e-ISSN: 2349-9745
p-ISSN: 2393-8161

Scientific Journal Impact Factor (SJIF): 1.711

International Journal of Modern Trends in Engineering and Research

www.ijmter.com

Cloud Search Service: Using Multi-Keyword Search Query

D.Packialakshmi*[1], R.K.Sindhuja*[2], K.Ramadevi*[3]

[1] Department Of Information Technology, S.K.P Engineering College
[2] Department Of Information Technology, S.K.P Engineering College
[3] Department Of Information Technology, S.K.P Engineering College

Abstract— Our project is summarized in two aspects: multi-keyword ranked search to achieve more accurate search results and synonym-based to support synonym queries. Extensive experiments on real-world dataset were performed to validate the approach showing that the proposed solution is very effective and efficient for multi-keyword ranked searching in a cloud environment. 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. Word Net method makes the search scheme even more reliable and better.

Keywords - multi-keyword, synonym, encrypted data.

I. INTRODUCTION

Improve search result accuracy as single keyword search often return coarse search results. In the real search scenario, it is quite common that cloud customers searching input Existing search approaches cannot accommodate such requirements like ranked search, multi-keywords search, semantics-based search etc. The ranked search enables cloud customers to find the most relevant information quickly. Ranked search can also reduce network traffic as the cloud server sends back only the most relevant data. Multi-keyword search is also very important to might be the synonyms of the predefined keywords, not the exact or fuzzy matching keywords due to the possible synonym substitution (reproduction of information content), such as commodity and goods, and/or her/his lack of exact knowledge about the data. The existing searchable encryption schemes support only exact or fuzzy keyword search. There is no tolerance of synonym substitution and syntactic variation which is on the other hand are typical user searching behaviours and happen very frequently. Therefore synonym –based multi-keyword ranked search over encrypted cloud data remains a very challenging problem. To meet the challenge of effective search system, this paper proposes a practically efficient and flexible searchable scheme which supports both multi-keyword ranked search and synonym based 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. To improve search efficiency, a tree-based index structure which is a balance binary tree is used. The searchable index tree is constructed with the document index vectors. So the related documents can be found by traversing the tree.

The contributions of this paper are summarized as follows:

(1) For the first time, a semantics-based multi-keyword ranked search technology over encrypted cloud data which supports synonym queries is proposed. The search results can be achieved when authorized cloud customers input the synonyms of the predefined keywords, not the exact or fuzzy matching keywords, due to the possible synonym substitution and/or her lack of exact knowledge about the data.

(2) By incorporating the state-of-art text feature extraction technique TFIDF (term frequency-inverse document frequency), an enhanced semantic feature extraction method E-TFIDF is proposed. The E-TFIDF algorithm, which can extract the most representative keywords from outsourced text documents, improves the accuracy of search results.

(3) Extensive experiments on the real-world dataset further show the effectiveness and efficiency of proposed solution. In the remainder of this paper, the following information is presented: In Section II, related research is discussed. Then, problem formulation is described in Section III. In Section IV, the proposed method for building keyword set extended by synonym in cloud is presented in detail. Section V presents the proposed search schemes. Performance analysis is presented in Section VI. Finally, in Section VII, the paper concludes with some suggestions for future work.

II. EXISTING SYSTEM

Existing system is based on synonym search. Synonym search-it shows related meaning only. Existing search approaches cannot accommodate such requirements like ranked search, multi-keyword search, semantics-based search etc. In the real search scenario the searching input might be the synonym of the predefined keyword. There is no tolerance of synonym substitution. It is difficult to search. Synonyms are words with the same or similar meanings. In order to improve the accuracy of search results, the keywords extracted from outsourced text documents need to be extended by common synonyms, as cloud customers' searching input might be the synonyms of the predefined keywords, not the exact or fuzzy matching keywords due to the possible synonym substitution and/or her lack of exact knowledge about the data. A common synonym thesaurus is built on the foundation of the New American Roget's College Thesaurus (NARCT). Then the keyword set is extended by using the constructed synonym thesaurus.

