Secure and privacy preserving keyword search over the large scale cloud data

Wei Zhang[1], Jie Wu[2], Yaping Lin[1]
1. Hunan Univeristy, China 2. Temple University, USA
Email: zhangweidoc@hnu.edu.cn

ABSTRACT

Cloud computing has attracted a lot of interests from both the academics and the industries, since it provides efficient resource management, economical cost, and fast deployment. However, concerns on security and privacy become the main obstacle for the large scale application of cloud computing. Encryption would be an alternative way to relief the concern. However, data encryption makes efficient data utilization a challenging problem. To address this problem, secure and privacy preserving keyword search over large scale cloud data is proposed and widely developed. In this paper, we make a thorough survey on the secure and privacy preserving keyword search over large scale cloud data. We investigate existing research arts category by category, where the category is classified according to the search functionality. In each category, we first elaborate on the key idea of existing research works, then we conclude some open and interesting problems.

KEYWORDS

Privacy preservation, security, keyword search, cloud computing, suvey.

INTRODUCTION

Cloud computing has attracted a lot of interests from both the academics and the industries, since it provides efficient resource management, economical cost, and fast deployment (Armbrust, et al, 2010). For both personal and enterprise users, the cloud computing creates a lot of opportunities for them to enjoy innovation, collaboration, and convenience. Due to the huge potential economical benefits of cloud computing, a lot of companies have deployed their cloud centers. For example, the Elastic Compute Cloud (EC2) of Amazon, the App Engine of Google, the Azure of Microsoft, and Blue Cloud of IBM.

Although the cloud computing has a lot of benefits, both individual and enterprise users are not willing to outsource their sensitive data (e.g., personal health records, financial records) to the cloud server. Because once these data are outsourced to the cloud server, the corresponding data owners will lose direct control over these data. The Cloud Service Provider (CSP) would promise that they can preserve the security of these data by using techniques like fireware, virtualization, and Intrusion Detection System(IDS). However, since the CSP takes full control of these data, these techniques cannot prevent employers of the CSP from revealing sensitive data. Encryption would be an alternative way to solve the problem. However, data encryption makes the traditional plaintext based search schemes impractical. A probable solution is downloading all these encrypted files and decrypting them locally to find the desired files. However, this is obviously unrealistic, since it would cause unacceptable communication and computation cost for the end users, whose communication and computation capabilities are often constrained. Therefore, devising a secure and privacy search scheme over encrypted cloud data would be grateful.

To address this problem, secure and privacy preserving keyword search over large scale cloud data is proposed and widely developed. A secure search system often includes three entities, i.e., data owner, cloud server, and data users. The data owner outsources encrypted files and indexes to the cloud server. Authorized data users generate secret trapdoors and submit them to the cloud server. The cloud server further returns the search results without knowing sensitive data. We can categorize existing secure search schemes based on different criterions.

First, based on the search functionality, existing researches include: secure conjunctive keyword search, secure ranked keyword search, secure fuzzy keyword search, and privacy preserving similarity keyword search.

Second, based on the adopted encryption method, these researches can be categorized into: symmetric encryption based schemes and asymmetric encryption based schemes. The symmetric encryption based schemes often achieve high efficiency while the asymmetric encryption based schemes achieve strong security.

Third, based on the threat model, existing research works consider two different models, one assumes the cloud server to be “curious but honest”, i.e., the cloud server will follow the proposed schemes, but they will try to reveal the sensitive data of both the data owner and data users. The other one assumes the cloud server to be “dishonest” (Zhang, et al, 2015), i.e., the cloud server would probably return false retrieval result, therefore, the corresponding research works seek to verify the retrieval results.

Forth, based on the number of data owner involved in the search system, existing research works are divided into single owner model and multi-owner model. In comparison, the multi-owner model would be more adapted to be deployed in reality.

Fifth, based on the number of cloud server, there are two kinds of research works, i.e., secure keyword search on the centralized cloud server, and secure keyword search among distributed cloud servers.

Finally, based on the design goals, different researches have different concerns: preserving data confidentiality and data privacy, achieving low computation and communication cost, hiding data access pattern (preventing the cloud server from knowing which data are actually returned), enabling scalability, and allowing data updating.

In this chapter, we make a thorough survey on secure and privacy preserving keyword search over large scale cloud data. We investigate research works category by category, where the category is classified according to the search functionality. In each category, we first elaborate on the key idea of existing research arts, then we conclude some open and interesting problems.

SEARCHABLE ENCRYPTION

Searchable encryption allows a client to outsource his data to a server secretly, while providing interface for different clients to search these secret data. Searchable encryption has attracted many researchers since it is proposed. We also note that, some searchable encryption schemes inspire the secure search schemes over large scale cloud data. In this section, we give an overview of the searchable encryption schemes.

Song et al. (2000) firstly propose a scheme for searches on encrypted data, where each word is encrypted independently under a special two-layered encryption construction. Their approach allows the server to find all the positions where a single query word occurs in a file. Both the files on the server and the query word are encrypted during the searching process. The proposed scheme prevents the server from deducing any information about the contents of the files and the words. Data user’s query privacy are preserved.

Goh (2003) propose to use Bloom filters to speed up the keyword search. With the help of the bloom filter, it is very easy to test whether a specific keyword is contained in a file. Specifically, the data owner constructs a bloom filter for each file, where all the possible trapdoors for a file are mapped into the bloom filter. Upon receiving a trapdoor from the data user, by checking whether all the positions of a bloom filter indicated by the hash of the trapdoor are 1, the server can easily check whether the trapdoor matches a file. Therefore the presence or absence of a keyword in a file can be tested in constant time. The server further returns the corresponding files when a keyword is matched.

