ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Efficient Group Based Data Retrieval from Cloud Storage using Data Mining Technique

Ankita V. Prajapati
Department of Computer engineering
U.V.Patel College of Engineering
Kherva, Mehsana, India

Prof. Dipak C. Patel
Department of Computer engineering
U.V.Patel College of Engineering
Kherva, Mehsana, India

Abstract: Now a day's user of computer system need everything on hand without location dependency with least cost and efficiently. The Cloud computing comes with lots of benefits like the user can store lots of information on Cloud server and able to access from anywhere, anytime. With this important advantage, this technology also has some issues like, Security of user information, quick retrieval of useful information from the various large amounts of data and many more. Among these when we consider retrieving of relevant information from the thousands of documents, then it became tedious process if we can't take care of this issue. When user need to search anything and the query is made to search in all the available documents, then it take large amount of time which make users job tough. So, to increase the use of Cloud and to deal with the various issues related to cloud environment as mentioned above, here an efficient searching mechanism from Cloud is proposed in which user can get required details quickly without getting any type of burden. Our scheme will provide efficient searching, Group based file retrieval using efficient mining technique which will reduce the time taken related to retrieval of important information.

Keywords: Data Mining, Cloud Computing, Vector Space Model, Keyword Based Retrieval, IaaS

I. INTRODUCTION

On the Internet, a large amount of data which is distributed, heterogeneous, dynamic, and more complex. Every day people have to deal with targeted advertising, by using data mining techniques organization become more efficient by lessen costs. Large amount of data are not handle by Traditional Data Storage systems and also difficult for traditional analytic tools to analyze the large amount of Data. So Cloud Computing is capable of solving the problem of storage, analyzing and handling the data on a distributed network. In recent years, Cloud computing service (CCS) is becoming one of the most important factors in our daily lives. With CCS, we can use a large number of applications via portable computing devices (PCDs) or personal computers. Cloud storage service (CSS) is a special form of CCS. With CSS, we can store various data in the CSS servers for free via public networks, and access the data anywhere and anytime [1].

In cloud computing environment applications and techniques of Data mining are very much needed because data privacy, data security and efficient retrieval of the data from the cloud storage is major issues while storing the data in a Cloud environment. So implementation of data mining techniques in Cloud computing will allow the users to quickly retrieve meaningful information from virtually integrated data warehouse that decreases the costs of storage and infrastructure [1], [7], [8].

A. Literature Review

Ning Cao, Cong Wang, Jin Li, Kui Ren proposed a scheme for securely keyword based searching mechanism. The user send a query and Cloud server retrieve matched file which contains the keyword and send back to client. They used privacy preserving mechanism to protect their data from unauthorized access. They compared proposed cryptographic primitive with the existing primitive cryptographic mechanism. The authors proposed data security but it is applicable to single keyword search. The proposed scheme name order-preserving symmetric encryption (OPSE) they guaranteed that this scheme works efficiently as compared to other similar approaches available. They also shows the practical experimental result for displaying efficiency of proposed scheme [1],

[2].Vijay Lakshmi P, D.Pratiba, Dr.G Shobha [3], proposed a model where they focus on file searching based on multiple keywords. They argued that the traditional keyword based search supports Boolean search which shows weather file may contain the keyword or not, without any relevance of data files. And the ranked based file retrieval using a single keyword having a poor result. Authors also mentioned that ranking on server side which is based on order-preserving encryption (OPE) breaks privacy of data. So, the authors provide scalable system with minimize information lekage.Their model prevent overload by working at user side for ranking files, where consume less bandwidth. They perform analysis which shows efficiency of their proposed solution.

Prof.G.M.Bhandari, Ms M.R.Girme, reviewed and described that, now today user remotely store own secret data on cloud .So in cloud computing, the authors focused on encrypted data which are remotely stored. In this authors argued that the user can search into the encrypted data using keywords without decrypting it in traditional searchable encryption schemes. These techniques of searchable encrypted data using keyword support Boolean search method, which is not sufficient. So the authors proposed secure ranked keyword search over large amount of data files which are in encrypted form in cloud, in which user retrive the rank-ordered file. So, authors also define OPSE technique and one to more order preserving mapping for retrieve efficient data from cloud [1],

[4].Krishna Challa, Rohit Handa, Rama Krishna Challa, reviewed that User can store infinite data on cloud by using limited setup and minimum usage cost. Due to the availability of resources at fast internet speed and low initial investment, the companies are motivated to store their data on the cloud. Authors propose a multi-key word search scheme which is a cluster based over encrypted cloud data. In this research paper

the formation of cluster is done by the client. Authors shows the simulation results based on average searching time and comparisons required to retrieve the desired documents from the cloud server. The proposed search scheme preserves the security requirements as proposed by the existing approaches in literature but provides efficient result [1], [5].

