Big Data Analytics In Clinical Data using Multi Keyword Search

Aswathy M M, Aathira Susan D’cruz, Hari Narayanan A G

Abstract*:** ***Paper Setup must be in A4 size with Margin: Top 1.78 cm, Bottom 1.78 cm, Left 1.78 cm, Right 1.65 cm, Gutter 0 cm, and Gutter Position Top. Paper must be in two Columns after Authors Name with Width 8.59 cm, Spacing 0.51 cm. Whole paper must be with: Font Name Times New Roman, Font Size 10, Line Spacing 1.05 EXCEPT Abstract, Keywords (Index Term), Paper Tile, References, Author Profile (in the last page of the paper, maximum 400 words), All Headings, and Manuscript Details (First Page, Bottom, left side).Paper Title must be in Font Size 24, Bold, with Single Line Spacing. Authors Name must be in Font Size 11, Bold, Before Spacing 0, After Spacing 16, with Single Line Spacing. Please do not write Author e-mail or author address in the place of Authors name. Authors e-mail, and their Address details must be in the Manuscript details. Abstract and Keywords (Index Term) must be in Font Size 9, Bold, Italic with Single Line Spacing. All MAIN HEADING must be in Upper Case, Centre, and Roman Numbering (I, II, III…etc), Before Spacing 12, After Spacing 6, with single line spacing. All Sub Heading must be in Title Case, Left 0.25 cm, Italic, and Alphabet Numbering (A, B, C…etc), Before Spacing 6, After Spacing 4, with Single Line Spacing. Manuscript Details must be in Font Size 8, in the Bottom, First Page, and Left Side with Single Line Spacing. References must be in Font Size 8, Hanging 0.25 with single line spacing. Author Profile must be in Font Size 8, with single line spacing. Fore more details, please download TEMPLATE HELP FILE from the website. Due to rapidly increasing clinical data, clinics are increasingly outsourcing local data to to cloud servers online– this achieves convenience and also reduces data management expenditure. To ensure privacy, data of a sensitive nature needs encryption before outsourcing. This renders impossible usual data recovery methods such as keyword - based document recovery. We study Big Data Analytics in clinical data employing multi – keyword search, which also supports active update operations such as document removal and addition. To be more specific, we build an index tree that is founded on the sculpture of vector space to search multi - keywords; this supports flexible update operations. In addition, cosine similarity measures are used to support the precise ranking of results emerging from the search. We proceed to formulate a search algorithm – this relies on a Greedy Depth-first strategy -to improve search efficiency. Clinical data experiments demonstrate the efficacy of the proposed scheme.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author

Aswathy M M*, Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi,Amrita Vishwa Vidyapeetham, India.

Aathira Susan D’cruz, Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi,Amrita Vishwa Vidyapeetham, India.

Hari Narayanan A G, Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi,Amrita Vishwa Vidyapeetham, India

Keywords*: *Clinical data, Greedy DFS, Keyword-NNE algorithm, Multi keyword search,

I. INTRODUCTION

The 21st century is an era of information technology, and humanity is faced with an explosion of information. In Wikipedia, the term Big Data is defined as: Big Data means vast collections of data sets that are so huge and complicated that it is hard to approach by means of standard database management tools or conventional applications for data processing. Such big data can help in exploring the underlying mechanisms of phenomenon varieties in all disciplines and facilitate further decision- making. For example, in biomedical field, big data also begin to show its important role. Instead of conventional experience and intuition, medical decision-making depends increasingly on data analysis. Data mining is more or less similar to big data analysis and attempts to extract interesting patterns that are non-trivial, often implicit, previously unanticipated and potentially of great utility from huge quantities of data. Clinical data is a key resource for the majority of medical and health research. Clinical data are collected either during proceeding patient care or in a formal clinical case programme. There are six important kind of clinical data:

- Electronic health Information
- Administrative report
- Declare data
- Patient / Disease registries
- Health reviews
- Clinical case data

Multi Keyword Search

Cloud computing allows customers to store data remotely in the cloud so that high- quality applications and services are available on demand. Cloud storage enables the user to access files from any computer while connected to the Internet. The system model of existing studies takes one data owner into account, which means that data owners and users can communicate and even exchange sensitive information with ease. In a scenario with large numbers of data owners, the exchange of restricted information causes significant overhead. We explore the problem of securing multiple keywords in the computing environment furnished by a cloud to handle the cases of many data owners and many data users. In this article, we address the secure search of multiple keywords over encrypted clinical cloud data. Multi-keyword search is when a user searches for several variations of the same keyword and lists them.

