International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Achieving Efficiency of Encrypted Cloud Data with Synonym Based Search and Multi-Keyword Ranked Search

Dipika Chavan[1], Dinesh Yadav[2]

[1]Computer Department, RSSOER, Narhe Technical Campus, Pune, India - 411041

[2]Director, RSSOER, Narhe Technical Campus, Pune, India – 411041

Abstract: ***Nowadays, cloud computing becomes efficient and flexible with reduced cost and utility of on-demand high quality applications and services, so internet usage strongly relies on cloud for privacy preserving and fast data retrieval. For consumers, they want to find the most relevant products or data, which is highly desirable in the "pay-as-you use" cloud computing paradigm. Sensitive data (such as photos, mails, health records, financial records, etc) is encrypted before outsourced to cloud. Although Searchable encryption scheme has been developed to conduct retrieval over encrypted data, these schemes only support exact or fuzzy keyword search, mainly evaluate the similarity of keywords from the structure but the semantic relatedness is not considered. This work focuses on realizing secure semantic search through query keyword semantic extension based on the co-occurrence probability of terms, the semantic relationship library is constructed to record the semantic similarity between keywords. To achieve efficiency of the search method we enhance the TFIDF algorithm by extending the keyword set with semantic words or natural language words for the keywords. This will ultimately support data retrieval on querying semantic query. Even when user doesn't know exact or synonym of keywords of encrypted data, he can try searching it by its meaning in natural language. WordNet method makes the search scheme even more reliable and better.***

Keywords: Cloud computing, multi-keyword search, semantic based search, TFIDF, anaphora resolution, WordNet Ontology.

1. Introduction

Today, consumer centric cloud computing is a new model of enterprise-level in IT infrastructure providing the on-demand high quality applications and services from a shared pool of computing resources. The Cloud Service Provider (CSP) has full control of the outsourced data; it may learn some additional information from that data therefore some problems arise in the circumstance. So, sensitive data is encrypted before outsourcing to the cloud. However the encrypted data make the traditional plaintext search methods useless. The simple and awkward method is downloading all data and decrypt it locally is obviously impractical, because the consumers want to search only the interested data rather all the data. Therefore it is essential to explore an efficient and effective search service over encrypted outsourced data.

The existing search approaches like ranked search, multi-keyword search that enables the cloud customers to find the most relevant data quickly. It also reduces the network traffic by sending the most relevant data to user request. But In real search scenario it might be possible that user searches with the synonyms of the predefined keywords not the exact or fuzzy matching keywords, due to lack of the user's exact knowledge about the data. These approaches supports only exact or fuzzy keyword search. That is there is no tolerance of synonym substitution and/or syntactic variation which are the typical user searching behaviors happens very frequently. Therefore synonym based multi-keyword ranked search over encrypted cloud data remains a challenging problem.

To overcome this problem of effective search system this paper proposes an efficient and flexible searchable scheme that supports both multi-keyword ranked search and semantic based search. The Vector Space Model is used to address multi-keyword search and result ranking. By using VSM document index is build i.e. each document is expressed as vector where each dimension value is the Term Frequency (TF) weight of each corresponding keyword. Another vector is generated in query phase. It has same dimension as that of document index and its each dimension value is the Inverse Document Frequency (IDF) weight. Then cosine measure is used to calculate the similarity between the document and the search query.

To enhance the efficiency of the search method we use the extended keyword set with semantic words or natural language words for the keywords. This will ultimately support data retrieval on querying semantic query. Even when user doesn't know exact or synonym of keywords of encrypted data, he can try searching it by its meaning in natural language. WordNet ontology is used to solve the problem of anaphora resolution. This makes the Semantic search more efficient and User doesn't need to worry about the keyword generated for each particular word on the cloud by adapting this method data will be retrieved from the cloud in well secure manner and also cost can be minimized by employing these scheme into the structure and also we are incorporating WordNet method which makes the search scheme even more reliable and Better.