II. PROBLEM IDENTIFICATION

A. *Existing Issues*

The traditional searchable schemes allow user to search in the cloud data through keywords, which support only Boolean search, i.e., whether a keyword exist in a file or not. As user try to search various pages/details by using multiple keyword (which can be too general), and as a result user may get unwanted pages/detail that may be retrieved due to general query submitted to search engine. So user have to go through every retrieved file which may results in visiting of unwanted pages which will increase time as well as increase network traffic. User also find some difficulty to fetch the accurate results because they need to search everywhere even the document is not at all related to their query or their area of interest because of lack of group based searching mechanism. The formation of the clusters are done at the client side so that, it becomes very time consuming and also the involvement of users is required and it becomes very difficult to handle [2],[3],[4],[6].

B. *Problem Statement*

Efficient cluster based documents retrieval system can be developed in which there should not be any overhead for cluster formation at client side and time of document retrieval should be minimized. So we have proposed an approach that can retrive data efficiently from massively large amount of cloud storage.

III. PROPOSED EFFICIENT GROUP BASED DATA RETRIEVAL FROM CLOUD STORAGE USING DATA MINING TECHNIQUE

The proposed system of efficient group based data retrieval from cloud storage using data mining technique can be described using following two models, first model is general description of interaction between the cloud server and users and the second model represents proposed keyword based retrieval from cloud storage step by step in detail.

A. *System Architecture*

As shown in the following figure 1, the cloud computing system having three different entities like Cloud server, Data User and Data owner. The Data Owner have a collection of the files that He/she wants to upload onto the cloud server. The Data Owner also expects that the cloud server provide the keyword based retrieval service to the data owner or the data user. The data user have an authorization to get the files which they want from the cloud server by using multikeyword search [4], [6].

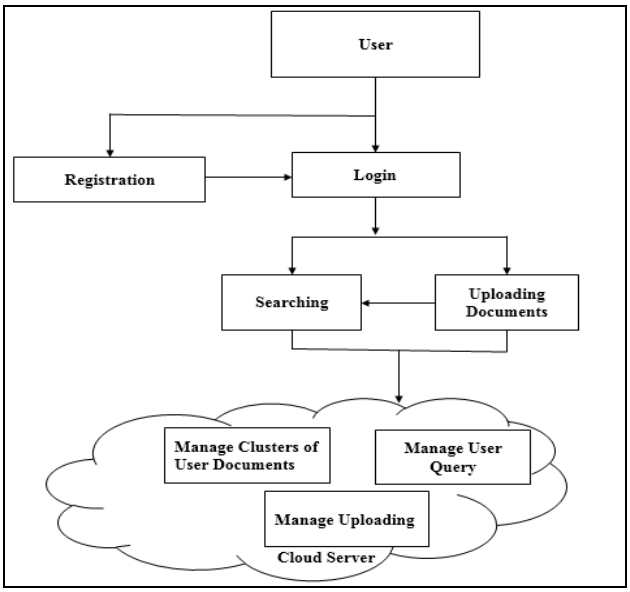

Fig.1: System Architecture

B. *Keyword Based Retrieval from The Cloud Server*

The following figure 2, represents the proposed keyword based searching mechanism from the cloud server. The user sends a query on the server at which first server checks the collection of documents which are already uploaded by different users or organizations. Now we apply the k-means clustering technique to form cluster which gives accurate result based on the user keywords given in various documents .After that the user query to retrieve desired documents is converted into the vector by using Tf-idf scoring. Later on the Euclidian distance is calculated between the query vector and centroid of the cluster and the documents in the cluster having a minimum distance is given to the user as a result [9],[10].

