Survey on Keyword Search in Encrypted Data With Privacy Preservation

Sonal Rahul Jamdade[1], **Jyoti N. Nandimath**[2]

[1]ME (Student) Dept. of Computer Engineering, SKN, Savitribai Phule Pune University, Pune, India

[2]ME (Assistant Professor) Dept. of Computer Engineering, SKN, Savitribai Phule Pune University, Pune, India

Abstract: ***Cloud computing is an on-demand computing. It is an Internet-based computing. In this area shared resources, data and information are provided on demand to computers and other devices. It also provides the services over the internet. In cloud computing, service providers have ability to provide storage at server according to users need. They allow users to store and retrieve the data in cloud server on demand from anywhere and on any type of device. This manipulation of data at cloud server gives rise to so many security issues because data is accessed over internet. For security purpose data is store in encrypted format. In this, client has no direct control on data once it is uploaded on cloud server. In this paper, we investigate the idea behind single keyword search over encrypted data and also multi keyword ranking. Cloud data owners want their documents in an encrypted form for the purpose of privacy preserving. Therefore it is necessary to develop efficient and reliable ciphertext search techniques. One challenge is that the relationship between documents will be normally concealed in the process of encryption, which will lead to significant search accuracy performance degradation.***

Keywords: Cloud computing, Encryption, Inner product similarity, Single Keyword Search, Multi-keyword search, ranking.

1. Introduction

As we step into the big data era, terabyte of data is produced world-wide per day. In the late 1960's the idea of "Utility computing" that was coined by MIT computer scientist and Turing award winner John McCarthy was preferably known as the concept of cloud computing over a network. Industries were looking for some sort of major solution, since utility computing ended up becoming something of a big business for companies such as IBM. Indeed, Martin Greenberger pointed out the concept that "advanced arithmetical machines of the future" were now being used not only institutionally for scientific calculation and research but also for business functions such as accounting and inventory. Further, he anticipated his piece of work in which computers would be universal almost like the major power companies running wires everywhere in due time. Enterprises and users who own a large amount of data usually choose to outsource their precious data to cloud facility in order to reduce data management cost and storage facility spending. As a result, data volume in cloud storage facilities is experiencing a dramatic increase. Although cloud server providers (CSPs) claim that their cloud service is armed with strong security measures, security and privacy are major obstacles preventing the wider acceptance of cloud computing service [1]. A traditional way to reduce information leakage is data encryption. However, this will make server-side data utilization, such as searching on encrypted data, become a very challenging task. In the recent years, researchers have proposed many ciphertext search schemes [35-38] [43] by incorporating the cryptography techniques. These methods have been proven with provable security, but their methods need massive operations and have high time complexity. Therefore, former methods are not suitable for the big data scenario where data volume is very big and applications require online data processing. In addition, the relationship between documents is concealed in the above methods. The relationship between documents represents the properties of the documents and hence maintaining the relationship is vital to fully express a document. For example, the relationship can be used to express its category. If a document is independent of any other documents except those documents that are related to sports, then it is easy for us to assert this document belongs to the category of the sports. Due to the blind encryption, this important property has been concealed in the traditional methods. Therefore, proposing a method which can maintain and utilize this relationship to speed the search phase is desirable. On the other hand, due to software/hardware failure, and storage corruption, data search results returning to the users may contain damaged data or have been distorted by the malicious administrator or intruder. Thus, a verifiable mechanism should be provided for users to verify the correctness and completeness of the search results. Due to a revolutionary change in the field of industries over past decade, there has been increase in demand of outsourcing of data over a wide range of network. In order to manipulate this huge amount of data in cost effective manner enterprise has adapted a prevalent technology called cloud computing that remove the burden of data management. In this data driven environment enterprise tend to store their data onto cloud that comprises of valuable asset of customer data like emails, personal health data etc. Cloud computing is turning out to be most essential paradigm in the development of information technology which offer flexible access, ubiquitous, on demand access and capital expenditure saving

2. Literature Review

Qin Liu et al. proposed Secure and privacy preserving keyword search in [1]. It provides keyword privacy, data privacy and semantic secure by public key encryption. The main issue of this search is that the communication and computational cost of encryption and decryption is more.