II. EXISTING SYSTEM

The encryption of the data before outsourcing is used to maintain data confidentiality. Search-supporting encryption strategies allow the customer access to the cloud for storing encrypted data and enables search using keywords through the cipher text field. Various threat models have proposed numerous works to achieve different search functionalities - for example, single keyword search, search for similarities, ranked search, Boolean multi keyword search, multiple keyword search, etc. Multi keyword classified search is increasingly focused on practical uses. Some dynamic systems have recently been introduced to support the insertion and removal of document collection operations. These task are vital, since data owners are likely to face the need to update cloud server data. Leading to huge costs in using the data. Existing information recovery techniques based on keywords, although widely used in plain text data, cannot be applied directly to encrypted information. It is obviously impractical to download and decrypt all data locally from the cloud. Because of their high overhead, existing system methods are not practical for the cloud server and user.

III. METHODS USED IN PROPOSED SYSTEM

Our primary intent is to assess the effectiveness of a set of algorithms in order to determine which is best suited for cloud platforms. A detailed study of the search algorithms like Sequential, Binary, MDB Tree and Greedy Depth First Search is presented here.

A. Data Set

We used a collection of clinical data sets available at http://www.cancerimagingarchive.net/, which includes information about particular diseases.

B. Search Algorithms

We have selected four algorithms for evaluation based on the comparative study in terms of performance measures. In the following paragraphs, we describe briefly each of the algorithms selected.

Sequential Search

Here, one examines the first element of the list and then moves sequentially along the list until a match occurs. This match can be a single word - aspiration or the minimum member in the list you are looking for [18]. This variation includes searching for a sorted list for all data value occurrences or for the first occurrence or each appearance of an unsorted list data value.

Binary Search

The target value and the selected arrays mid - element value are compared. If the target value equals the centre, the search concludes returning position. If the target value is lower (higher) than the middle element, one continues the search in the lower (higher) part of the array recursively until either the sought value is found (and the corresponding element position returned), or until a "not found" conclusion is reached [19][20].

MDB Tree

The multidimensional binary search tree (or k-d tree, where k is the dimension of search space) as a data structure for data storage obtained through associative searches. It defines the k-d tree and gives examples. In its storage requirements, it is proven to be quite efficient. An important advantage of this structure is that multiple types of queries can be handled masterly by a single data structure [17]. Several auxiliary algorithms can be conceived; and the average running times while processing a file with n records can be shown to be as follows: Insertion, O(log n); root deletion, O(n (k-1) / k); removal of a random node, O(log n); and optimization, O(n log n). Search algorithms deal with partial match queries with specified t keys and also neighboring queries. An algorithm is presented to tackle any general intersection queries.

Greedy DFS

Greedy DFS must build a result list known as RList, the element of which is defined as (RScore; FID). The RScore is the fFID documents relevant score for the query. The RList stores the documents accessed by the k with the most relevant query scores. The list elements are classified in descending order in accordance with the RScore and are updated in search time.

RScore (Du, Q)- Calculating the value of the query vector Q and the index vector Du stored in node u.

Kthscore- The smallest score of relevance in the current RList, initialized as 0.

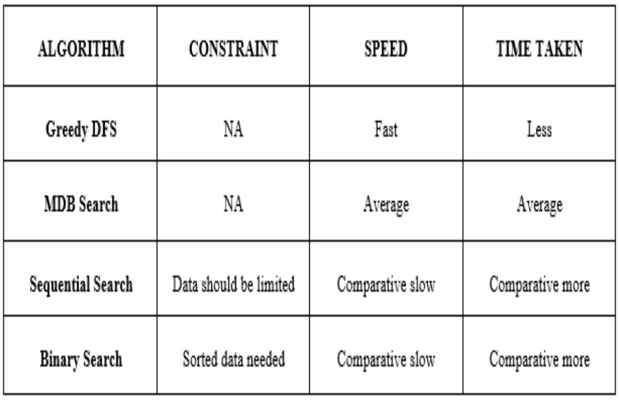
Hchild- A tree node child with a higher score of relevance.