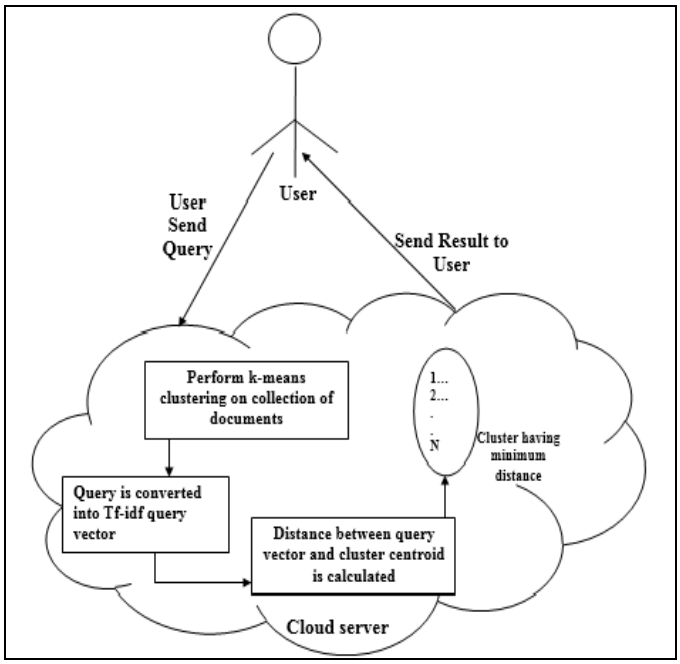

Fig.2: Keyword Based Retrieval From the Cloud Server

IV. EXPERIMENTAL SETUP AND RESULTS

An experiment has been conducted by using the different documents from the given system. For the performance analysis the different number of documents with different number of keywords are used. The experiment were performed on Core 2 Duo Processor with 4GB RAM. The entire implementation is done using JAVA language. For simulation we have used CloudSim simulator. The results for the existing search scheme and the proposed search scheme are obtained on this system and thus used for the comparison.

A. Average time required to perform a search

As the query vector is compared with the centroid of the cluster instead of directly with the document Tf-idf vectors, So there are the fewer number of comparisons are required as compared to the existing approach. The following Figures represent the average search time required to search the desired document by using the different number of keywords as query.