Ming Li et al. proposed Authorized Private keyword Search (APKS) in [2]. It provides keyword privacy, Index and Query Privacy, Fine-grained Search Authorization and Revocation, Multi-dimensional Keyword Search, Scalability and Efficiency. This search method increases the search efficiency using attribute hierarchy but in practice all the attributes are not hierarchical.

Cong Wang et al in [3] proposed Secure and Efficient Ranked Keyword Search which solves processing overhead, data and keyword privacy, minimum communication and computation overhead. It is not useful for multiple keyword searches, Also there is a little bit of overhead in index building.

Kui Ren et al. [4] proposed Secured fuzzy keyword search with symmetric searchable encryption (SSE). It does not support fuzzy search with public key based searchable encryption, also it cannot perform multiple keywords semantic search. The updates for fuzzy searchable index are not efficiently performed.

Ming Li et al. [5] proposed Privacy assured searchable cloud Storage method. It is implemented using SSE, Scalar-Product-Preserving Encryption and Order-Preserving Symmetric Encryption. It supports the privacy and functional requirements. This scheme does not support public key based searchable encryption.

Wei Zhou et al. [6] proposed K-gram based fuzzy keyword Ranked Search. In this owner create k-gram fuzzy keyword index for files D and tuple <I, D> is uploaded to search server (SS) which is inserted to bloom filter for size controlling. The encrypted file D is uploaded to storage server. But the problem is that, the size of the k-gram based fuzzy keyword set depends on the jacquard coefficient value.

J. Baek et al. in [7] proposed Secure Channel Free Public Key Encryption with Keyword Search (SCF-PEKS) method. In this method cluster servers creates its own public and private key pair but this method suffers from outside attacker by KGA.

H. S. Rhee et al. [8] proposed Trapdoor in distinguishability Public-Key Encryption with Keyword Search (IND-PEKS). In this outsourcing is done as SCF-PEKS. It suffers from outside attacker using KGA and analyzing the frequency of occurrence of keyword trapdoor.

Peng Xu et al. [9] proposed Public-Key Encryption PEFKS with Fuzzy Keyword Search, in this user creates fuzzy keyword trapdoor Tw and exact keyword trapdoor Kw for W. User requests Tw to CS. Then CS checks Tw with fuzzy keyword index and sends superset of matching cipher texts by Fuzz Test algorithm that is executed by CS. The user process ExactTest algorithm for verifying ciphertexts with Kw and retrieve the encrypted files. The process of creating fuzzy keyword index and exact keyword index is difficult for large size database.

Ning et al. [10] proposed Privacy Preserving Multi Keyword Ranked Search (MRSE). It is useful for known cipher text model and background model over encrypted data. It provides low computation and communication overhead. The coordinate matching is selected for multi-keyword search. The drawback is that MRSE have small standard deviation which reduces the keyword privacy.

