ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Secure Fuzzy Multi-Keyword Ranked Search over Encrypted Cloud Data

Preethi Mathew, Dr. S. Sasidhar Babu

M. Tech Student, Dept. of CSE, SNGCE, Kolenchery, Kerala, India

Professor, Dept. of CSE, SNGCE, Kolenchery, Kerala, India

ABSTRACT: As cloud computing has become prevalent data owners tend to outsource their sensitive information to the cloud. The data have to be encrypted before outsourcing to protect the data privacy. Related works on searchable encryption support either single keyword search or Boolean keyword search which focuses on exact keyword. It does not support minor typos or format inconsistencies. Sorting of results is also rare. In this paper we propose secure fuzzy multi keyword ranked search over encrypted cloud data. This scheme allows multiple keyword in the search request and returns the documents in the order of their relevance to these keywords. A fuzzy keyword set is built from a predefined set of words based on edit distance so that it can return the matching files or closest possible matching files. Here coordinate matching is used based on secure inner product computation to measure similarity. The proposed scheme is secure and privacy preserving while introducing low overhead on computation and communication**.**

KEYWORDS: Cloud computing, fuzzy keyword search, ranked search, encryption

I. INTRODUCTION

In recent years cloud computing is gaining much popularity in IT industry. Cloud has virtually unlimited data storage capabilities and elastic resource provisioning. Both individuals and enterprises are motivated to outsource their data to the cloud storage server to reduce cost of management. To prevent unauthorized access in the cloud, sensitive data should be encrypted by data owners before outsourcing to the commercial public cloud [11]; thus makes traditional data utilization service based on plaintext keyword search unsuitable for cloud computing. Data encryption will make effective data utilization a very challenging task as there are large numbers of outsourced data files. Downloading all the data and decrypting locally is not practical, as it results in huge amount of bandwidth cost in cloud scale systems. Data encryption also demands the protection of keyword privacy since keywords usually contain important information related to the data files. Thus, exploring privacy preserving and effective search service over encrypted cloud data is important.

In cloud computing, data owners share their outsourced data with a number of authorized users. Keyword-based retrieval allows users to retrieve files they are interested in. Keyword-based retrieval is widely used in plaintext search schemes, in which user can retrieve relevant files based on the keyword in the search request. However, it is a difficult task in ciphertext scenario due to limited operations on encrypted data. The existing searchable encryption techniques allows performing searches securely and effectively but is not suitable in cloud computing scenario as they support only exact keyword search and does not support minor typos and format inconsistencies are not supported. Sometimes users searching input might not exactly match those pre set keywords due to the possible typos, representation inconsistencies and lack of exact knowledge about the data. Simple spell check mechanisms are used to support fuzzy keyword search. However, this approach will not completely solve the problem and sometimes can be ineffective as it requires additional interaction of user to determine the correct word from the candidates generated by the spell check algorithm, which costs extra computation effort for the users. If a user types some other valid keywords by mistake the spell check algorithm will not work because it cannot differentiate between two actual valid words. Due to these new techniques that has searching flexibility which support both minor typos and format inconsistencies is required. In this paper, we use edit distance to evaluate keywords similarity for the construction of fuzzy keyword sets and a search scheme based on this set.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Numerous searchable symmetric encryption schemes have been proposed to enable search on cipher text. However, these traditional schemes enable users to securely retrieve the cipher text, but they support only Boolean keyword search. Schemes presented in [10], [12], [13] show that they support top-k single keyword retrieval under various scenarios. The files should be ranked in the order of relevance by user's interest and only the files with the highest relevance are sent back to users. Ranked search can eliminate unnecessary network traffic by returning only the most relevant data. This ranking operation should not leak any keyword related information for privacy protection. To improve the search result accuracy and to enhance the user searching experience, it is necessary for such ranking system to support multiple keywords search, as single keyword search often yields coarse results. Data users may provide a set of keywords to indicate their search interest in order to retrieve the most relevant data. Each keyword in the search request helps to narrow down the search result further. "Coordinate matching" [14] is an efficient similarity measure to refine the relevance of result, used in the plaintext information retrieval community widely. However, the application of this measure in the encrypted cloud data search system remains a very challenging task because of security and privacy reasons.

