International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

Nearest Keyword Set Search Queries on Multi-Dimensional Datasets

Ms Rachael Christina M.S
M.Tech (CNE) Department of CSE
B.N.M. I.T-Bengaluru, India

Dr. Vimuktha Evangeleen Sails
Department of CSE,
B.N.M.I.T – Bengaluru, India

***Abstract* – Data mining is an important aspect in refining the data, this makes the data simpler and easy to use and it is a widely used technique to extract the useful information from huge chunks of data. The application developed is a user friendly application which helps to extract the informative keywords from text files and associates keywords with the file and the searching and retrieval of file is made easier with nearest keywords in a multidimensional data set of files. The integration of data mining concepts along with the cloud storage security techniques lead to the development of an efficient search application to retrieve files from the repository using easily remembered keywords.**

Keywords: Data security, keyword extraction, multidimensional data set.

I.INTRODUCTION

Present world is a technology centric world where each and every small works carried out in day to day life is dependent on the technology. Technology is rapidly growing such that it has occupied major portion in human routine. In this technological era the dependency of people over the information is growing hand in hand with the technology. Each and every activity is dependent on accessing the information and processing with it to get the required result. Starting from a small home to large organisation storing and retrieving of information has become an inseparable part of one's routine. Researches' have stated that on a whole the collection of data includes 10% of structured data and the remaining 90% are unstructured data. Information retrieval from structured data set is simpler compared to information retrieval from unstructured datasets. In order to discover the hidden patterns in the unstructured dataset data mining techniques are used.

Data mining is the process in which the patterns are extracted from the data set and the discovered patterns are analysed and used in studies. Machine learning, statistical analysis are some of the domains which uses the data mining techniques. In this application the data mining is applied over the text documents. The text documents with varied contents without any specific patterns to extract are being considered. A multidimensional data is the type of data in which there is heterogeneous data type objects that are grouped under certain attributes. But when the text document with no specified patterns are considered the pattern recognition becomes impossible, in such cases the keywords extracted from every document is one dimension of the document that describes the feature of the document.

II. MOTIVATION

Over time the collection of documents increases and remembering all file names is impossible for normal human brains and retrieval of file is impossible if file name is unknown, but one can actually be aware of the searching content that is present in a file. In this application the searching based on the keywords document with the associated keywords are easier to search as data consumer can easily interpret about the content of the document rather than the file name. An application to search the documents along with the secured storage for documents is very necessary in every organisation and its implementation is shown in section four and the literature study is shown in section three, design in section four and conclusion in section six.

III. LITERATURE SURVEY

First technique is and it mainly on computing exact nearest and farthest neighbour which is a challenging task, especially in the case of high-dimensional data. Many techniques are used to solve the nearest neighbor problem but not much importance is on farthest neighbor problem. By the calculation of the farthest neighbour a clear idea is obtained for the elimination of unrelated objects to the query there by it helps in giving the result more accurately.[1]

Multidimensional text cube analysis is another technique which is used to analyse the textual documents and the analysis is done by applying the data mining technique over the documents in order to extract the hidden patterns out of it.[2]

Collective spatial keyword technique is another technique which is used to derive the result such that more than one object is required to satisfy the user's query in such cases one node is being considered as owner and other two nodes as sub objects that matches closely with query keyword. This is helpful in developing navigational search query applications where more than one object is necessary to satisfy the user's need.[3]

Cloud computing being one of the blooming technologies provides various services one such facility is the storage at minimum cost which makes most of the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

organizations to store their data inside the cloud but there are chances of cloud being vulnerable to attacks. There is a technique called n-keyword search where the n is total number of distinct keywords present in all the documents and scalar product is performed over the co-ordinate keywords and homomorphic key exchange is done between the participating entities such that transmission is done securely. [4]

The literature survey of these papers helped in developing an idea to implement a search application over multidimensional data set.

