Cloud Computing and Data Mining

Priyanka Garg
Assistant Professor Department of Computer Science & Engg. Gian Jyoti Group of Institutions,Shambhukalan,Patiala

Abstract:
The storage of data is moving from personal hard disks to cloud , where the amount of data is ever increasing. In this paper we look at the challenges of mining the data in cloud storage. We then analyze how the streaming K-means and D stream algorithm can be used to effectively mine the data in the cloud storage.

Keywords: Data Redundancy; Dynamic clustering; streaming K-means; D-stream.

INTRODUCTION

The cloud computing is a rapidly growing technology . It is growing at a annual rate of 36% [1] and is estimated to have a market size of 19.5 billions $ by the year 2106. From home users to small businesses to large corporations, everyone are migrating to cloud storage solutions. Popular examples of cloud storage include Dropbox, Google drive, Live Drive etc. large enterprises use a private cloud. This is due the following reasons.

1. Data stability: The data in cloud is secure and maintenance free. Companies just need to pay and store the data.
2. They need not worry about the hardware to store the data.
3. Anytime anywhere access: The data on cloud can be shared instantaneously to anyone in any part of the world and can be accessed at anytime.
4. Low cost: The storage solutions are economically viable as one pays only for what one uses.
5. Scalability: The data space can be upgraded or downgraded at any time without any hassle.

This growth is the storage leads to a mammoth size of data. The problem of carrying statistical analysis on the data becomes a huge problem. In the next section we look at the problems encountered in data mining in the cloud storages.

2. PROBLEMS WITH DATA MINING IN CLOUD STORAGES

Traditionally data mining was done on databases that were static or fixed. The data analysis was done usually using clustering algorithm such as K-means or K-Medoids. The clustering algorithms required that the no. of cluster was fixed before the algorithm was run. The clustering algorithm were unsupervised machine learning algorithm and in an iterative process, they clustered the data. The iterative process of the algorithm was repeated till the files did not switch clusters.

The above solution cannot be applied to the cloud storages because of the following problems.

1. The data storage will not be static i.e. data will be coming at huge rates into the database. Hence a clustering algorithm will not converge on increasing data.
2. Redundant data: The incoming data into the cloud might sometimes be the files already stored. Running algorithm like K-

means on redundant data simply increases the algorithm time.

We look at the methods to solve the above problems in the next section.

3. PROPOSED METHOD

A. Solving Data Redundancy problems

To reduce the size of the data for clustering, we propose a method to reduce data redundancy (multiple files with the same data). In algorithm like K-Means a file is represented as a vector and is represented as a single point in a vector space. Having similar files makes the clustering algorithm calculate two distances for the same points got by same content files, stored in two different places.

To solve this problem, a table with the columns filename, its hash and keywords is created. Format of the table is as shown in Table 1. As soon as a file enters the cloud the process as shown in the flowchart in Figure 1 is run on it.

File Name	Hash Value	Keywords

Table1: Index Table Format

The file is first pre processed to remove articles, pronouns and prepositions. Then the keywords in the document are listed. The file's hash value is calculated using any hash algorithm such as MD5. The values are stored in the hashtable.