In this paper we define and solve the problem of secure fuzzy multi-keyword ranked search over encrypted cloud data (FMRSE) while preserving strict system wise privacy in the cloud computing paradigm. Coordinate matching specifically inner product similarity is used to quantify similarity of a document to search query. Each document is associated with a binary vector as a sub index where each bit represents whether the corresponding keyword is contained in the document. The search query is also associated with a binary vector where each bit represents whether the corresponding keyword appears in this search request. Thus similarity can be measured exactly by the inner product of the query vector with the data vector. Directly outsourcing the data vector or the query vector will violate the index privacy or the search privacy. Thus both the data vector and query vector have to be encrypted before outsourcing .Also, to meet the challenge of supporting minor types, format inconsistencies and multi keyword semantic without privacy breaches we propose FMRSE using secure inner product computation, adapted from a secure k-nearest neighbour (kNN) technique [15].

II. RELATED WORK

A. *Searchable encryption*

Existing searchable encryption schemes allow a user to securely search over encrypted data through keywords without decrypting it. The first construction of searchable encryption was proposed by Song et al. [16], where every word in the document is encrypted independently by a special two-layered encryption construction. Goh [8] proposed to construct the indexes for the data files using Bloom filters. For more efficient search, Chang et al. [17] and Curtmola et al. [18] both proposed a single encrypted hash table index to be built for the entire file collection. Traditional single keyword searchable encryption schemes [6],[8],[16],[17],[18],[19] usually build an encrypted searchable index such that its content is hidden to the server unless it is given appropriate trapdoors generated by secret key. The works [10], [20] utilizes keyword frequency to rank results instead of returning undifferentiated results. Boneh et al. [6] present the first searchable encryption construction, where anyone with public key can write to the data stored on server but only authorized users with private key can search Public key solutions are usually computationally expensive. And the keyword privacy could not be protected in the public key setting since server could encrypt any keyword with public key and then use the received trapdoor to evaluate this cipher text. All these existing schemes support only exact keyword search therefore not suitable for cloud computing.

B. *Boolean keyword searchable encryption*

Designs that have been proposed to support Boolean keyword search [21]-[26] are still not adequate to provide users with acceptable result ranking functionality. Conjunctive keyword search returns "all-or-nothing," i.e.it only returns those documents where all the keywords specified by the search query appear. Disjunctive keyword search returns only that document that contains a subset of the specific keywords. Predicate encryption schemes [24], [25], [26] are recently proposed to support both conjunctive and disjunctive search. None of existing Boolean keyword searchable encryption schemes support multiple keywords ranked search over encrypted cloud data while preserving

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

privacy. Most of these schemes are built upon the expensive evaluation of pairing operations on elliptic curves. Such inefficiency disadvantage also limits their practical performance when deployed in the cloud.

