International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
Index Copernicus Value (2015): 56.67 | Impact Factor (2017): 5.156

Top-K Similarity Search over Encrypted Medical Data Set for Disease Evaluation

Peter Morris[1], Ghilby Jaison[2]

[1]KTU University, Mar Baselios Christian College of Engineering & Technology, Kuttikanam, India.
peterjosemorris[at]live.com
[2]KTU University, Mar Baselios Christian College of Engineering & Technology, Kuttikanam, India.
ghilbyvarghesejaison[at]mbcpeermade.com

Abstract: ***The Data owners using big Data applications outsource their data on cloud servers for the convenience of data management and mining. To protect privacy of sensitive data, documents are encrypted and uploaded to the cloud servers. Searchable encryption is used to encrypt information in documents so that the information can be queried by the authorized users. Symmetric encryption is used here. Based on the relevance of words in search query with the term frequency and inverse document frequency, the documents are retrieved from cloud and provided to the authorized users. The proposed system uses Winner Takes All (WTA) concept of LAMSTAR (Large Memory Storage and Retrieval) to classify and retrieve the documents as per the query requested by the authorized users. Here LAMSTAR is used for medical diagnosis problems that concern large data sets of many categories.***

Keywords: cloud computing, privacy preserving, multi-keywords search, disease evaluation**,** similarity search, Lamstar algorithm, Big data analysis

1. Introduction

Cloud provides easy access to resources and services. Cloud is preferred over local storage of information by data owners as they do not have to worry about data storage and maintenance. Cloud servers are only partially trusted. Documents with secret details cannot be uploaded just like that which would allow other cloud users to access the information that requires privacy. Health records of famous people are supposed to be given privacy. For the privacy of information stored in the cloud servers, data is encrypted before being uploaded to cloud. AES algorithm is used for encryption of data set. But searching the information stored in cloud becomes an issue when data is encrypted. Searchable encryption techniques are used to search data. Multi keyword search is preferred over single keyword search. Based on the relevance scores, the documents are provided to the users.

2. Literature Survey

Cong Wang and Ning Cao[1] uses a Ranked search which makes the system more efficient by providing files in a ranked order based on term frequency. They also use symmetric encryption of documents that are outsourced to the cloud. They also use Inverse Document frequency for calculation of relevance scores can make this a more efficient search. M. Sathya, J. Jayanthi [2] uses k-means to cluster similar documents and retrieve documents based on query. The data is calculated by counting co-occurrence of each word in the document and finding Euclidean distance between documents. Jin Li and Qian Wang, Cong Wang [3] uses a fuzzy keyword sets are used for keyword similarity where the system accepts fuzzy keywords from users to retrieve documents. Mehmet Kuzu, Mohammad Saiful Islam [4] uses a similarity search over encrypted files using Locality sensitive hashing. Xiaofeng Ding, Peng Liu [5] accepts multi keyword from users and performs a similarity search to rank documents.

3. Proposed System

As depicted in Figure1, the system model consists of a User, Data Owner and a cloud server. Data set is encrypted and uploaded to the cloud by the Data Owner for privacy reasons. Data owner shares a token to the registered user. User enters multiple symptoms for analyzing the disease associated with the symptoms. Based on it, the cloud server checks the relevance and documents of diseases are provided to the User.

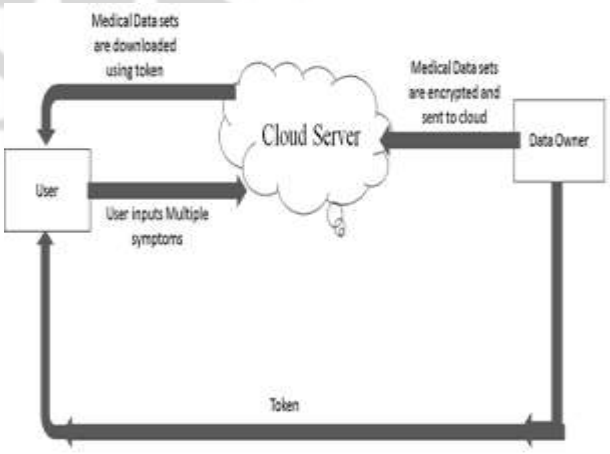

Figure 1: Proposed System

A medical dataset is encrypted by data owner. The users enter multiple symptoms from which the minimum weight is calculated and used to analyze the diseases associated with the symptoms. Based on the minimum weighted symptom, the disease documents are selected using the Winner Takes All (WTA) concept of LAMSTAR. The winner of the multiple symptoms is used to retrieve documents. We need to look

only in the selected documents for the symptoms other than the minimum weighted symbol which users provide. This helps to reduce the execution time as we do not have to perform search in all the documents. From a medical dataset of 4000 documents related to diseases, let's say user enters 3 symptoms. So in total 12000(4000*3) executions are required. In our system we search all documents based on the minimum weighted symbol and only in the documents where the minimum weighted symptom is found, we search for the other symbols provided by the user. If within 300th execution, documents containing minimum weighted symbol are found, we do not have to perform the rest of the search as the other symptoms will also be present in the retrieved documents. Top-K documents based on the weight of symptoms are retrieved. A token is provided by the data owner to the registered users to download the documents that are listed based on the multiple symptoms provided by the user.