Chang and Mitzenmacher (2005) also propose to build index on the keyword dictionary before outsourcing the files and keyword dictionary to the server. To stop the server from knowing sensitive data, they adopt pseudo-random bits to mask the node in the index. When an authorized data user wants to search over the index, he will send a short seed to help server recover the selective part without revealing other parts to the server.

Curtmola et al. (2006) further propose an inverted-list traversal based scheme. In their scheme, they construct an encrypted hash table index for the whole file set. Specifically, the table is composed of a series of inverted keyword lists, where the head of each list corresponds to a keyword. The order of these lists are randomly permuted. Therefore, only the authorized user can search and get the header of the linked list, and obtain the whole list by decrypting nodes in the list iteratively. On the contrary, without knowing the header node, it is computationally infeasible to derive any useful information. Obviously, the security can be achieved since the server does not know the header node. In their scheme, they also consider a setting where one data owner and multiple data users are involved. The data owner controls to grant or revoke searching capabilities to data users efficiently.

Boneh et al. (2004) present the first public key scheme for keyword search over encrypted data. In their constructions, a sender encrypts the keywords with the public key of the receiver, then the receiver can distribute the server a key that enables the server to test whether a specific keyword is involved in the encrypted keyword set without revealing any other information to the server.

Chai and Gong (2012) propose to address the problem of verifiable secure search over encrypted data in cloud computing. Since the cloud server would probably return forged or incomplete search results, they define a rationale called verifiable searchability. Their scheme achieves privacy, verifiability, and efficiency. End users with constrained resources can also decrypt and verify search results efficiently.

Fang et al. (2013) point out that existed public encryption with keyword search(PEKS) can only defend the keyword guessing attack under the random oracle model. They define the strongest security model which is secure channel free and secure against keyword guessing attack, chosen keyword attack, and chosen ciphertext attack. They also propose a secure channel free PEKS scheme which is secure without requiring random oracles.

Remark: Most of these searchable encryption schemes are concerned mostly with single or boolean keyword search. Extending them to large scale cloud data will incur heavy computation and storage cost. Compared with the searchable symmetric encryption schemes, the public key based solutions usually requires relatively more computational overhead. Additionally, almost all the searchable symmetric encryption schemes expose the relationship between the file ID and the trapdoor, while the public key based schemes are easy to preserve this leakage by introducing randomness in the trapdoor generation. Readers can also refer to (Bosch, et al, 2014) for more detailed introduction on the provable secure searchable encryption.

RANKED SINGLE KEYWORD SEARCH

The ranked single keyword search is described as follows, given a keyword, the cloud server returns the most relevant files to the keyword. Compared with returning all undifferentiated files that match a keyword, the ranked keyword search has two main advantages. First, it saves numerous communication cost for the search system. Second, it provides more excellent user experience. In this section, we investigate the existed researches on ranked single keyword search.

Wang et al. (2010) first define and solve the secure ranked keyword search over encrypted cloud data. Due to the fact that large number of data files are stored on the cloud, a ranked keyword search only returns the most relevant *k* files. Authors first weaken their security guarantees, and then derive an efficient one-to-many order-preserving mapping function. Therefore, a relevance score would be mapped to different encoded values, while the corresponding order is still preserved. With this smart property, the cloud server can rank the files according to the order of the encoded relevance scores without knowing the actual data of relevance scores.

Ananthi et al. (2011) propose a secure ranked keyword search scheme. The data owner encrypts the search indexes and files before outsourcing them to the cloud. The data user generates the encrypted search request with the data owner's secret key. The cloud server ranks the search results based on the keyword frequency occurred in each file. However, the frequency of the keyword in each file is revealed to the cloud server. Once the cloud server knows some background information, he can deduce the actual value of some keywords. On the other hand, taking keyword frequency as the unique ranking metric would bring inaccuracy for the ranking.

Remark: Ranked single keyword search is the foundation of ranked multi-keyword search. A practical search system should enable the capability for ranked multi-keyword search instead of only supporting just a ranked single keyword search.

MULTI-KEYWORD SEARCH

Compared with single keyword search, multi-keyword search would be more practical. A multi-keyword search allows data users to input more than one keyword, which describes user's search request more accurately. In this section, we first review the art of ranked multi-keyword search, then we review the multi-keyword search without ranking.

1) Ranked Multi-keyword Search

The ranked multi-keyword search allows data users to submit a search request with multiple keywords, which enables them to search the most relevant files corresponding to their search request.

Cao et al. (2011), and Cao et al. (2014) first seek to solve the challenging privacy-preserving multi-keyword ranked search problem in cloud computing. First of all, they choose the “coordinate matching” as the principle to evaluate the similarity between files and search requests. Then they propose to use the “inner product similarity” of file vectors (where each bit denotes whether the corresponding keyword exists in the file or not) and query vectors to quantify the similarity. Further, they design to protect the privacy of file vectors, query vectors, and the score of similarities. Consequently, their schemes allow the cloud to find top-k relevant files corresponding to a multi-keyword query. Both the files stored in the cloud and the multi-keyword query are encrypted throughout the search process. The cloud cannot know the plaintext of the query or the contents of files stored in the cloud. data owners’ data privacy and data users' query privacy are both preserved. This paper attracts the interests of many researchers, they find that this paper suffers from three drawbacks (Xu, et al, 2012). First, this scheme does not support keyword update. If data owners want to insert new keywords into the search, the entire encrypted keyword dictionary has to be rebuilt. Second, the trapdoor generation algorithm proposed in this scheme brings an out-of-order problem, i.e., files with more matching keywords could be ranked lower than files with less matching keywords but some keywords have been matched many times. Third, the time and storage complexity of the scheme needs to be further improved.