Lchild- the tree's child node with a lower relevance score

IV. .PROPOSED SYSTEM

We examine a program based on a secure tree - based search for encrypted cloud data - it supports search and dynamic operations on collection documents with multi - keyword classification. In particular, the vector space model and the popular "Term Frequency (TF) * reverse document frequency (IDF)" model are combined to provide multi - keyword search in index construction and generation of queries. To maximize search efficiency, we build an index structure based on the tree and propose an algorithm that relies on this index tree called "Greedy Depth- First Search." The secure kNN algorithm encrypt index and query vectors, while ensuring precise calculation of the relevance score between encrypted index and query vectors. We build two secure search schemes to withstand diverse attacks occurring in various threat models: the basic dynamic multi- keyword ranked search (BDMRS) scheme in the known cipher text model and the enhanced dynamic multi- keyword ranked search (EDMRS) scheme in the known background model.

We intend to carry out a work that analyzes a set of search algorithms in cloud platforms to search for clinical data using multi-keyword search. The scheme that we propose implements the first search for clinical data in sequential, binary, MDB Tree and Greedy Depth. We analyzed these algorithms on the basis of clinical data sets based on constraints, speed and time taken.



ALGORITHM	CONSTRAINT	SPEED	TIME TAKEN
Greedy DFS	NA	Fast	Less
MDB Search	NA	Average	Average
Sequential Search	Data should be limited	Comparative slow	Comparative more
Binary Search	Sorted data needed	Comparative slow	Comparative more

I. Comparison between Searching Techniques

A. *System Architecture*

According to the data owner and data user cloud computing model of the block diagram, three entities, such as data owner, data user, cloud server etc... are involved. Data owner possesses the file collection. He builds secure keyword search index and keywords are extracted from files. He submits the server keyword index. He then encrypts files and outsources files encrypted by the server to the cloud. Server receives keyword index encrypted. When data users are looking for files in a cloud server, they first calculate and submit the corresponding trapdoors to the server. A cloud server searches the data owner's encrypted index and proceeds to return the top-k encrypted files to data user. When data users receive top- K files from their cloud server, they download files and decrypt them.

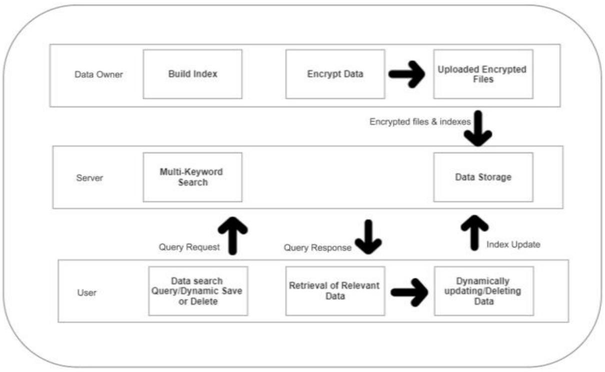


Fig 1: Block diagram of proposed system.

B. Keyword-NNE Algorithm

In preceding work, the BKC algorithm suffers from reduced performance caused by a proliferation of query keywords. To address this issue, a more efficient keyword nearest neighbor expansion (keyword- NNNE) has been developed that uses a different approach. In this algorithm, a query is taken to be a keyword for the query. These objects are linked to the main query keyword and are regarded as the main objects. Keyword-NNE calculates the best local solution corresponding to each main object. The BKC algorithm finds and returns the lbkc's highest assessment. The lbkc can simply select the viewer / customer for each of the main objects few closest and highly rated objects. Compared to the clustering of k-means, the keyword is significantly reduced. This keyword covers an optimal process in the keyword NNE algorithm, and every processed keyword candidate generates very few new keywords. It is an efficient multi- keyword classified search algorithm. This secure algorithm is used for encrypting the vector index and query. We suggest an algorithm that is founded on this index tree "Greedy Depth-first Search." The KNNE performs better than linear search in terms of search efficiency but suffers from achieving a loss of precision.

C. *Greedy Depth First Search Algorithm*

Greedy algorithms make the best choice at every step, as they try to find the optimal way to solve the whole problem. At each stage it uses local optimum to find a global optimum. A greedy algorithm finds one step at a time the best solution for a problem. In each step, the algorithm makes the choice that most improves the solution, even if this choice is less fruitful in future. This sometimes gives you the right solution to the problem. There are several conditions in an algorithm at a time. If it leads to the best optimal solution when we choose the one optimal condition at a single step in the hope, this algorithm is called a greedy algorithm.
The covetous algorithm works on this property:
1. Greedy Property: In every step, it makes the locally optimal solution in the hope that it leads to an optimal global solution.
2. Optimal substructure: Optimal substructure contains an optimum substructure solution.
Adv: easy to install, easy to run and fast.