3.1 Algorithm

1. Weights of symptoms are calculated using the formulae
WTF (t) = TF * IDF (t) for term t.
IDF= log 2(N/n).
Where:
TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document).
IDF (t) = log (Total number of documents / Number of documents with term t in it).
2. Weights of the symptom keywords are compared and the symptom with least weight is selected.
3. Weights of symptoms that were earlier used for disease evaluation are stored in database along with the diagnosed disease.
4. In the documents selected based on minimum weighted symbol, the other symptoms are searched.

4. Experimental Analysis

To analyze the performance, the program is tested over a medical data set of 3971 data's. As we can see in Figure2, multiple symptoms are entered by a user which is headache and fever. Weights of Headache and Fever are shown as already Exist as they were calculated for disease evaluation earlier and saved in database. The minimum weight of fever and headache is taken in next stage and documents are retrieved based on the minimum weighted symptom, i.e. headache. This is shown in Figure 3.

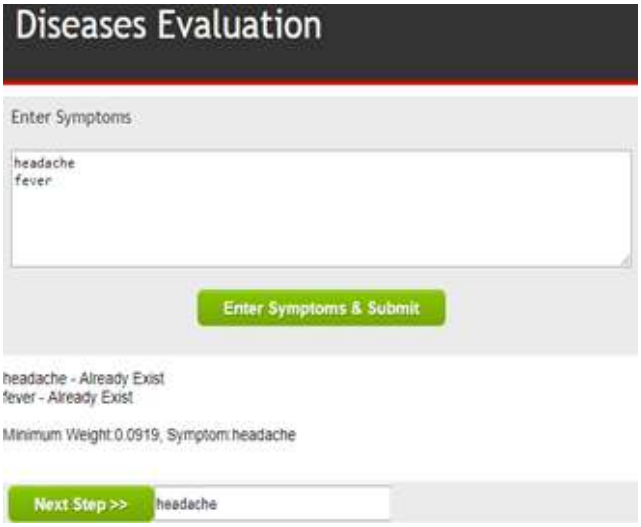

Figure 2: User entering multiple symptoms

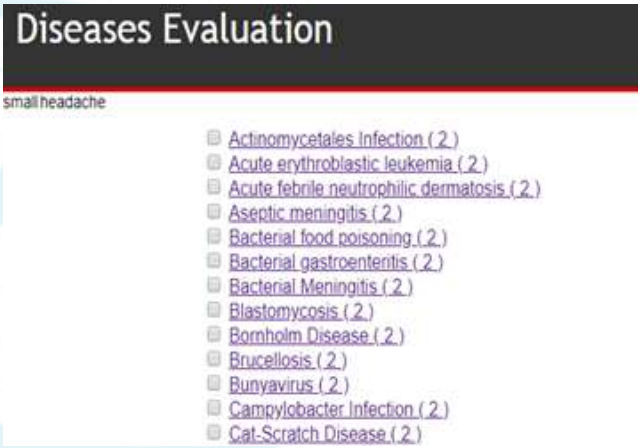

Figure 3: Displaying diseases

5. Conclusion

Efficiency and the security of multi-keyword search over encrypted data is improved. Large Memory Storage and Retrieval (LAMSTAR) concept is used here. Initially multiple keywords are entered as symptoms. The keyword with least weight is selected and this is used to check whether other keywords are contained in diagnosing for the disease.

References

[1] Cong Wang, Ning Cao, Jin Li, Kui Ren, and Wenjing Lou." Secure Ranked Keyword Search over Encrypted Cloud Data." 2010.
[2] M.Sathya, J.Jayanthi, N. Basker." link based k-means clustering algorithm for information retrieval."2011.
[3] Jin Li, Qian Wang, Cong Wang, Ning Cao, Kui Ren, and Wenjing Lou. "Fuzzy keyword search over encrypted data in cloud computing."2010.
[4] Mehmet Kuzu, Mohammad Saiful Islam, Murat Kantarcioglu. "Efficient similarity search over encrypted data"2012.
[5] Xiaofeng Ding, Peng Liu and Hai Jin "privacy-preserving multi-keyword top-k similarity search over encrypted data".2017