Xu et al. (2012), Li et al. (2014) propose MKQE (Multi-Keyword ranked Query on Encrypted data). First of all, they propose the idea of partitioned matrices. With the design of partitioned matrices, the expansion of keyword dictionary can be dynamically and efficiently achieved without touching the contents in the original dictionary. Therefore, their scheme enjoys wonderful scalability and flexibility. To avoid the out-of-order problem appears in (Cao, et al, 2011), they design a novel trapdoor generation algorithm, which can effectively reduce the impacts of dummy keywords on the ranking scores. Furthermore, they take the keyword access frequencies and keyword weights in the index file into consideration when computing the ranking scores. Consequently, files containing more frequently accessed keywords and higher weighted keywords would have higher probabilities to be returned, i.e., the data users have higher probabilities to retrieve their desired files.

To simultaneously achieve accurate result ranking, efficient multi-keyword search, and results verification, Sun et al. (2014) introduce a verifiable and privacy-preserving multi-keyword text search (MTS) scheme with similarity ranking functionality. For accurate result ranking, they propose to use the vector model, where each item is a TF×IDF weight. Therefore, the vector product of the search vector and the file vector are used to measure the similarity between the search request and each file. The search results are further ranked based on these similarity scores. For search efficiency, they propose to divide each vector into sub-vectors and build a tree based index based on these sub-vectors. When an authorized data user wants to submit a search, he also constructs a preference search vector, which records his preference for each keyword. Then the data user adopts the same method to split the search vectors. The cloud server returns the top matched files based on the similarity scores. For result verification, they propose to sign the index tree with a method adapted from the Merkle hash tree (Merkle, 1989). Therefore, any forged search results can easily be detected.

Ibrahim et al. (2012) explore a ranked searchable encryption scheme of multi-keyword queries over cloud data, which is based on information retrieval systems and cryptography approaches. In order to avoid of leaking access pattern (the association between documents and keywords), encrypted keywords and files are outsourced to two different cloud servers, which effectively hides the association between them. To prevent the statistical inference attacks, they use the probabilistic asymmetric Paillier cryptosystem to implement the encryption function. Further, they use the privacy preserving mapping (PPM) (Tang, 2010) to preserve the relevance score between each document and keyword.

Hore et al. (2012) focuses on creating an index I to enable the server to efficiently retrieve files against a query that is issued by the user. The index of an encrypted file is a set of colors which encode the presence of the keywords while not giving out their exact identities. The server can search the encrypted files by their color codes. Specifically, given a query keyword, a user first computes its color code and sends it to the server. The server then determines all files whose index entries have at least one of the specified colors and returns them back.

Xu et al. (2012) propose to use a two-step-ranking method to achieve secure ranked multi-keyword search over encrypted cloud data. In their scheme, they adopt the order preserving encryption (OPE) technique to encoded relevance scores between keywords and files. For the first step, they divide files into groups through coordinate matching, as a result, files with the same number of searched keywords, whose relevance scores are the top-k highest, would be classified into the same group, and all files in the group with higher number would be ranked higher than those in the group with lower number. For the second step, they rank files in each group based on the summation of encoded relevance scores of the searched keywords.

Orencik and Savas (2012) also propose to solve the secure ranked multi-keyword search over encrypted cloud data. In their scheme, only authorized data users can get the secret keys for trapdoors. Upon receiving data users' search request, the cloud server can return the top matches with a ranking function. They also propose to obfuscate the searched files so that other parties cannot know the search results.

Orencik et al. (2013) propose to deal with the problem of secure multi-keyword search in cloud computing. They first use the minhash (Leskovec, 2014) technique to construct the signatures of documents. Then each document is mapped to different buffers based on its signatures. Further, each document is attached with an encrypted relevance score between itself and the signature (buffer). Once a data user wants to issue a multi-keyword search, he generates the query signatures with minhash. Upon receiving the search requests, the cloud server finds the documents in buffers corresponding to the query signatures, and sends the ID and encrypted relevance scores of these documents to the data user. The data user further decrypts the relevance scores and ranks the results based on the summation of the corresponding relevance scores. However, the decryption of the relevance scores and the ranking are done by the data user, which would cause great burden for the resource limited data users. Although they propose to use two server model to solve this problem, i.e., one server is in charge of searching, and the other is responsible for decrypting and ranking the search results. Once the two cloud servers collude with each other, all sensitive data are revealed to the cloud servers.

Shen et al. (2013) depict a preferred keyword search over encrypted data. They first use the occurrence times of a keyword to denote its weight in a file vector. Then they propose to transform a search request to a polynomial form, adopt the Lagrange polynomial to denote data user's preference, and transform the preference polynomial into a search vector. Finally, they use the secure inner product of file vector and search vector to indicate the relevance between files and search request.

Yu et al. (2013) address the problem of privacy preserving multi-keyword top-k retrieval over encrypted cloud data. They first present their observation, i.e., ranking search results on the cloud server inevitably reveals data privacy. Therefore, they propose a two-round searchable encryption(TRSE) to support the top-k multi-keyword search. In TRSE, they also adopt the vector space model to ensure accurate ranking for the search results. Specifically, the file vector records the relevance score between keywords and files, the search vector records the search preference of data users. The vector product of these two vectors indicate their similarity. To preserve the privacy of the similarity, they propose to adopt the homomorphic encryption, which helps the cloud server compute the similarity without knowing the actual value of the similarity. Upon receiving all the encrypted similarity scores from the cloud server, the data user decrypts and ranks these scores, and then launches a second round file retrieval, by submitting the corresponding file IDs.