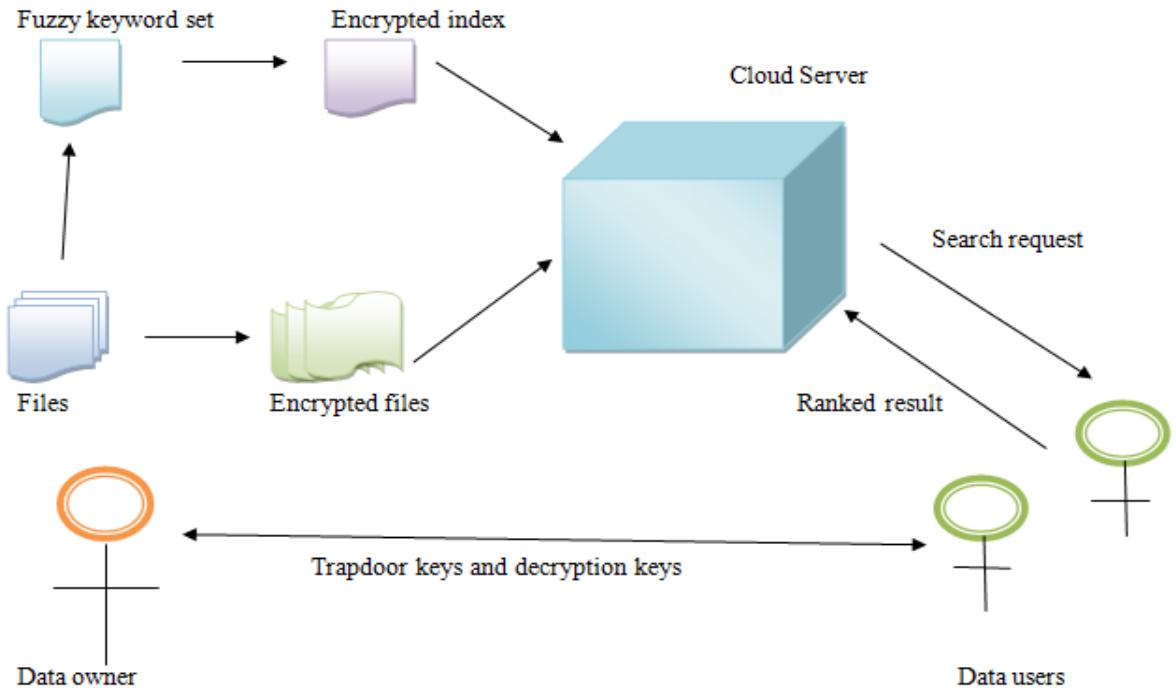

Fig.1. System Architecture

C. *Fuzzy keyword search*

Kui [27] analyzed that users have various typing behaviors for keywords and are known as typos, representation inconsistencies and typing habits. The importance of fuzzy search has gained attention in the context of plain text searching in information retrieval community [28]–[30].They addressed this problem in the traditional information access paradigm for finding relevant information based on approximate string matching. This construction suffers from the dictionary and statistics attacks and fails to achieve the search privacy. Wei [7] created k-gram based fuzzy keyword set for W of encrypted files C and Jaccard coefficient to calculate keyword similarity. The size of the k-gram based fuzzy keyword set depends on the Jaccard coefficient value. Jianfeng [31] discuss that the keyword contains file sensitive information, so keyword privacy must be protected. He proposed this search using symbol tree. The verification is done by users by checking the proof set and ID set created from index. Verifiable fuzzy keyword search requires more space for storing the symbol tree fuzzy searchable index Peng [32] found that third party could access files by knowing keyword search trapdoor. The process of creating fuzzy keyword index and exact keyword index is too difficult if the database is very large.

III. PROBLEM FORMULATION

A. *System model*

As shown in Fig.1, a cloud computing system hosting data service is considered here in which three different entities are present, data owner, data user and cloud server. The data can contain many sensitive information. As the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

cloud servers cannot be completely trusted to protect data, the files must be encrypted before outsourcing. The cloud server will provide keyword retrieval service to authorized users. There is a predefined set of keywords W and a new fuzzy keyword set F is built based on W. The data owner builds a searchable index I from F and then outsources the encrypted index and the encrypted files onto the cloud server. The computing power on user side is limited i.e. the operation on user side should be simplified .The data user at first generates a query and the keywords are kept concealed for privacy reasons. To search the document collection for the given keywords, an authorized user acquires a corresponding trapdoor T through search control mechanisms. Corresponding set of encrypted documents is returned upon receiving T from data user after searching index I. By ranking the search result according to coordinate matching, the document retrieval accuracy can be improved. An optional number k can be send along with the trapdoor T by the user to reduce the communication cost, as it sends back only top-k documents that are most relevant to the query. The data user can use the files after decrypting it.

