encrypted data, in Proc. ICICS, 2005, pp. 414426.

[7] A. Broder and M. Mitzenmacher, Network applications of bloom filters: A survey, Internet Math., vol. 1, pp. 485509, 2005.

[8] D. Boneh and B. Waters, Conjunctive, subset, range queries on encrypted data, in Proc. Theory of Cryptogra-phy Conf. TCC, 2007, pp. 535554.

[9] Marinescu, Dan C, ”Cloud Computing - Theory and Practice- ”