IV.SYSTEM DESIGN

Any project before being implemented must be designed such that it gives a complete view of the entire project. The workflow is the main part of project design which is step wise explanation to the flow of project. The flow diagram for the project is shown below in Figure 1.

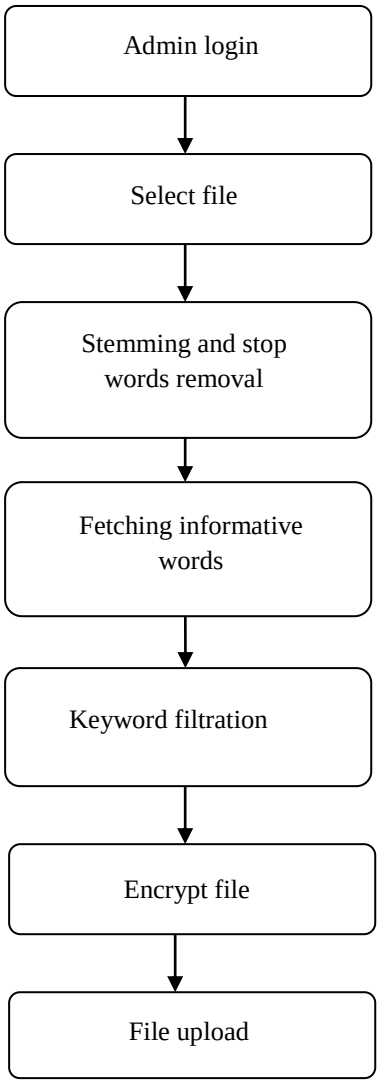


Figure 1 Work flow of admin

Figure 1 shows the work flow of admin module. There are six operations carried out at the time of uploading a file. The specified file is selected and file is being sent in to stop words removal and stemming process informative keywords are fetched and keyword filtration is done. The file is uploaded along with its associated keywords. The workflow of user is shown in figure 2 once user registers and gets the user ID and password and decryption key to the mail box from admin, then user can search the file by decrypting the key given to him and search file with the keywords and download if needed.

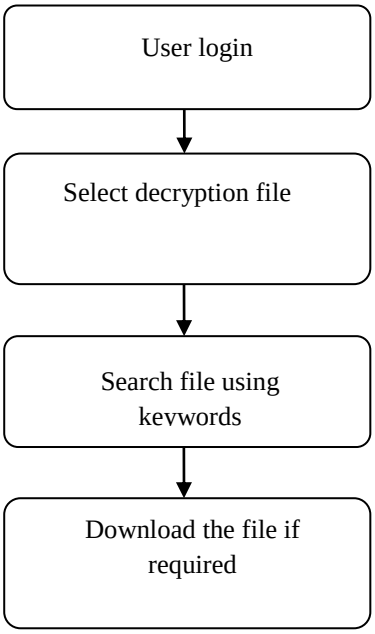


Figure 2 Work flow of user

Algorithm for upload process

Step 1: Start
Step 2: Read File (F) and access permission based on category(C).
Step 3: Remove unnecessary words and special characters.
Step 4 : Shortlist the Keywords.
Step5:UsingTermFrequency(TF) calculates weight age for Keywords.
Step 6: Let N be the number of category allowed to access.
Step 7: For I=1 To N.
Step 8: Fetch the hash key of I^{th} category.
Step 9: Using hashing technique with fetched hash key, generate keyword hash tags for all the keywords.
Step 10: Insert all keyword hash tag into index.
Step 11: Repeat from Step 7 to Step 10 up to I=N.
Step 12: Stop

Algorithm for search process

Step 1: Start
Step 2: Get the Keyword (K) from User (U).
Step 3: Find the User Category (UC).
Step 4: Fetch Hash key for UC.
Step 5: Generate Trapdoor using UC hash key.
Step 6: Search trapdoor on index array.
Step7: Filter all the matched index elements.
Step 8: Shortlist filter from filtered index.
Step 9: Using Inverse document Frequency Rank the files.
Step 10: Display the file list to user.
Step 11: Stop

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

The above are the algorithms for upload and search processes respectively. The upload process is carried out by the data provider and downloading happens by the data consumer.