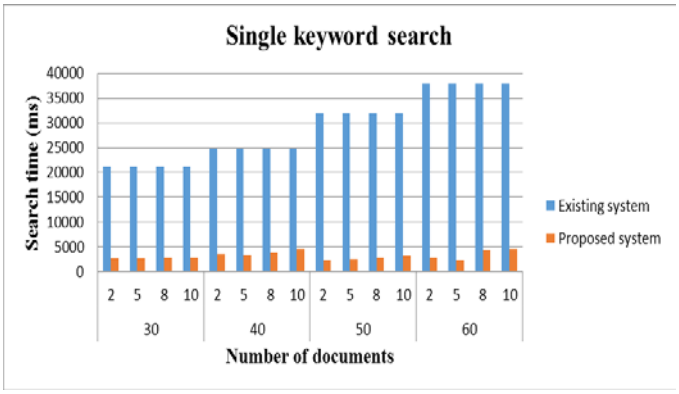

Fig.3: Average Search Time for Single Keyword Search

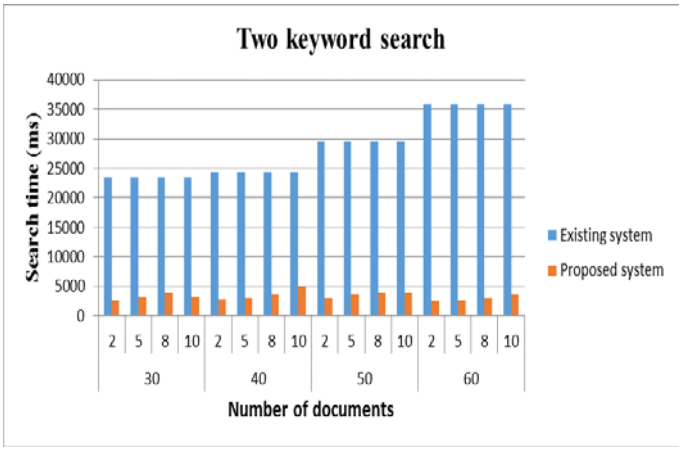

Fig.4: Average Search Time for Two Keyword Search

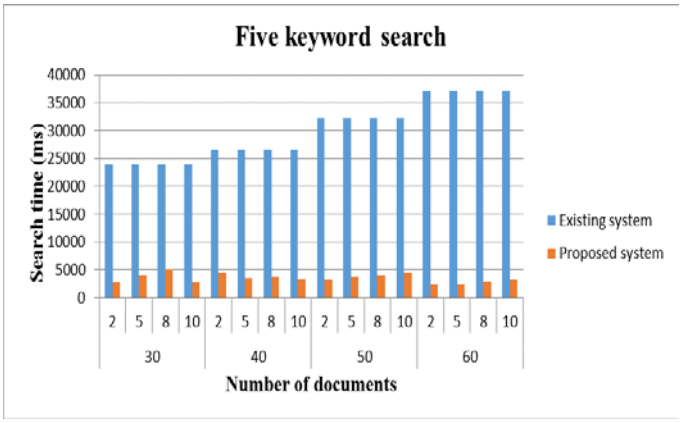

Fig.5: Average Search Time for Five Keyword Search

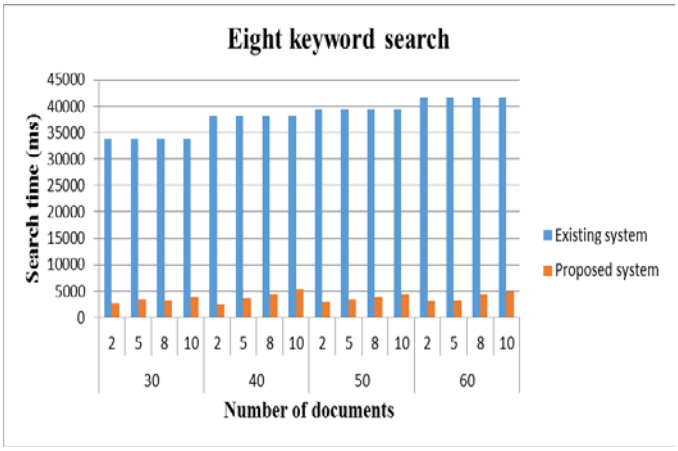

Fig.6: Average Search Time for Eight Keyword Search

B. Analysis of performance

Based on the above results we must say that as we create clusters and put the document in cluster based on the Euclidian distance and it gives efficient result. The system work faster because the number of documents in which search is to be performed is minimized. The analysis also shows even we increase the number of cluster the system performance is stable while comparing with existing system where there is no any clustering mechanism applied. We also analyze the system by increasing the number of keywords which is sent as a query by the user to search the required documents in the available documents. As we increase the number of keywords the system time will be little bit increase but the performance is not degrades its show that the proposed system work effectively even the case of number of cluster or number of keyword or when we can take multiple combination of this two parameter. The following figure 7, shows the number of keywords given by user as a query to search the desired documents and the next figure 8, represents the list of the documents given to the user as a result to the query.