B. *Threat model*

A semi trusted cloud server is considered in our model. The cloud server acts in an "honest" manner as it follows the designated protocol specification but it is "curious" to analyze the data in its storage and search requests from users so as to learn additional information. Care must be taken to conduct search in a secure manner, so that the retrieval of data files reveals only little information as possible to the cloud server. In this paper ,the cloud server is supposed to know only the encrypted data set and the searchable index both of which are outsourced from the data owner and it is required that nothing should be leaked from the data set and index beyond the outcome and pattern of search queries.

C. *Design goals*

In this paper, we address the problem of efficient fuzzy multi keyword ranked search over encrypted cloud data. For effective utilization of outsourced cloud data, we have the following goals.1) to design a mechanism for constructing fuzzy keyword sets which are storage efficient.2) to design a multi keyword search scheme based on the constructed fuzzy keyword sets.3) to provide result similarity ranking for effective data retrieval.4) to meet privacy requirements the cloud server is prevented from learning additional information from data set and index.5)to provide low computation and communication overhead.

D. *Preliminaries*

- *Fuzzy keyword set construction based on edit distance*

In this paper we use the concept of edit distance [33] to measure keyword similarity. This technique will reduce the need to mention all fuzzy keywords one by one and moreover the size of fuzzy keyword set is further reduced. The number of operations required to transform one word w1 to another word w2 is the edit distance ed (w1, w2) between two words. Mainly three basic operations are performed for transformation of one word to another. They are

1) Substitution - In a word one character is changed to another.
2) Deletion – From a word any one character is deleted.
3) Insertion – In a word any one character is inserted.

A keyword w_i is given, then F denote the set of fuzzy keywords satisfying ed $(w_i , w_i^{'}) \leq d$ for a certain integer d.

- *Coordinate matching*

"Coordinate matching" [14] is a similarity measure widely used in plain text Information Retrieval community. To apply it in the encrypted cloud data is a very challenging task, due to security and privacy reasons. In coordinate matching the relevance of a document to the query is determined by the number of query keywords appearing in that document. In cloud computing users does not know the exact subset of the data set to be retrieved, due to huge amount of outsourced data. Thus, it is more convenient for the users to specify a list of keywords that indicates their interest .The users can retrieve the most relevant top-k documents with a rank order.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

- *Secure kNN computation*

In the secure kNN scheme [15], to select k nearest database records, the Euclidean distance between a data record and a query vector is used. If d is the number of fields of each record, then the secret key is composed of one (d+1) bit vector as S and two (d+1)×(d+1)invertible matrices. At first, every data vector and query vector are extended to (d+1) dimension vectors. Besides, the query vector is scaled by a random number. Then the (d+1) dimension query and data vectors are split into two random vectors respectively based on the vector S. When i^{th} bit of S is 0,then the bits of split vectors are set as the same bit in the (d+1) dimension data vector whereas the bits in split query vector is set to two random numbers, so that their sum is equal to that of i^{th} bit in the (d+1) dimension query vector. If the i^{th} bit of S is 1, the splitting process is same except that the process is switched between the data and query vector. The split data vector pair is encrypted by multiplying them with the transpose of the matrices of the secret key whereas the split query vector pair is encrypted by multiplying them with the inverse of matrices of the secret key .To select k nearest neighbours the product of data vector pair and query vector pair is taken.