Fu et al. (2014) propose to solve the ranked multi-keyword search over encrypted cloud data supporting synonym query. Similar to (Cao, 2011), they also use the vector model to retrieve the top-k ranked search results. To achieve the synonym-based keyword search, they construct a common synonym thesaurus based on the foundation of the New American Roget’s College Thesaurus (NARCT) (Morehead, 2002). Basically, they enlarge the keyword dictionary, any synonym query defined in the dictionary can obtain the corresponding results.

Zhang et al. proposed to ensure secure ranked multi-keyword search while supporting multiple data owners in (Zhang, et al, 2014), (Zhang, et al, 2015). In their scheme, different data owners use their own (potentially different) secret keys to encrypt both the documents and keywords. Any authenticated data users can submit an encoded multi-keyword search request without knowing different data owners' secret keys. The cloud server achieves the secure multi-keyword search without knowing the actual data of keywords, files, and users' search requests. To rank the search results while preserving the privacy of relevance scores between keywords and files, they propose a novel encoding scheme called additive order and privacy preserving function family, which enables the cloud server to return the most relevant search results to data users without divulging any sensitive information.

2) Multi-keyword Search without Ranking

Liu et al. (2012) propose a privacy preserving COoperative Private Searching (COPS) protocol. They introduce a middleware layer called the Aggregation and Distribution Layer (ADL) to combine queries from data users and divide search results to corresponding data users. The COPS saves both computation and communication cost at the expense of some search delays. Particularly, the COPS adopts the Paillier encryption to encode the search requests and the indexes of files. With the homomorphic property of Paillier encryption, the cloud server only executes operations on cipher-texts. Consequently, the cloud server returns the search results without knowing which files are actually returned, therefore, the access pattern is well preserved.

Zhang et al. proposed to achieve secure distributed keyword search in geo-distributed clouds in (Zhang, et al, 2014). To overcome essential problems existed in the centralized cloud model, e.g., single point of failure, loss of availability, they consider the search problem in a distributed model, where multiple cloud servers are involved. Their proposed schemes achieve secure and efficient distributed multi-keyword search, while the robust, availability, and usability of the search system are maximized.

Remark:Although a lot of research works put their efforts on proposing an applicable multi-keyword search scheme, some problems remain to be solved.

First, Most of these researches do not preserve the access pattern(list of returned files), which should be considered to defend statistical attacks(e.g., the cloud server would deducing vital data by observing which files are most frequently returned). Therefore, ensuring secure ranked keyword search while preserving access pattern is very important. An outstanding research which enables ranked multi-dimensional query with access preservation is presented in (Elmehdwi, et al, 2014). In this paper, authors aim to deal with the secure k-nearest neighbor query over encrypted cloud data. They assume two non-collusive cloud servers are involved. The techniques used in this paper include secure multi-party computation (Goldreich, 2004), e.g., Secure Multiplication (SM) Protocol, and Paillier Encryption (Paillier, 1999). As a result, the cloud server can only see new random numbers or newly generated random encryptions. The cloud server makes comparison on two encrypted items without knowing which one is really larger, since the larger of the two is randomly generated on cipher-text. The cloud server also returns search results to authorized data users without knowing which entries are actually returned. However, as shown in the performance evaluation, preserving access pattern requires considerable computation overhead. We need to come up with a tradeoff between security and efficiency.

Second, most of existing ranked multi-keyword search schemes do not support efficient incremental updates. Additionally, many schemes need to pre-define a keyword dictionary. This property, to some extent, also affects efficient data update. However, a flexible multi-keyword search scheme should achieve a dynamic update on both files and keywords.

Third, most of these systems only support one data owner. To deploy the search system into reality, it should support multiple data owners. When multiple data owners are involved, how to distribute and manage secret keys for different data owners, how to efficiently generate trapdoors for different data owners' data, and how to achieve decryption capability are very challenging.

FUZZY KEYWORD SEARCH

To tolerant minor type errors and format inconsistencies, fuzzy keyword search is proposed. The fuzzy keyword search not only improves the robustness, but also increases the usability of the search system. In this section, we review the researches of fuzzy keyword search.

Li et al. (2010) first propose to solve the secure and efficient fuzzy keyword search over encrypted data in cloud computing. To tolerate minor type errors and format inconsistencies, they propose to construct a fuzzy keyword set, which contains all the probable fuzzy keywords. Specifically, they use the edit distance to indicate the distance between a fuzzy keyword and a predefined keyword. When the distance is smaller than a predefined threshold, the fuzzy keyword is added to the fuzzy keyword set. Additionally, instead of enumerating all possible fuzzy keywords, they develop two novel and storage efficient techniques, i.e., a wildcard based and a gram based technique. Upon receiving data user's search request, the cloud server returns the exactly matched files when the searched keyword exactly matches the predefined keywords, or the closest possible files corresponding to the keywords in the fuzzy keyword set. To speed up the search process, they also propose a symbol-based trie-traverse searching method, which constructs a multi-way tree by using symbols transformed from the fuzzy keyword set. The proposed scheme has two limitations. First, they have to construct a very large fuzzy keyword set, this is very computationally intensive. Second, it does not support efficient multi-keyword fuzzy search.