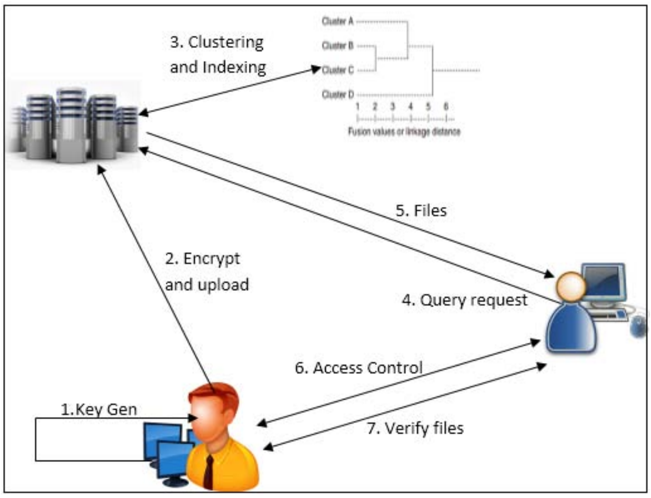
3. Proposed Work

Cloud data owners prefer to outsource documents in an encrypted form for the purpose of privacy preserving. Therefore it is essential to develop efficient and reliable ciphertext search techniques. One challenge is that the relationship between documents will be normally concealed in the process of encryption, which will lead to significant search accuracy performance degradation. Also the volume of data in data centers has experienced a dramatic growth. This will make it even more challenging to design ciphertext search schemes that can provide efficient and reliable online information retrieval on large volume of encrypted data.

The next generation web called the Semantic Web will help the user to retrieve the useful data that is stored on the cloud in the form of ontology and make the data visible to the user which is hidden behind the cloud. The aim of the proposed ranking algorithm is to provide users the result set of relevant data.

A hierarchical clustering method is proposed to support more search semantics and also to meet the demand for fast cipher text search within a big data environment. The proposed hierarchical approach clusters the documents based on the minimum relevance threshold, and then partitions the resulting clusters into sub-clusters until the constraint on the maximum size of cluster is reached. In the search phase, this approach can reach a linear computational complexity against an exponential size increase of document collection. In order to verify the authenticity of search results, a structure called minimum hash sub-tree is designed. We extend this notion of semantic similarity to consider inherent relationships between concepts using ontologies. We propose ranking algorithm with multi keyword and ontology.

4. Architectural View



A hierarchical clustering method is proposed to support more search semantics and also to meet the demand for fast cipher text search within a big data environment. The proposed hierarchical approach clusters the documents based on the minimum relevance threshold, and then partitions the resulting clusters into sub-clusters until the constraint on the maximum size of cluster is reached. In the search phase, this approach can reach a linear computational complexity against an exponential size increase of document collection. In order to verify the authenticity of search results, a structure called minimum hash sub-tree is designed. We extend this notion of semantic similarity to consider inherent relationships between concepts using ontology's. We proposed ranking algorithm with multi keyword and ontology.

Table 1: Survey Table

Sr. No	Paper	Author	Technique	Advantage	Disadvantage	Result
1	Practical techniques for searches on encrypted data [11]	D. X. D. Song, D. Wagner, and A. Perrig,	cryptographic schemes	Single Keyword Searchable Encryption	high searching cost due to the scanning of the whole	remote searching on encrypted data using an untrusted server
2	Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data [12]	C. Wang, N. Cao, K. Ren, and W. J. Lou	***keyword search and concept based search methods***	Order-preserving techniques are utilized to realize the rank mechanism	Not efficient relevance score calculation method	improve the efficiency of ranked keyword search Algorithms
3	Privacy-preserving multi-keyword text search in the cloud supporting similarity -based ranking [13]	. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li	a tree-based index structure and various adaption methods for *multi-dimensional (MD) algorithm*	Better search efficiency	At the stage of index building process, the relevance between documents is ignored	similarity-based ranking for more accurate search result and a tree-based search algorithm that achieves better-than linear search efficiency.
4	An Efficient Privacy-Preserving Ranked Keyword Search Method[14]	R. X. Li, Z. Y. Xu, W. S. Kang, K. C. Yow, and C. Z. Xu	Authentication based document retrieval	This approach can reach a linear computational complexity against an exponential size increase of document collection.	Algorithm is described only for adding the files, but not for how it work after files are updated for deleted	clusters the documents based on the minimum relevance threshold
5	Ranking Algorithm of Web Documents using Ontology[15]	Gurdeep Kaur, Poonam Nandal,	Ranking algorithm using ontology	computes the joint probability of page with respect to ontology by matching concepts	Not provides any kind of security or privacy to documents	providing the relevant result set to the users of the internet according to the query or keyword specified