IV. FMRSE FRAMEWORK

In our more design, first we construct storage efficient fuzzy keyword sets. Then an efficient privacy preserving search scheme is proposed. Basically, secure kNN computation is used with more advanced design. The dimension extending operation is preserved and in addition a new random number t is assigned to the extended dimension in each query vector. The newly added randomness increases the difficulty for the cloud server to learn the relationship among the received trapdoors. Also, a dummy keyword is inserted to each data vector and a random value is assigned to it. Each individual vector is extended to (n+2) dimension instead of (n+1).The whole scheme to achieve fuzzy ranked search with multiple keywords over encrypted data is as follows:

Build fuzzy keyword list : First create a wordlist W. Taking each word w_i in the word list, create a set of words w_i' satisfying $ed(w_i, w_i') \leq 1$ and place that in fuzzy word set F. For example, the following is the listing variants after a substitution operation on the first character of keyword CAST: {AAST, BAST, DAST, · · · YAST, ZAST}. Combine the words in the wordlist W and fuzzy set F and create a new fuzzy word list N. Based on the new list N the following operations are performed.

Setup :The data owner will generate an (n+2) bit vector randomly as V and two (n+2) × (n+2) invertible matrices {M1,M2}.Now this will be the secret key S.

Build Index: The corresponding document and secret key is given as input to this module. The owner generates a binary data vector B_i for every document D_i, where each binary bit in B_i represents whether the corresponding keyword in document D_i is in the new fuzzy word list N. The plain text sub index B_v will be generated by implementing dimension extending and splitting procedures in B_i .It's similar to secure kNN computation except that $(n+1)^{th}$ entry in B_v is set to a random number and $(n+2)^{th}$ entry is set to 1.By splitting B_v as in secure kNN we get B_v' and B_v'' . The sub index X_i is generated by multiplying the split vectors with the transpose of matrices M1 and M2. Thus $X_i = \{M1^T B_v' .M2^T B_v''\}$.

Trapdoor: Here n keywords is given as input in W_q .A binary vector Q is generated where each of binary bit represent whether the keyword is present in the new fuzzy keyword set N.Q is extended to (n+2) dimension vector Q_v .In Q_v the $(n+1)^{th}$ bit is set to 1 then its scaled by a random number r where $r \neq 0$, after which it's extended to (n+2) dimension with another random number. Again we apply the splitting operation as in secure kNN computation to get Q_v' and Q_v''. Then trapdoor T is generated by multiplying the split vectors with the inverse of matrices M1 and M2. Thus trapdoor T = $\{Q_v' M1^{-1}, Q_v'' M2^{-1}\}$.

Rank: With the trapdoor T, the cloud server will compute the similarity scores of each document D_i as X_i .T. Similarity score will have a larger value if the query keywords are in a document. The documents will be ranked based on the similarity score. Top-k documents will be returned to the user where k is specified by the client.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

V. SIMULATION RESULTS

An experimental evaluation of the proposed technique is performed on real world documents. The experimental system is implemented by Java language with Intel Core i5 Processor, 2.30GHz, 4 GB RAM.

First we randomly select different number of documents and form different data sets. Then a query is given to each data set and the execution time for score calculation is noted for each data set. Here, the number of query keywords is kept constant as 2. The corresponding graph is illustrated in Fig.2. Next we can determine the effect of query keywords on the execution time by keeping the number of documents in the data set constant. Here the number of documents in the data set is kept as 5.Then the execution time for score calculation is determined for different number of query keywords. Fig .3 illustrates the corresponding graph. From both the graph it is clear that the execution time increases as number of documents in the data set and/or the number of keywords in the query increases.