For Example: If we enter any input to the cloud server it gives related synonym of that word. Let us assume Education as a keyword it gives result as the synonym of that word.

III. PROPOSED SYSTEM

Our project proposes a practically efficient and flexible searchable scheme which supports both multi-keyword ranked search and synonym based search. To address multi -keyword search and result ranking, Vector Space Model (VSM) is used to build document index. The vector has the same dimension with document index and its each dimension value is the Inverse Document Frequency (IDF) weight. To improve search efficiency, a tree based index structure which is a balance binary tree is used. Our paper propose about multi- keyword and fuzzy keyword. In information retrieval, a ranking function is usually used to evaluate relevant scores of matching files to a request. Among lots of ranking functions, the "*TF*×*IDF*" rule is most widely used, where *TF* (term frequency) denotes the occurrence of the term appearing in the document, and *IDF* (inverse document frequency) is often obtained by dividing then total number of documents by the number of files containing the term. That means, *TF* represents the importance of the term in the document and *IDF* indicates the importance or degree of distinction in the whole document collection. Each document is corresponding to an index vector *Dd* that stores normalized *TF* weight, and the query vector *Q* stores normalized *IDF* weight. Each dimension of *dD* or *Q* is related to a keyword in *W*, and the order is same with that in *W*, that is, *D* [*i*] *d* is corresponding to keyword *i w* in *W*. The similarity evaluation function [15] is employed for cosine measure.

3.1. Basic scheme

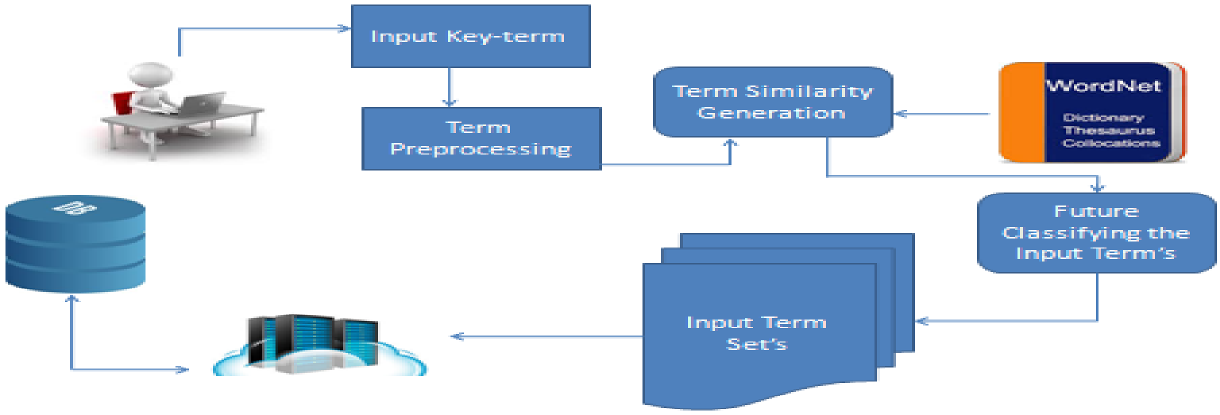
In this phase, the system is initialized. The data owner generates the secret key *SK* and picks a random key *sk* .The *SK* includes: (1) A n-bit randomly generated vector *S* ;

(2) Two n*n invertible matrices { , } 1 2 *M M* . Hence, *SK* is in the form of a 3-tuple as { , , } 1 2 *S M M* .

3.2. Enhance scheme

In the basic scheme, the keyword privacy leakage is possible in the known background model because the cosine value calculated from encrypted vector *dD* ~ and *Q*~ is equal to the one form vector *dD* and *Q*. For the purpose of eliminating such equality property, some dummy keywords can be used.

To be specific, all vectors (including document index vectors and query user vectors) are extended to (*n*+*U*)-dimensions, where *U* is the number of dummy keywords and each extended dimension is corresponding to a dummy keyword.



3.3. User Interface

1. Search space

After user login process, cloud user can enter the search space page. This is the environment for user to search the content from the cloud server. This Search Space is the interface for user and cloud servers.