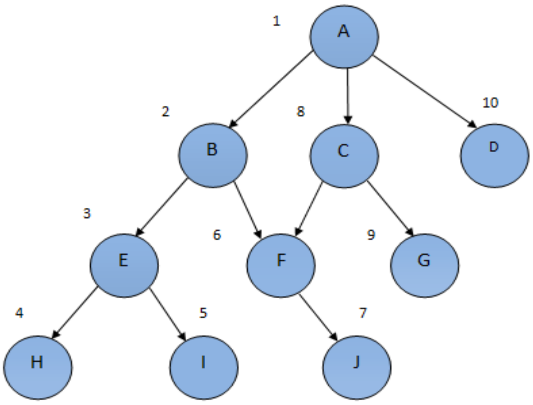


Fig 2: Depth First Search Traversal

Depth First Search algorithm traverses a graph in a deep motion (left subtree first then right subtree) and uses a stack to remember to start a search with the next vertex when an impasse occurs in any iteration.

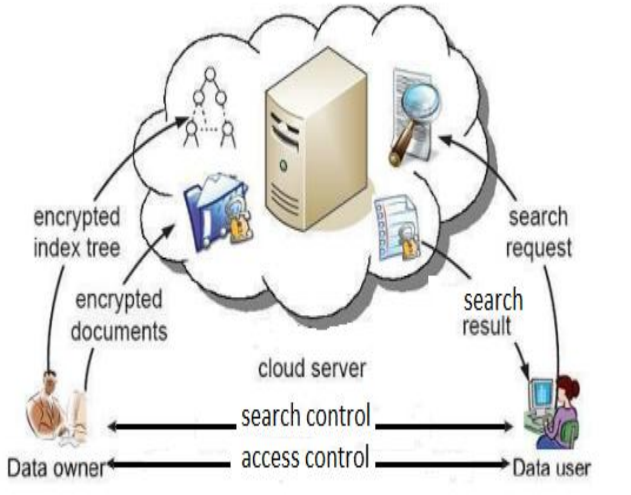


Fig 3: The architecture of multi-keyword ranked search over encrypted clinical cloud data.

We have proposed a secure tree- based search system for encrypted clinical cloud data. This scheme supports multi keyword searching and also dynamic document collection operations.

D. Advantages

Given the specialties in the structure of our tree-based index, our suggested search scheme flexibly achieves sub linear search times; it can also handle the deletion and insertion of documents. Our encryption scheme that supports efficient searches also supports both accurate multi- keywords ranking and flexible dynamic document collection operations. The search complexity of the proposed scheme is essentially kept logarithmic because of the carefully chosen structure of our tree-based index. The proposed scheme can achieve higher search efficiency by implementing our algorithm "Greedy Depth-first Search." In addition, parallel searches can be carried out with greater flexibility and this can lead to a further reduction in search time.

V. Literature Review

Big data consists of a collection of data sets that are both huge and highly complex. The data size is generally Petabyte and Exabyte. This large amount of data cannot be captured, stored and analyzed by traditional database systems. With the internet boom, big data too grows. Big data changes the processing and utilization of data. Applications include healthcare, traffic management, banking, retail, education, etc. Few types of data are also expected to present increasing challenges [12].

In this paper, the problem of multi- keyword searching for encrypted cloud data is defined and solved; we also establish a various privacy requirement. We have chosen the efficient "coordinate matching" similarity measure, i.e., as many matches as possible, to capture effectively the relevance of outsourced documents to query keywords and use "inner product similarity" to quantitatively evaluate this similarity measure [13].

It describes the search for multiple keywords for outsourced cloud data. Specifically, the user specifies multiple keywords considering only text data. From the cloud, files that contain more than a threshold number of keywords or keywords with sufficient similarity are returned; to quantify the concept of similarity, a suitably defined distance metric of editing is employed [14].

In this paper, we have noted a number of shortcomings that have a serious impact on the effectiveness our approach. Driven by this, we develop a novel access methodology called the spatial inverted index that extends the standard inverted index to handle information of a multidimensional nature; there are also associated algorithms that answer keywords in real time to nearest neighboring queries [15].

It describes an efficient search method for data protection over encrypted cloud data using minhash functions. The majority of works available in the literature can only support a single search for features in queries – this reduces efficiency. The ability to search multiple keywords for a single query is one of the major benefits of this method of ours. The proposed method has been shown to meet the adaptive definition of semantic security [16].