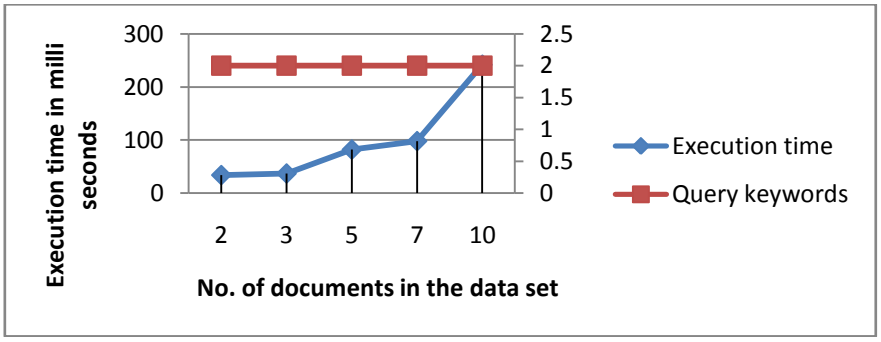

Fig.2. Execution time graph for different number of documents in data set

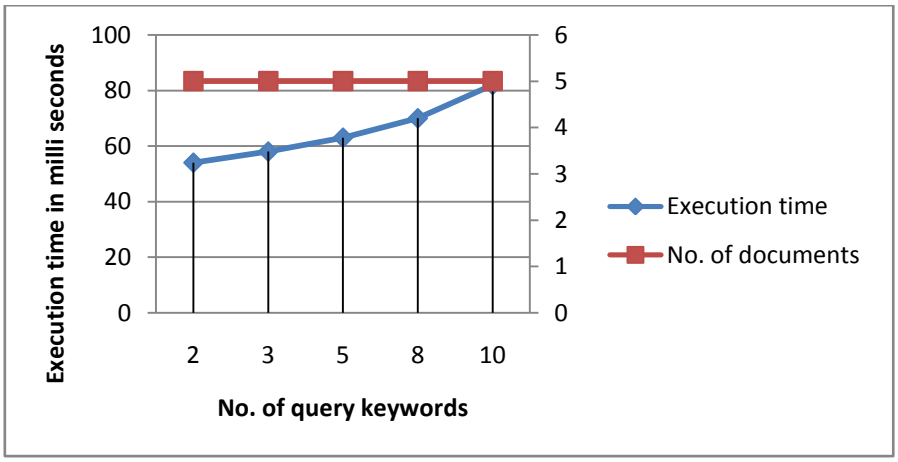

Fig.3. Execution time graph for different number of query keywords

VI. CONCLUSION

In this paper, we propose the method of secure fuzzy multi keyword ranked search over encrypted cloud data for efficient utilization of remotely stored encrypted cloud data. An advanced technique is designed to construct fuzzy keyword set based on similarity metric of edit distance. Based on the fuzzy keyword set an efficient privacy preserving keyword search scheme is proposed. The efficient similarity measure of "inner product computation" is used to capture the relevance of outsourced documents to the query keywords. Top-k documents are returned to the user based on similarity score, where k is specified by the user. The proposed solution is secure and impose low overhead on computation and communication.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