2. Input from User

Get the input text from the user for the search process.

3.4. Data Preprocessing

1. Stop word Removal

Stop words are words which are filtered out prior to, or after, processing of natural language data (text). It is controlled by human input and not automated. These are some of the most common, short function words, such as *the*, *is*, *at*, *which* and *on*.

2. Poster Stemming

Stemmers employ a lookup table which contains relations between root forms and inflected forms. To stem a word, the table is queried to find a matching inflection. If a matching inflection is found, the associated root form is returned. Eg: A stemming algorithm reduces the words "fishing", "fished", "fish", and "fisher" to the root word, "fish".

3.5. Ontology Clustering:

Words ending in nym's are often used to describe different classes of words, and the relationships between words.

Hypernym: A word that has a more general meaning than another.
Hyponym: A word that has a more specific meaning than another.
Synonym: One of two (or more) words that have the same (or very similar).

The Artificial-Intelligence literature contains many definitions of ontology (Word net).It includes machine-interpretable definitions of basic concepts in the domain and relations among them. The featured results produced by the sentence-based, document-based, corpus-based, and the combined approach concept analysis have higher quality than those produced by a single-term analysis similarity.

IV. CONCLUSION

This paper, for the first time, proposes an effective approach to solve the problem of synonym-based multi-keyword ranked search over encrypted cloud data. The main contributions are summarized in two aspects: synonym-based search and similarity ranked search. The search results can be achieved when authorized cloud customers input the synonyms of the predefined keywords, not the exact or fuzzy matching keywords, due to the possible synonym substitution and/or her lack of exact knowledge about the data. The vector space model is adopted combined with cosine measure, which is popular in information retrieval field, to evaluate the similarity between search request and document. Finally, the performance of the proposed schemes is analyzed in detail, including search efficiency and search accuracy, by the experiment on real-world dataset. The results show that the proposed solution is very efficient and effective in supporting synonym-based searching. The next work is to research semantics-based search approaches over encrypted cloud data that support syntactic transformation, anaphora resolution and other natural language processing technology. The aim is that cloud consumers can search the most relevant products or data by using the designed system.

REFERENCES

[1] P.A.Cabarcos, F.A. Mendoza, R.S. Guerrero, A.M. Lopez, and D. Diaz-Sanchez,"SuSSo: seamless and ubiquitous single sign-on for cloud service continuity across devices," IEEE Trans. Consumer Electron.,vol. 58, no. 4, pp. 1425-1433, 2012.

[2] D. Diaz-Sanchez, F. Almenarez, A. Marin, D. Proserpio, and P.A. Cabarcos, "Media cloud: an open cloud computing middleware for content management," IEEE Trans. Consumer Electron., vol. 57, no. 2, pp. 970-978, 2011.

[3] S. G. Lee, D. Lee, and S. Lee, "Personalized DTV program recommendation system under a cloud computing environment," IEEE Trans. Consumer Electron., vol. 56, no. 2, pp. 1034-1042, 2010.

[4] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A break in the clouds: towards a cloud definition," ACM SIGCOMM Compute. Commune. Rev., vol. 39, no. 1, pp. 50-55, 2009.

[5] S. Kamara, and K. Lauter, "Cryptographic cloud storage," FC 2010 Workshops, LNCS 6054, PP. 136-149, Jan. 2010.

[6] I. H. Witten, A. Moffat, and T. C. Bell, Managing gigabytes: Compressing and indexing documents and images, Morgan Kaufmann Publishing: San Francisco, May 1999, PP. 36-56

[7] S. Grzonkowski, and P. M. Corcoran, "Sharing cloud services: user authentication for social enhancement of home networking," IEEE Trans. Consumer Electron., vol. 57, no. 3, pp. 1424-1432, 2011.

[8] R. Sanchez, F. Almenares, P. Arias, D. Diaz-Sanchez, and A. Marin,"Enhancing privacy and dynamic federation in IdM for consumer cloud computing," IEEE Trans. Consumer Electron., vol. 58, no. 1, pp. 95-103, 2012.

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.