VI. Conclusion

We can contribute to two main aspects following the study and analysis of all methods: Multi - keyword search for more precise search results and Greedy Depth First Search algorithm to search more efficiently and dynamically. In large databases, our proposed search technique Greedy DFS is useful. The results show that the proposed algorithm exhibits improvements in search efficiency and time complexity. Finally, by experimenting with real-world data set, we analyze the system's performance in detail. However, there are still some issues, such as reducing the construction time of index tree, etc. In the future, we will carry out more research. In addition, we will scrutinize the etiquette of our suggested multi- user systems. We therefore proposed the problem of searching multiple keywords over encrypted clinical cloud data and establish a variety of safety requirements. We choose the effective principle of coordinate matching from various multi-keyword concepts. We propose first secure internal data calculation. We also achieve an effective ranking result using neighboring technology as close as possible. Currently, this system works on a single cloud and offers greater security in multiuser systems.

References

1. What is big data in clinical https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4283332/
2. What is clinical data https://en.wikipedia.org/wiki/Clinical_data_management
3. How to implement multi keywords in big data https://ieeexplore.ieee.org/document/8254990
4. Advantages and disadvantages https://www.datamation.com/big-data/big-data-pros-and-cons.html
5. Real time and future https://www.ciklum.com/blog/pros-and-cons-of-big-data/
6. Analysis https://www.degruyter.com/view/j/jib.ahead-of-print/jib-2017-0030/jib- 2017-0030.xml
7. Applications
8. https://www.datapine.com/blog/big-data-examples-in-healthcare/
9. Survey of Clinical Data Mining Applications on Big Data https://ieeexplore.ieee.org/document/6786154
10. Challenges https://www.omicsonline.org/open-access/big-data-science-and-its-applications-in-health-and-medical-research-challenges-and-opportuniti

es-2155-6180-1000307.php?aid=75506

11. Existing system written from challenges https://healthitanalytics.com/news/top-10-challenges-of-big-data-analytics-in-healthcare
12. DFS Algorithm applied in multi keyword search https://www.ijert.org/a-secure-greedy-depth-first-search-algorithm-for-encrypted-data-in-cloud-computing-environment
13. "Big Data Security Issues and Challenges"-Raghav Toshniwal, Kanishka Ghosh Dastidar, Asoke Nath-International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 2, Volume 2 (February 2015).
14. "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data "- Ning Cao, Cong Wan, Ming Li,Kui Ren,Wenjing Lou-2011 Proceedings IEEE INFOCOM.
15. "Privacy-Preserving Multi Keyword Similarity Search Over Outsourced Cloud Data "-Chia-Mu Yu, Chi-Yuan Chen, and Han-Chieh Chao-IEEE Systems Journal (Volume: 11 , Issue: 2 , June 2017).
16. "Fast Searching With Keywords Using Data Mining"-Chandrashekhar-International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)Vol. 2, Issue 2, pp: (82-99), Month: April-June 2014.
17. "A Practical and Secure Multi-Keyword Search Method over Encrypted Cloud Data"-Cengiz Orencik, Murat Kantarcioglu† and Erkay Savas-2013 IEEE Sixth International Conference on Cloud Computing
18. Ondreicka, M, Pokorn´y J,"Extending Fagin's algorithm for more users based on multidimensional B-tree", In: Proc. of ADBIS 2008, LNCS 5207, 2008, pp. 199-214.
19. Heydari, J, "Quickest sequential search over correlated sequences", ECSE Dept., Rensselaer Polytech. Inst., Troy, NY, USA.
20. Zhenzheng Ouyang, Quanyuan Wu, Tao Wang "An Efficient Decision Tree Classification Method Based on Extended Hash Table for Data Streams Mining", Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on, On page(s): 313 - 317 Volume: 5, 18-20 Oct. 2008.
21. Tao Wang, "A new fuzzy decision tree classification method for mining high-speed data streams based on binary search trees" ,Nat.Univ. of Defense Technol., Changsha, Second international conference on System , 2007. ICONS '07.

AUTHORS PROFILE

Aswathy M M PG Student, Master of Computer Application, Department of Computer Science & IT, Amrita School of Arts & Sciences, Kochi, Amrita Vishwa Vidyapeetham, India.

Aathira Susan D'cruz PG Student, Master of Computer Application, Department of Computer Science & IT, Amrita School of Arts & Sciences, Kochi, Amrita Vishwa Vidyapeetham, India.

Hari Narayanan A G Assistant Professor, Master of Computer Application, Department of Computer Science & IT, Amrita School of Arts & Sciences, Kochi, Amrita Vishwa Vidyapeetham, India.