V. PROPOSED MODEL

The system architecture of the search application developed is shown in figure 3. The two main modules are admin and the user. This is a multiuser environment based application. Once the user is authorised the user can get access to the files through the internet if the application is installed in the user's system. Admin is the data owner who maintains the user details and uploads the files in to the database. Once the registration is done the user gets a decryption sent to the mail id. If a user wants to retrieve the files then the decryption key sent to the user's mail is decrypted then the search box opens up and the user can search the file using the predicted keywords.

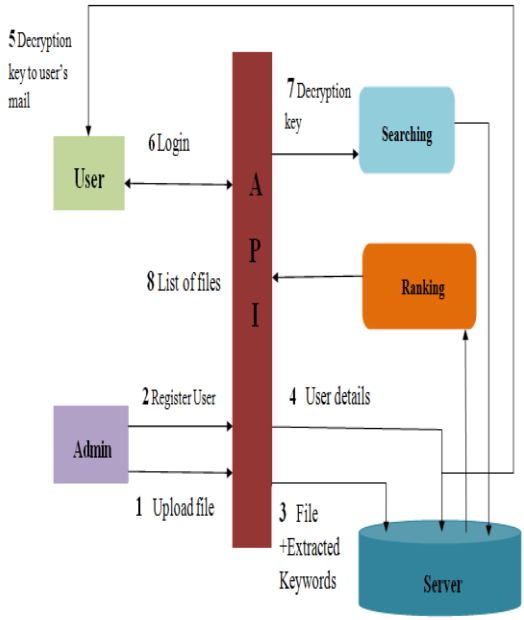
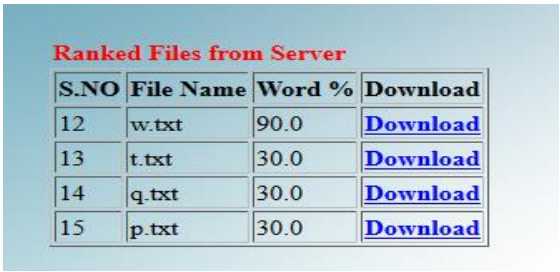


Figure 3: System architecture

Admin being the data owner maintains records and at the time of uploading files the admin is given the authority to select the grade such that the specific file being uploaded can be viewed only by the user who belongs to same grade. This facility is provided with a security point of view and also to reduce the searching process which happens only within the particular grade to which user belongs. The searching process need not consider all the files inserted in the database instead it is sufficient to carry out the searching only within the particular grade. Once the user logs in the user needs to decrypt the key which is provided by the admin at the time of the registration. Once the decryption of key is done then the search box is being opened to the user, where the user can type the keywords either multiple or single keyword. Once the matching of keywords is done the respected files are displayed such that the file with the highest rank will be at the top followed by files with least ranking.

VI. RESULT ANALYSIS

The result of the work carried out is to get the list of files based on the keywords weightage in the form of ranked list as shown in figure 4. The test cases for the work are shown in figure 5.



Ranked Files from Server

S.NO	File Name	Word %	Download
12	w.txt	90.0	Download
13	t.txt	30.0	Download
14	q.txt	30.0	Download
15	p.txt	30.0	Download

Figure 4 : List of files

Test cases	Input	Result
Keywords	Keywords matched with index keywords	PASS
	Keywords spelt wrong or not matched	FAIL
Number of keywords	Keywords less than five	PASS
	Keywords more than five	FAIL
Type of documents	Text file	PASS
	Pdf, word document, excel, image file.	FAIL

Figure 5 : Table of test cases

The analysis is done based on the uploading and downloading time. The uploading time take for any file size is always greater than the downloading time. Uploading takes more as the stemming and stop words removal are happening at the time of uploading. And the downloading time is lesser than the uploading time this because there is no much steps involved in downloading a file hence file search is efficient using this application.