Wang et al. (2012) propose to verify the search results of the secure fuzzy keyword search. To construct the fuzzy keyword set, they also adopt the wildcard based technique. To ensure efficient fuzzy keyword search, they use the multi-way tree to store the fuzzy keyword, where the symbols concatenated from the root node to the leaf node form a specific keyword. To verify the search results, each internal node is accompanied by a signed bloom filter, which records the prefixes of all of its children. As a result, when the cloud server return forged or incomplete search results, the authorized data user will soon detect inconsistency from the signed bloom filters or the signature of the leaf nodes.

Chuah et al. (2011) propose to achieve a fuzzy multi-keyword search, which also assures data update, with a solution based on a privacy-aware bedtree. First, they propose to extract a predefined keyword set from a list of files, and identify useful multi-keywords with a co-occurrence probability scheme. Then they construct a fuzzy keyword set with a predefined edit distance. Next, they accompany each predefined keyword with several bloom filters, where all the fuzzy keywords with a specified edit distance of a predefined keyword are mapped. Further, they construct an index tree for all the files, where the leaf node is composed of a hash value of keyword, one or two data vectors, and some bloom filters. With this index tree, any new data can be easily inserted. Once an authorized data user wants to search, he submits the hash value of keywords, the edit distance, and a list of hash value of the corresponding fuzzy keywords to the cloud server. The cloud server searches the index tree, finds the matched keywords, and returns the corresponding encrypted files.

Wang et al. (2014) address the problem of multi-keyword fuzzy search over the encrypted cloud data. Instead of performing single fuzzy search for multiple rounds, schemes proposed in this work achieve multi-keyword fuzzy search efficiently and simultaneously. Specifically, they first transform each keyword into a bigram vector. Then they adopt the LSH function to substitute the normal hash function to map the bigram vector into bloom filters, as a result, the misspelling keyword and the original keyword (correspondingly, the two bigram vectors with small Euclidean distance) are mapped to the same position of the bloom filter. The data user will use the same method to convert his search request to a bloom filter. Finally, to rank the search results, they propose to use the secure inner production technique on two encoded bloom filters to measure the similarity between the search request and the index. Schemes proposed in this work not only ensure privacy preserving and efficient multi-keyword fuzzy search, but also eliminate the requirement of a predefined keyword dictionary, which supports frequent keyword and file update. However, if the proposed scheme is extended to allow the keyword to contain numbers or special characters, the bigram vector would be very large.

Remark: Existing researches concerned with fuzzy keyword search only help search the files that contain the original keyword of a probable misspelling keyword. They do not consider how to rank the search result (Wang, et al, 2014) makes an attempt, but it does not consider the relevance between the original keyword and the files, which may lead to some incorrect ranking.

CONJUNCTIVE KEYWORD SEARCH

A conjunctive keyword search is described as follows, when a data user has several keywords, he first generates the trapdoor for each keyword. The search results are the intersection of search results of each keyword. In this section, we review the existing conjunctive keyword search schemes.

Ballard et al. (2005) aim to solve the secure conjunctive keyword search over encrypted data. They propose two schemes, one is based on the Shamir Secret Sharing (Shamir, 1979), which achieves high efficiency. But the trapdoor size of this scheme increases linearly with the number of files increases. The other scheme is based on bilinear pairings, which is also very efficient. Both schemes achieve provable security in the standard model.

Golle et al. (2004) propose to solve the secure conjunctive keyword search over encrypted data. They construct two public key based scheme, both schemes achieve a defined security. The second scheme only requires constant communication cost, while the communication cost for the first scheme is linear with the number of files.

Boneh and Waters (2007) depict a secure framework for constructing and analyzing public key systems for various searches over encrypted data. Based on the Hidden Vector Encryption(HVE), they construct a public key based search system that support any conjunctive searches, where no individual conjunction is revealed.

Most of existing researches assume a pre-known keyword set, which is impractical in many applications. To relief this assumption, Wang et al. (2008) propose a keyword field-free conjunctive keyword search scheme over encrypted data. They also extend their scheme to a dynamic group setting. Rigorous security proof and analysis are also presented.

Cai et al. (2013) identify a security threat in the secure conjunctive keyword search, i.e., IR attack, where the cloud server can deduce an inclusion relation(IR) by observing the relationship between the trapdoor and search results. To defend this attack, they propose to use bloom filters and probability numbers to build secure index, and integrate randomness into the trapdoors, which ensures the cloud server cannot know the actual relationship between different queries and the search results, correspondingly, the cloud server cannot deduce an inclusion relation.

Wang et al. (2015) propose a scheme that protects search pattern and supports efficient conjunctive multi-keyword search based on the inverted index. They also design an efficient oblivious transfer protocol to prevent the cloud from knowing the access pattern.

In (Sun, et al, 2015), Sun et al. construct a cryptographic design, which supports the secure conjunctive keyword search, efficient file collection update, and search results verification.

Remark: Most of existing researches assume to construct a keyword dictionary in advance, it would be very interesting if we can relief this assumption. Additionally, if the conjunctive search scheme can support multiple data owners to share their data, it would be very attracting.

SIMILARITY KEYWORD SEARCH

A similarity keyword search asks the cloud server to return all possible files that are similar (possibly matched) with data user' search requests, which has many applications nowadays. Due to the security and privacy obstacles, devising a secure similarity keyword search protocol is challenging.

Park et al. (2007) propose to solve secure similarity search over encrypted data. In their scheme, they propose to encrypt each character in a keyword. To prevent a dictionary attack, they propose to form a large object by padding each character with data user's secret key, file ID, and field ID. Then they encrypt the large object, which not only stops the dictionary attack, but also achieves 'Cell Privacy', i.e., the encryption of a character differs from cell to cell. To prevent the server from knowing the relationship between trapdoors, they propose to introduce randomness during the trapdoor generation. Based on these techniques, they propose two similarity searchable schemes. The first scheme achieves perfect similarity search privacy, while the second scheme ensures high efficiency at the expense of security guarantee.