REFERENCES

[1] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," Proc. IEEE INFOCOM, pp. 829-837, Apr, 2014.
[2] Jiadi Yu, Peng Lu, Yanmin Zhu, Guangtao Xue, Member, IEEE Computer Society, and Minglu Li, "Toward Secure Multikeyword Top k Retrieval over Encrypted Cloud Data", IEEE Transactions, July 2013.
[3] 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.
[4] J. Li et al., "Fuzzy Keyword Search Over Encrypted Data in Cloud Computing," Proc. IEEE INFOCOM, 2010, pp. 441–45.
[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] D. Boneh, G.D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public Key Encryption with Keyword Search," Proc. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), 2004.
[7] 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.
[8] E.-J. Goh, "Secure Indexes," Cryptology ePrint Archive, http://eprint.iacr.org/2003/216. 2003.
[9] M. Li, S. Yu, N. Cao, and W. Lou, "Authorized Private Keyword Search over Encrypted Data in Cloud Computing," Proc. 31st Int'l Conf. Distributed Computing Systems (ICDCS '10), pp. 383-392, June 2011.
[10] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure Ranked Keyword Search over Encrypted Cloud Data," Proc. IEEE 30th Int'l Conf. Distributed Computing Systems (ICDCS'10), 2010.
[11] S. Kamara and K. Lauter, "Cryptographic Cloud Storage," Proc. 14th Int'l Conf. Financial Cryptograpy and Data Security, Jan. 2010.
[12] S. Zerr, D. Olmedilla, W. Nejdl, and W. Siberski, "Zerber +r: Top-k Retrieval from a Confidential Index," Proc. 12th Int'l Conf. Extending Database Technology: Advances in Database Technology (EDBT), 2009.
[13] A.Swaminathan , Y. Mao, G.-M. Su, H. Gou, A.L. Varna, S. He, M. Wu, and D.W. Oard, "Confidentiality-Preserving Rank-Ordered Search," Proc. Workshop Storage Security and Survivability, 2007.
[14] I.H. Witten, A. Moffat, and T.C. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann Publishing May 1999.
[15] W.K. Wong, D.W. Cheung, B. Kao, and N. Mamoulis, "Secure kNN Computation on Encrypted Databases," Proc. 35th ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 139-152, 2009.
[16] D. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in Proc. of IEEE Symposium on Security and Privacy'00, 2000.
[17] Y.-C. Chang and M.Mitzenmacher ,"Privacy preserving keyword searches on remote encrypted data," in Proc. of ACNS'05, 2005.
[18] R. Curtmola, J. A. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: improved definitions and efficient constructions," in Proc. of ACM CCS'06, 2006.
[19] M. Bellare, A. Boldyreva, and A. ONeill, "Deterministic and Efficiently Searchable Encryption," Proc. 27th Ann. Int'l Cryptology Conf. Advances in Cryptology (CRYPTO '07), 2007.
[20] C. Wang, N. Cao, K. Ren, and W. Lou, "Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 8, pp. 1467- 1479, Aug. 2012.
[21] D. Boneh and B. Waters, "Conjunctive, Subset, and Range Queries on Encrypted Data," Proc. Fourth Conf. Theory Cryptography (TCC), pp. 535-554, 2007.
[22] R. Brinkman, "Searching in Encrypted Data," PhD thesis, Univ. of Twente, 2007.
[23] Y. Hwang and P. Lee, "Public Key Encryption with Conjunctive Keyword Search and Its Extension to a Multi-User System, "Pairing, vol. 4575, pp. 2-22, 2007.
[24] J. Katz, A. Sahai, and B. Waters, "Predicate Encryption Supporting Disjunctions, Polynomial Equations, and Inner Products," Proc. 27th Ann. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), 2008.
[25] A. Lewko, T. Okamoto, A. Sahai, K. Takashima, and B. Waters, "Fully Secure Functional Encryption: Attribute-Based Encryption and (Hierarchical) Inner Product Encryption," Proc. 29th Ann.Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT '10), 2010.
[26] E. Shen, E. Shi, and B. Waters, "Predicate Privacy in Encryption Systems" Proc. Sixth Theory of Cryptography Conf. Theory of Cryptography (TCC), 2009.
[27] Kui Ren et al., "Towards Secure And Effective Data utilization in Public Cloud", IEEE Transactions on Network, volume 26, Issue 6, November / December 2012.
[28] C. Li, J. Lu, and Y. Lu, "Efficient merging and filtering algorithms for approximate string searches," in Proc. of ICDE'08, 2008.
[29] A. Behm , S.Ji, C. Li and J. Lu, "Space-constrained gram-based indexing for efficient approximate string search," in Proc. of ICDE'09.
[30] S. Ji, G. Li, C. Li, and J. Feng, "Efficient interactive fuzzy keyword search," in Proc. of WWW'09, 2009.
[31] Jianfeng Wang et al., "Efficient Verifiable Fuzzy Keyword Search over Encrypted Data in Cloud Computing" , Journal of Computer Science and Information system, volume 10, Issue 2, April 2013.
[32] 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.
[33] V. Levenshtein, "Binary codes capable of correcting spurious insertions and deletions of ones," Problems of Information Transmission, vol. 1,no. 1, pp. 8–17, 1965.