The time recorded for various sizes is shown in figure 6 and the bar chart in figure 7 clearly depicts that for any size of files the uploading time is always greater than downloading time.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

File size in KB	Uploading time in milliseconds	Downloading time in milliseconds
4	10	6.46
20.5	20.206	14.567
30.5	15.157	13.634
2.8	6.399	5.637
16.2	11.518	7.42
11	8	4.027
21	12.672	8.9
31	12.573	8.1
10	8.736	3.9

Figure 6: Table of time records

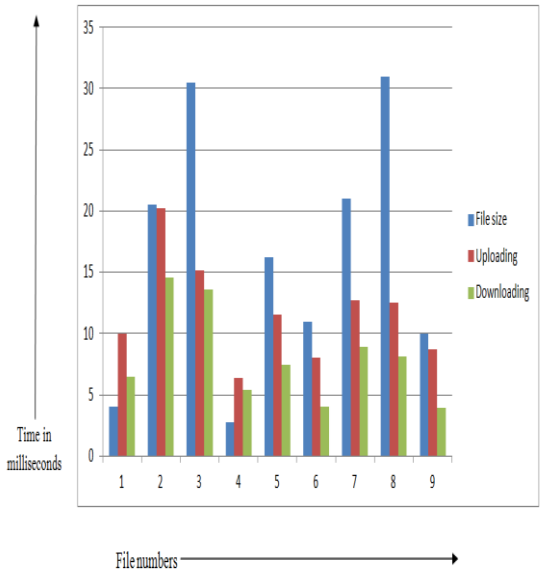


Figure 7 : Bar chart representation for recorded time

VII. CONCLUSION

The application developed is useful in any organization where the huge amount text data is stored in the form of files, so late at the time of retrieval the work of user becomes easier by searching the file with the keywords. The application can further be developed to take inputs such as images, audio, video and documents other than text files and also the application can be developed in to an android app which reduces the burden of the user to a greater level.

VIII. REFERENCES

[1] Yifan Hao, Huiping Cao, Yan Qi, Chuan Hu, Sukumar Brahma, Jingyu HanNew Mexico State University, Las Cruces, "Efficient Keyword Search on Graphs using MapReduce" 2015 IEEE International Conference on Big Data (Big Data)

[2] Cheng Long Raymond Chi-Wing Wong Ke Wang Ada Wai-Chee Fu The Hong Kong University of Science and Technology Simon Fraser University The Chinese University of Hong Kong "Collective Spatial Keyword Queries: A Distance Owner-Driven Approach" SIGMOD'13, June 22–27, 2013, New York, New York, USA. Copyright 2013 ACM 978-1-4503-2037-5/13/06

[3] Shafin Rahman Department of Electrical & Computer Engineering North South University Dhaka, Bangladesh Mrigank Rochan Department of Computer Science University of Manitoba Winnipeg, Canada "Fast Farthest Neighbor Search Algorithm for Very High Dimensional Data" 19th International Conference on Computer and Information Technology, December 18-20, 2016, North South University, Dhaka, Bangladesh

[4] Fangbo Tao, Kin Hou Lei, Jiawei Han, ChengXiang Zhai, "EventCube: Multi-Dimensional Search and Mining of Structured and Text Data" KDD'13, August 11–14, 2013, Chicago, Illinois, USA. Copyright 2013 ACM 978-1-4503-2174-7/13/08

[5] Gandeevan Raghuraman Pooja Nilangekar Pallavi Vijay Kavya Premkumar Saswati Mukherjee "Cloud based Privacy Preserving Efficient Document Storage and Retrieval Framework"1234Department of Computer Science,Anna University, Chennai 5Department of Information Technology, Anna University Chennai 978-1-4673-8200-7/15/ ©2015 IEEE