Wang et al. (2012) propose a solution for privacy preserving similarity search in cloud computing. To construct a storage efficient similarity keyword set for a given file collection, they adopt the edit distance to measure the similarity between keywords, and use a suppressing technique. Consequently, all the keywords within a specific edit distance are put into a similarity keyword set. Then, they encrypt all the keywords in the similarity set and files, and outsource them to the cloud. When an authorized data user wants to perform a similarity keyword search, the data user first generates a similarity keyword set, encrypts them and submits them to the cloud. The cloud searches all the files according to the received encrypted keywords set and returns the corresponding result set. The data user further decrypts and obtains the desired files. To improve the search efficiency on the cloud, they also build a private trie-traverse searching index, where a multi-way tree is constructed for storing the similarity keyword elements. All similar words in the trie-tree can be found by a depth-first search. The scheme proposed in this paper also supports a fuzzy keyword search.

Kuzu et al. (2012) also propose to solve the secure similarity search over encrypted data. To achieve efficient search over encrypted data, they first propose to construct a secure index based on a novel technique called Locality sensitive hashing(LSH), which is widely used for fast plain-text based similarity search. Then they propose the similarity searchable symmetric encryption scheme based on the secure index. As a result, the server can efficiently perform secure similarity search on the secure index without knowing any sensitive data. To illustrate how their theoretical schemes work in reality, they give a real application, i.e., error tolerant keyword search over encrypted data. To prevent the server from knowing the relationship between the file ID and search request, they propose to involve multiple non-collusive servers.

Remark: Secure similarity search attracts many interests recently. But we still need to put some efforts before it can be deployed into reality. A practical similarity search scheme should consider the efficiency, security, and robustness. Moreover, it should take user preference into consideration, which helps achieve personalized search service.

ATTRIBUTE BASED KEYWORD SEARCH

A practical secure search scheme should support multiple data owners and multiple data users before being deployed. However, most of existing schemes do not support multiple data owners to securely and efficiently share data. The attribute based keyword search scheme is a hopeful candidate to support multiple data owners to share data in the large scale cloud computing.

Xhafa et al. (2014) seek to achieve fine grained access control for PHR data and efficient PHR data retrieval in a hybrid cloud computing. To ensure privacy preserving and fine grained access control, they adopt an anonymous attributed based encryption scheme to encrypt the secret key used to encrypt the sensitive PHR data. As a result, when a data user's attributes do not meet the access policy, the data user cannot decrypt the cipher-text, or know the access policy. To ensure stronger usability, they adapt their scheme to support fuzzy keyword set. Finally, they also use a symbol based trie-traverse search scheme to speedup the search process.

Zheng et al. (2014) propose a verifiable attribute-based keyword search (VABKS) scheme over outsourced encrypted cloud data. They make full use of techniques including attribute-based encryption, digital signature, and bloom filter. They also propose a novel technique called attribute-based keyword search (ABKS). As a result, by setting the access control policy, the proposed scheme allows the data owner to control the search capability of its outsourced data. The authorized data users can also outsource the search operation to the cloud server, and then verify whether the cloud server returns the search results faithfully. However, this paper only supports a static data set, when the outsourced data are frequently updated, new schemes need to be developed.

Sun et al. (2014) propose an attribute-based keyword search scheme with user revocation. The scheme ensures fine-grained search authorization, which is secure even if numerous data owners are involved. Data users with the corresponding attributes(secret keys) can launch a secure search without depending on an always online trusted authority. To achieve efficient data update for user revocation, they propose to combine proxy re-encryption with lazy re-encryption techniques to delegate the computationally intensive workload of data update to the semi-trusted cloud server. Authors further prove that their scheme is selectively secure against the chosen-keyword attack.

Remark: Attributed based keyword search schemes usually support multiple data owners to share data securely. An interesting problem is how to achieve secure and efficient ranked multi-keyword search in the multiple owner paradigm. Can we incorporate the attribute based keyword search scheme with a ranked multi-keyword search scheme? Another problem is how to achieve efficient data owner or data user registration or revocation? We need to put more efforts on discovering and solving potential problems before we can deploy these schemes in reality.

CONCLUSION

In this chapter, we survey outstanding researches of secure and privacy preserving keyword search over large scale cloud data. Specifically, we first classify existing research arts into categories based on the search functionalities. Then we investigate these works category by category. For each category, we first elaborate on the key idea of these research works, then we conclude some open and interesting problems. We also need to put more efforts on discovering and solving potential problems before we deploy existing search schemes in reality.

REFERENCES

Ananthi, S., Sendil, M. S., & Karthik, S. (2011). Privacy preserving keyword search over encrypted cloud data. In *Advances in Computing and Communications* (pp. 480-487). Springer Berlin Heidelberg.

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM*, *53*(4), 50-58.

Ballard, L., Kamara, S., & Monrose, F. (2005). Achieving efficient conjunctive keyword searches over encrypted data. In *Information and Communications Security* (pp. 414-426). Springer Berlin Heidelberg.

Boneh, D., Di Crescenzo, G., Ostrovsky, R., & Persiano, G. (2004, January). Public key encryption with keyword search. In *Advances in Cryptology-Eurocrypt 2004* (pp. 506-522). Springer Berlin Heidelberg.

Boneh, D., & Waters, B. (2007). Conjunctive, subset, and range queries on encrypted data. In *Theory of cryptography* (pp. 535-554). Springer Berlin Heidelberg.