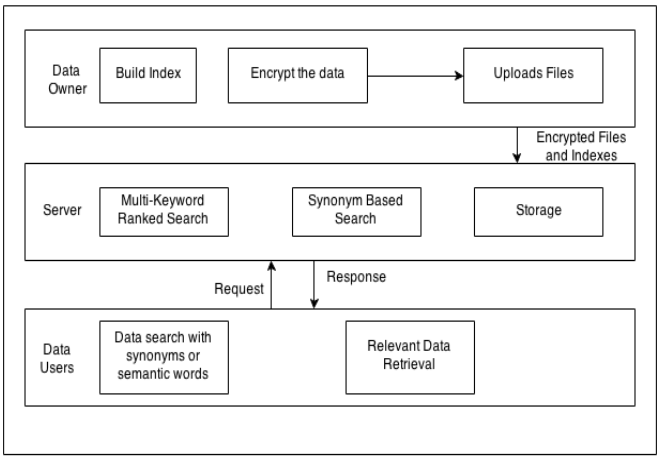

Figure 1: System Architecture

2. Literature Survey

J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, uses the Fuzzy keyword search method that enhances system usability by returning the matching files containing exact match of the predefined keywords or the closest possible matching files based on keyword similarity semantics, when *exact* match fails. They exploit edit distance to quantify keywords similarity and develop an advanced technique on constructing fuzzy keyword sets, which greatly reduces the storage and representation overheads [2].

C. Wang, N. Cao, J. Li, K. Ren, and W. Lou proposes the Ranked search that enhances system usability by returning the matching files in a ranked order regarding to certain relevance criteria (e.g., keyword frequency). It gives a straightforward yet ideal construction of ranked keyword search under the state-of-the-art searchable symmetric encryption (SSE) security definition, and demonstrates its inefficiency. To achieve more practical performance, they propose a definition for ranked searchable symmetric encryption, and give an efficient design by properly utilizing the existing cryptographic primitive, order-preserving symmetric encryption (OPSE) [3].

N. Cao, C. Wang, M. Li, K. Ren, and W. Lou designed a system that solves the challenging problem of privacy-preserving multi-keyword ranked search over encrypted cloud data (MRSE), and establish a set of strict privacy requirements for such a secure cloud data utilization system to become a reality. Among various multi-keyword semantics, they choose the efficient principle of "coordinate matching", i.e., as many matches as possible, to capture the similarity between search query and data documents, and further use "inner product similarity" to quantitatively formalize such principle for similarity measurement [4].

W. Sun, B. Wang, N. Cao, M. Li, W. Lou, and Y. T. Hou present a privacy-preserving multi-keyword text search (MTS) scheme with similarity-based ranking to address this problem. To further enhance the search privacy, they propose two secure index schemes to meet the stringent privacy requirements under strong threat models. In particular, to support multi-keyword queries and search result ranking functionalities, they proposes to build the search index based on the vector space model, i.e., cosine measure, and incorporate the $TF \times IDF$ weight to achieve high search result accuracy[6].

Zhangjie Fu, Xingming Sun, Nigel Linge and Lu Zhou proposes an effective approach to solve the problem of multi-keyword ranked search over encrypted cloud data supporting synonym queries. To address multi-keyword search and result ranking, Vector Space Model (VSM) is used to build document index, that is to say, each document is expressed as a vector where each dimension value is the Term Frequency (TF) weight of its corresponding keyword. A new vector is also generated in the query phase. The vector has the same dimension with document index and its each dimension value is the Inverse Document Frequency (IDF) weight. Then cosine measure can be used to compute similarity of one document to the search query. To improve search efficiency, a tree-based index structure which is a balance binary tree is used [1].

3. Methodology

3.1 Multi-Keyword Ranked Search:

The existing systems like exact or fuzzy keyword search, supports only single keyword search. These schemes doesn't retrieve the relevant data to users query therefore multi-keyword ranked search over encrypted cloud data remains a very challenging problem. To meet this challenge of effective search system, an effective and flexible searchable scheme is proposed that supports multi-keyword ranked search. To address multi-keyword search and result ranking, Vector Space Model (VSM) is used to build document index, that is to say, each document is expressed as a vector where each dimension value is the Term Frequency (TF) weight of its corresponding keyword. A new vector is also generated in the query phase. The vector has the same dimension with document index and its each dimension value is the Inverse Document Frequency (IDF) weight. Then cosine measure can be used to compute similarity of one document to the search query [1].

To improve search efficiency, a tree-based index structure used which is a balance binary tree is. The searchable index tree is constructed with the document index vectors. So the related documents can be found by traversing the tree.