5. Conclusion

This paper studies various techniques of searching in the encrypted cloud data storage. We have systematically presents the security and data utilization issues in the cloud storage related to all available searching techniques. Thus identified the main issues that are to be satisfied for secured data utilization are keyword privacy, Data privacy, Index privacy, Query Privacy, Fine-grained Search, Scalability, Efficiency, Result ranking, Index confidentiality, Query confidentiality, Query Unlinkability, semantic security and Trapdoor Unlinkability. Most of the searching techniques mainly focus on security and some on data utilization. The limitations of all the searching techniques are also discussed. By the above survey, security can be provided by Public-Key Encryption and effective data utilization by fuzzy keyword search. We believe that this survey will make the researchers to shape their problem in the area of data utilization in cloud storage.

References

[1] Qin Liuy, Guojun Wangyz, and Jie Wuz,"Secure and privacy preserving keyword searching for cloud storage services", ELSEVIER Journal of Network and computer Applications, March 2011

[2] Ming Li et al.," Authorized Private Keyword Search over Encrypted Data in Cloud Computing,IEEE proc. International conference on distributed computing systems,June 2011,pages 383-392

[3] Cong Wang et al.,"Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data", IEEE Transactions on parallel and distributed systems, vol. 23, no. 8, August 2012

[4] Kui Ren et al., "Towards Secure and Effective Data utilization in Public Cloud", IEEE Transactions on Network, volume 26, Issue 6, November / December 2012

[5] Ming Li et al.,"Toward Privacy-Assured and Searchable Cloud Data Storage Services", IEEE Transactions on Network, volume 27, Issue 4, July/August 2013

[6] Wei Zhou et al., "K-Gram Based Fuzzy Keyword Search over Encrypted Cloud Computing "Journal of Software Engineering and Applications, Scientific Research , Issue 6, Volume 29-32,January2013

[7] J. Baek et al., "Public key encryption with keyword search revisited", in ICCSA 2008, vol. 5072 of Lecture Notes in Computer Science, pp. 1249 - 1259, Perugia, Italy, 2008. Springer Berlin/Heidelberg.

[8] H. S. Rhee et al., “Trapdoor security in a searchable public-key encryption scheme with a designated tester," The Journal of Systems and Software, vol. 83, no. 5, pp. 763-771, 2010.

[9] Peng Xu et al., Public-Key Encryption with Fuzzy Keyword Search: A Provably Secure Scheme under Keyword Guessing Attack”,IEEE Transactions on computers, vol. 62, no. 11, November 2013

[10] Ning Cao et al.,” Privacy-Preserving Multi- Keyword Ranked Search over Encrypted Cloud Data”, IEEE Transactions on parallel and distributed systems, vol. 25, no. 1, jan 2014

[11] D. X. D. Song, D. Wagner, and A. Perrig, ”Practical techniques for searches on encrypted data,” in Proc. S & P, BERKELEY, CA, 2000, pp. 44.

[12] C. Wang, N. Cao, K. Ren, and W. J. Lou, Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data, IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 8, pp. 1467-1479, Aug. 2012.

[13] W. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li, ”Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking,” in Proc. ASIACCS, Hangzhou, China, 2013, pp. 71-82.

[14] R. X. Li, Z. Y. Xu, W. S. Kang, K. C. Yow, and C. Z. Xu, Efficient multi-keyword ranked query over encrypted data in cloud computing, Futur. Gener. Comp. Syst., vol. 30, pp. 179-190, Jan. 2014.

[15] Gurdeep Kaur, Poonam Nandal, “Ranking Algorithm of Web Documents using Ontology”, *IOSR Journal of Computer Engineering (IOSR-JCE) eISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. VIII (May-Jun. 2014), PP 52-55*