Bösch, C., Hartel, P., Jonker, W., & Peter, A. (2014). A survey of provably secure searchable encryption. *ACM Computing Surveys (CSUR)*, *47*(2), 18.

Cai, K., Hong, C., Zhang, M., Feng, D., & Lv, Z. (2013, December). A Secure Conjunctive Keywords Search over Encrypted Cloud Data Against Inclusion-Relation Attack. In *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on* (Vol. 1, pp. 339-346). IEEE.

Cao, N., Wang, C., Li, M., Ren, K., & Lou, W. (2014). Privacy-preserving multi-keyword ranked search over encrypted cloud data. In *INFOCOM, 2014 Proceedings IEEE* (pp. 829-837). IEEE.

Cao, N., Wang, C., Li, M., Ren, K., & Lou, W. (2014). Privacy-preserving multi-keyword ranked search over encrypted cloud data. *Parallel and Distributed Systems, IEEE Transactions on*, *25*(1), 222-233.

Chai, Q., & Gong, G. (2012, June). Verifiable symmetric searchable encryption for semi-honest-but-curious cloud servers. In *Communications (ICC), 2012 IEEE International Conference on* (pp. 917-922). IEEE.

Chang, Y. C., & Mitzenmacher, M. (2005, January). Privacy preserving keyword searches on remote encrypted data. In *Applied Cryptography and Network Security* (pp. 442-455). Springer Berlin Heidelberg.

Chuah, M., & Hu, W. (2011, June). Privacy-aware bedtree based solution for fuzzy multi-keyword search over encrypted data. In *Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on* (pp. 273-281). IEEE.

Curtmola, R., Garay, J., Kamara, S., & Ostrovsky, R. (2006, October). Searchable symmetric encryption: improved definitions and efficient constructions. In *Proceedings of the 13th ACM conference on Computer and communications security* (pp. 79-88). ACM.

Elmehdwi, Y., Samanthula, B. K., & Jiang, W. (2014, March). Secure k-nearest neighbor query over encrypted data in outsourced environments. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on* (pp. 664-675). IEEE.

Fang, L., Susilo, W., Ge, C., & Wang, J. (2013). Public key encryption with keyword search secure against keyword guessing attacks without random oracle. *Information Sciences*, *238*, 221-241.

Fu, Z., Sun, X., Linge, N., & Zhou, L. (2014). Achieving effective cloud search services: multi-keyword ranked search over encrypted cloud data supporting synonym query. *Consumer Electronics, IEEE Transactions on*, *60*(1), 164-172.

Goh, E. J. (2003). Secure Indexes. *IACR Cryptology ePrint Archive*, *2003*, 216.

Goldreich, O. (2004). *Foundations of cryptography: volume 2, basic applications*. Cambridge university press.

Golle, P., Staddon, J., & Waters, B. (2004, January). Secure conjunctive keyword search over encrypted data. In *Applied Cryptography and Network Security* (pp. 31-45). Springer Berlin Heidelberg.

Hore, B., Chang, E. C., Diallo, M. H., & Mehrotra, S. (2012). Indexing encrypted documents for supporting efficient keyword search. In *Secure Data Management* (pp. 93-110). Springer Berlin Heidelberg.

Ibrahim, A., Jin, H., Yassin, A., & Zou, D. (2012, December). Secure rank-ordered search of multi-keyword trapdoor over encrypted cloud data. In *Services Computing Conference (APSCC), 2012 IEEE Asia-Pacific* (pp. 263-270). IEEE.

Kuzu, M., Islam, M. S., & Kantarcioglu, M. (2012, April). Efficient similarity search over encrypted data. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on* (pp. 1156-1167). IEEE.

Rajaraman, A., & Ullman, J. D. (2012). *Mining of massive datasets* (Vol. 77). Cambridge: Cambridge University Press.

Li, J., Wang, Q., Wang, C., Cao, N., Ren, K., & Lou, W. (2010, March). Fuzzy keyword search over encrypted data in cloud computing. In *INFOCOM, 2010 Proceedings IEEE* (pp. 1-5). IEEE.

Li, R., Xu, Z., Kang, W., Yow, K. C., & Xu, C. Z. (2014). Efficient multi-keyword ranked query over encrypted data in cloud computing. *Future Generation Computer Systems*, *30*, 179-190.

Liu, Q., Tan, C. C., Wu, J., & Wang, G. (2012). Cooperative private searching in clouds. *Journal of Parallel and Distributed Computing*, *72*(8), 1019-1031.

Merkle, R. C. (1990, January). A certified digital signature. In *Advances in Cryptology—CRYPTO'89 Proceedings* (pp. 218-238). Springer New York.

Morehead, P. D. (2002). *New American Roget's College Thesaurus in Dictionary Form (Revised &Updated)*. Penguin.

Orencik, C., Kantarcioglu, M., & Savas, E. (2013, June). A practical and secure multi-keyword search method over encrypted cloud data. In *Cloud Computing (CLOUD), 2013 IEEE Sixth International Conference on* (pp. 390-397). IEEE.

Örencik, C., & Savaş, E. (2012, March). Efficient and secure ranked multi-keyword search on encrypted cloud data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops* (pp. 186-195). ACM.

Paillier, P. (1999, January). Public-key cryptosystems based on composite degree residuosity classes. In *Advances in cryptology—EUROCRYPT'99* (pp. 223-238). Springer Berlin Heidelberg.

Park, H. A., Kim, B. H., Lee, D. H., Chung, Y. D., & Zhan, J. (2007, November). Secure similarity search. In *Granular Computing, 2007. GRC 2007. IEEE International Conference on* (pp. 598-598). IEEE.