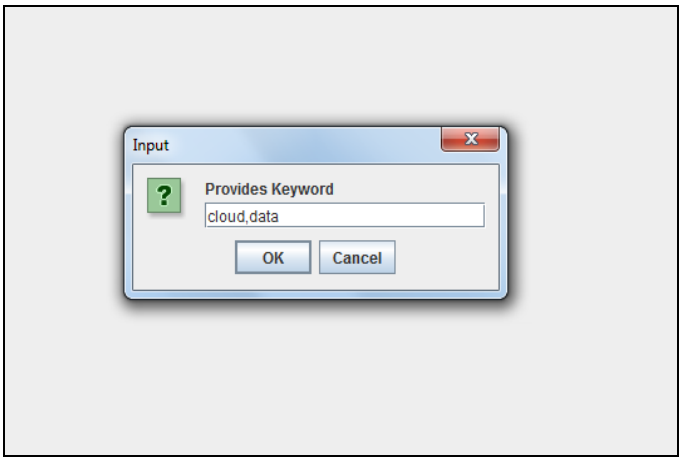

Fig.7: Searching Of Document Based On Keyword

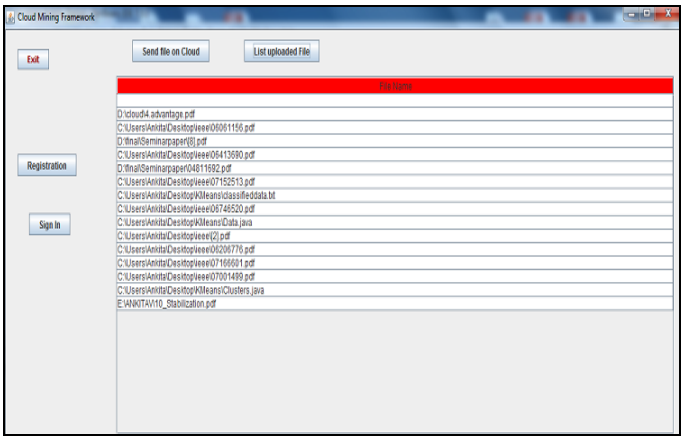

Fig.8: List of The File Retrieved To User As A Result

V. CONCLUSION AND FUTURE SCOPE

Proposed scheme uses Data Mining technique for grouping of documents/information which will retrieve needed documents quickly. The system is scalable because proposed scheme produce efficient result if more users/documents are added. As the Cloud servers are high performance system which gives higher performance with least execution cost, we have formed clusters at server side to reduce the burden at client side. Multiple keyword searching is also one of the important points in proposed scheme. So, overall the proposed method lessens the burden of user by giving scalability and provides the time efficient searching of relevant documents by its own side.

We suggest as a future work that, Instead of uploading keywords manually system can be made to find important keywords of documents automatically. Security to the uploaded documents can also be provided by introducing efficient security mechanism.

VI. REFERENCES

[1] Dipak C.Patel, Ankita Prajapati, "A Survey: Efficient Group Based Data Retrieval from Cloud Storage Using Data Mining Technique", International Journal of Innovative Research in Computer and Communication Engineering-Volume 5.Issue 4, pp.6998- 7001, April 2017.

[2] Ning Cao, Cong Wang, Jin Li, Kui Ren, "Secure ranked keyword search over encrypted cloud Data", IEEE International Conference of Distributed Computing Systems (ICDCS), Genoa, Italy, pp.253–262, 2010.

[3] D.Pratiba, Dr.G Shobha, Vijay Lakshmi P,"Efficient Data Retrieval from Cloud Storage Using Data Mining Technique", International Journal on Cybernetics and Informatics-Volume 4.No 2, pp.271- 279, April 2015.

[4] Ms. M.R.Girme, Prof.G.M.Bhandari, "Efficient Ranked Keyword Search for Achieving Effective Utilization of Remotely Stored Encrypted Data in Cloud", International Journal of Application or Innovation in Engineering and Management –Volume 3, Issue 6, pp.105-113, June2014.

[5] Rohit Handa,Rama Krishna Challa,"A Cluster Based Multi-KeywordSearchon IEEE International Conference on Computing for Sustainable Global Development, pp.115-120, 2015.

[6] Akshay D Kapse, Piyush K Ingole, "Secure and Efficient Search Technique in Cloud Computing", IEEE Fourth International Conference on Communication System and Network Technologies, pp.743-747, 2014.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[7] R.Kabilan, Dr.N.Jayaveeran,"Survey of Data Mining Techniques in Cloud Computing", International Journal of Scientific Engineering and Applied Science -Volume 1, Issue-8, pp.123-127, November 2015.

[8] Zhu Jia, A, Zhang Ping,"Design and Implementation of Data Mining Platform Based On the Cloud Computing", IEEE Workshop on Advanced Research and Technology in Industry Application, pp.163-165, 2014.D. Kornack and P. Rakic, "Cell Proliferation without Neurogenesis in Adult Primate Neocortex," Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.

[9] Ms Aishwarya S. Patil, Ms Ankita S. Patil," A Review on Data Mining Based Cloud Computing", International Journal of Research in Science & Engineering E -Volume 1, Special Issue: 1, pp.1-4, 2014.

[10] Zdravko Markov, Daniel Larose, "Data Mining the Web – Uncovering Patterns in Web Content, Structure and Usages", April 2007.