3.2 Semantic Based Search:

While user searching the data on cloud server it might be possible that the user is unaware of the exact words to search, i.e. there is no tolerance of synonym substitution or syntactic variation which are the typical user searching behaviors and happen very frequently. To solve this problem semantic based search method is used. To improve the search for information it is necessary that search engines can understand what the user wants so they are able to answer objectively. To achieve that, one of the necessary things is that the resources have information that can be helpful to searches.

The Semantic Web proposed to clarify the meaning of resources by annotating them with metadata data over data. By associating metadata to resources, semantic searches can

be significantly improved when compared to traditional searches. It allows users the use of natural language to express what he wants to find. Here the enhanced E-TFIDF algorithm is proposed for improving documental searches optimized for specific scenarios where user want to find a document but don´t remember the exact words used, if plural or singular words were used or if a synonym was used. The defined algorithm takes into consideration: 1) the number of direct words of the search expression that are in the document; 2) the number of word variation (plural/singular or different verbs conjugation) of the search expression that are in the document; 3) the number of synonyms of the words in the search expression that are in the document; weights to each one of this components as the fuzziness part of the algorithm [7].

4. Conclusion

The proposed Semantic Search with WordNet methodology makes the Search process more efficient. The proposed scheme could return not only the exactly matched files, but also the files including the terms semantically related to the query keyword. The concept of co-occurrence probability of terms is used to get the semantic relationship of keywords in the dataset. It offers appropriate semantic distance between terms to accomplish the query keyword extension. To guarantee the security and efficiency, the data is encrypted before outsourced to cloud, and provides security to datasets, indexes and keywords also. Then the data owner groups the indexes and forms the ontology based on the documents which is having syntactically and semantically similar words.

The overall performance evaluation of this scheme includes the cost of metadata construction, the time necessary to build index and ontology construction as well as the efficiency of search and WordNet methodology which makes the search scheme still more efficient to the user and by employing this technique keyword that we used for searching will also protected and better search mechanism can be achieved.

References

[1] Zhangjie Fu, Xingming Sun, Nigel Linge and Lu Zhou, “Achieving Effective Cloud Search Services: Multi-keyword Ranked Search over Encrypted Cloud Data Supporting Synonym Query”, IEEE Transactions on Consumer Electronics, Vol. 60, No. 1, February 2014.
[2] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, “Fuzzy keyword search over encrypted data in cloud computing,” *Proceedings of IEEE INFOCOM’10 Mini-Conference*, San Diego, CA, USA, pp. 1-5, Mar. 2010.
[3] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, “Secure ranked keyword search over encrypted cloud data,” *Proceedings of IEEE 30th International Conference on Distributed Computing Systems (ICDCS)*, pp. 253-262, 2010.
[4] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, “Privacy-preserving multi-keyword ranked search over encrypted cloud data,” *Proceedings of IEEE INFOCOM 2011*, pp. 829-837, 2011.
[5] Q. Chai, and G. Gong,“Verifiable symmetric searchable encryption for semi-honest-but-curious cloud servers,” *Proceedings of IEEE International Conference on Communications (ICC’12)*, pp. 917-922, 2012.
[6] W. Sun, B. Wang, N. Cao, M. Li, W. Lou, and Y. T. Hou, “Privacy preserving multi-keyword text search in the cloud supporting similarity based ranking,” *ASIACCS 2013*, Hangzhou, China, May 2013, pp. 71-82, 2013.
[7] Sara Paiva, ”A Fuzzy Algorithm for Optimizing Semantic Documental Searches”, International Conference on Project Management / HCIST 2013.
[8] Automatic Pronominal Anaphora Resolution in English Texts, Tyne Liang and Dian-Song Wu.

Author Profile

Dipika Chavan received Diploma degree in Computer Engineering from I.O.P.E. Lonere, Maharashtra and B.E. in Computer Engineering from RMCET, Ratnagiri, Maharashtra in 2007 and 2010 respectively. She is now pursuing M.E. in Computer Engineering at RSSOER,Pune, Maharashtra. Her area of interest is Cloud Computing, Data Mining.

Dinesh Yadav is Director of Rajarshi Shahu College of Engineering and Research, JSPM NTC Pune, India. He obtained Bachelor of Engineering, Masters of Engg. and PhD in Electronics and Telecommunication Engineering. His area of interest is Image Processing, Digital Signal Processing and Networking.