Shamir, A. (1979). How to share a secret. *Communications of the ACM*, *22*(11), 612-613.

Shen, Z., Shu, J., & Xue, W. (2013, June). Preferred keyword search over encrypted data in cloud computing. In *Quality of Service (IWQoS), 2013 IEEE/ACM 21st International Symposium on* (pp. 1-6). IEEE.

Song, D. X., Wagner, D., & Perrig, A. (2000). Practical techniques for searches on encrypted data. In *Security and Privacy, 2000. S&P 2000. Proceedings. 2000 IEEE Symposium on* (pp. 44-55). IEEE.

Sun, W., Liu, X., Lou, W., Hou, Y. T., & Li, H. (2015, April). Catch you if you lie to me: Efficient verifiable conjunctive keyword search over large dynamic encrypted cloud data. In *Computer Communications (INFOCOM), 2015 IEEE Conference on* (pp. 2110-2118). IEEE.

Sun, W., Wang, B., Cao, N., Li, M., Lou, W., Hou, Y. T., & Li, H. (2014). Verifiable privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking. *Parallel and Distributed Systems, IEEE Transactions on*, *25*(11), 3025-3035.

Sun, W., Yu, S., Lou, W., Hou, Y. T., & Li, H. (2014, April). Protecting your right: Attribute-based keyword search with fine-grained owner-enforced search authorization in the cloud. In *INFOCOM, 2014 Proceedings IEEE* (pp. 226-234). IEEE.

Tang, Q. (2010, October). Privacy preserving mapping schemes supporting comparison. In *Proceedings of the 2010 ACM workshop on Cloud computing security workshop* (pp. 53-58). ACM.

Wang, B., Song, W., Lou, W., & Hou, Y. T. Inverted Index Based Multi-Keyword Public-key Searchable Encryption with Strong Privacy Guarantee. In *INFOCOM, 2015 Proceedings IEEE* (pp. 2092-21110). IEEE.

Wang, B., Yu, S., Lou, W., & Hou, Y. T. (2014, April). Privacy-preserving multi-keyword fuzzy search over encrypted data in the cloud. In *INFOCOM, 2014 Proceedings IEEE* (pp. 2112-2120). IEEE.

Wang, C., Cao, N., Li, J., Ren, K., & Lou, W. (2010, June). Secure ranked keyword search over encrypted cloud data. In *Distributed Computing Systems (ICDCS), 2010 IEEE 30th International Conference on* (pp. 253-262). IEEE.

Wang, C., Ren, K., Yu, S., & Urs, K. M. R. (2012, March). Achieving usable and privacy-assured similarity search over outsourced cloud data. In *INFOCOM, 2012 Proceedings IEEE* (pp. 451-459). IEEE.

Wang, J., Ma, H., Tang, Q., Li, J., Zhu, H., Ma, S., & Chen, X. (2012). A new efficient verifiable fuzzy keyword search scheme. *Journal of Wireless Mobile Networks, Ubiquitous Computing and Dependable Applications*, *3*(4), 61-71.

Wang, P., Wang, H., & Pieprzyk, J. (2008). Keyword field-free conjunctive keyword searches on encrypted data and extension for dynamic groups. In *Cryptology and Network Security* (pp. 178-195). Springer Berlin Heidelberg.

Xhafa, F., Wang, J., Chen, X., Liu, J. K., Li, J., & Krause, P. (2014). An efficient PHR service system supporting fuzzy keyword search and fine-grained access control. *Soft Computing*, *18*(9), 1795-1802.

Xu, J., Zhang, W., Yang, C., Xu, J., & Yu, N. (2012, November). Two-step-ranking secure multi-keyword search over encrypted cloud data. In *Cloud and Service Computing (CSC), 2012 International Conference on* (pp. 124-130). IEEE.

Xu, Z., Kang, W., Li, R., Yow, K., & Xu, C. Z. (2012, December). Efficient multi-keyword ranked query on encrypted data in the cloud. In *Parallel and Distributed Systems (ICPADS), 2012 IEEE 18th International Conference on* (pp. 244-251). IEEE.

Yu, J., Lu, P., Zhu, Y., Xue, G., & Li, M. (2013). Toward secure multikeyword top-k retrieval over encrypted cloud data. *Dependable and Secure Computing, IEEE Transactions on*, *10*(4), 239-250.

Zhang, W., Lin, Y., Xiao, S., Liu, Q., & Zhou, T. (2014, May). Secure distributed keyword search in multiple clouds. In *Quality of Service (IWQoS), 2014 IEEE 22nd International Symposium of* (pp. 370-379). IEEE.

Zhang, W., Lin, Y., Xiao, S., Wu, J., & Zhou, S. (2015). Privacy preserving ranked multi-keyword search for multiple data owners in cloud computing. *Computers, IEEE Transactions on*.

Zhang, W., Lin, Y., & Gu, Q. (2015). Catch You if You Misbehave: Ranked Keyword Search Results Verification in Cloud Computing. *Cloud Computing, IEEE Transactions on*.

Zhang, W., Xiao, S., Lin, Y., Zhou, T., & Zhou, S. (2014, June). Secure ranked multi-keyword search for multiple data owners in cloud computing. In *Dependable Systems and Networks (DSN), 2014 44th Annual IEEE/IFIP International Conference on* (pp. 276-286). IEEE.

Zheng, Q., Xu, S., & Ateniese, G. (2014, April). Vabks: Verifiable attribute-based keyword search over outsourced encrypted data. In *INFOCOM, 2014 Proceedings IEEE* (pp. 522-530). IEEE.