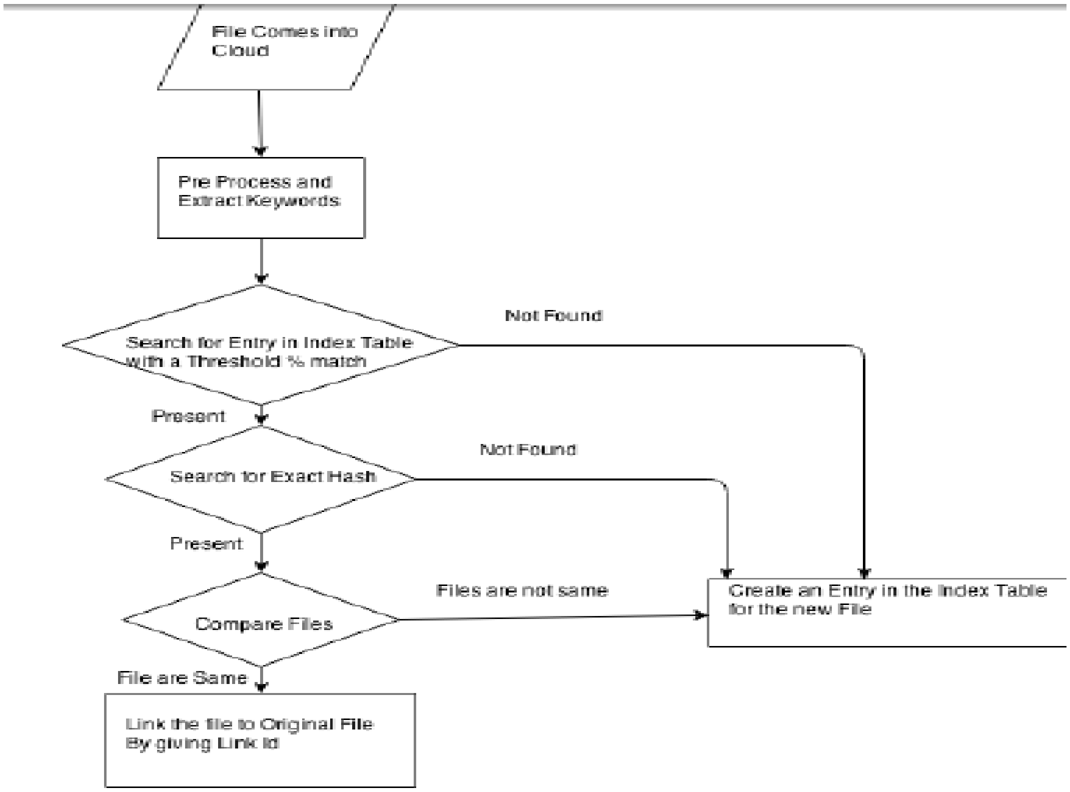
File Comes into Cloud
Pre Process and Extract Keywords
Search for Entry in Index Table with a Threshold % match
Not Found
Present
Search for Exact Hash
Not Found
Present
Compare Files
Files are not same
Create an Entry in the Index Table for the new File
File are Same
Link the file to Original File By giving Link Id

Figure 1: Flow Chart for Reducing Data Redundancy

The above process is done if the file is new. To determine that , first the keywords of the file are extracted and are searched in the index table. A threshold percentage match is fixed. With the files returned after the threshold search, the hash values of the files are compared. If a match is found, then the files are compared word by word. If the file is a perfect match then the file is linked to the original file by giving a link Id and maintaining a table with link id and the file name and the paths as shown in Table2. From now on, only one file will be used for clustering and other files will be automatically added to the same cluster without adding them as points to the clustering algorithm.

Link Id	File Names	File Paths

B. Clustering of data

As discussed in section II, Static clustering algorithms cannot be applied to data on cloud due to the rapid incoming data. Therefore dynamic clustering methods have to be used to cluster the data. Dynamic Clustering algorithms cluster the data instantaneously.

There are many dynamic clustering algorithms existing and have been used for various applications. We look at the two algorithms that can be used in the context.

- **Steaming K-Means**

The Steaming K-Means algorithm works on principle of Divide and conquer. It is a batch clustering algorithm and therefore data stream is divided into blocks and clustered, the results of clustering of each block are finally combined.

- **D Stream Algorithm**

D stream analyzes the cluster based on the density of the data. D stream algorithm will assume each input data record is in a separate space and each space is divided into partitions. A density grid maintain the Density of the data record. Based on the density of the data record in the density grid, a coefficient is assigned for each data. These coefficients will be the timestamp of occurring of data record in the data stream for each data record. The overall density of the grid can be calculated as the sum of all density coefficients of all the data record that are mapped to density grid. Finally the grid cluster is connected to the group which has got the higher density than the other grids.

4. ADVANTAGES

Mining the data in the cloud storages is very important because most of the data from photos to financial document, everything is being stored in the cloud. When the data in the cloud's are clustered, they are much more organized to do statistical analysis. For example, this may include findings the consumption patterns of consumers which will help companies in the development of new products and services.

Clustering of data also helps in organization of the vast data on the cloud. This helps in reducing the search time for searching files as files are searched only in specified clusters.

5. CONCLUSION AND FUTURE WORK

Use of stream clustering algorithm for clustering the data in the cloud storage is advantageous as compared to the static clustering algorithm as the stream clustering algorithm are light- weight algorithms and can therefore operate on huge quantities of data. The static clustering algorithm such as K-Means have been used for forensic analysis. As a future work, the clustering algorithms can be tested for different applications such as Forensics. Digital Forensics has been a fast growing application in the age of Cyber Crime. Attacks on system are being carried out and the data for these attacks are being shared in clouds. With a light- weight algorithm and more computing power, it will be possible to cluster real time data and prevent attacks.

REFERENCES

[1] Louis colubus " Predicting Enterprises Cloud Computing Growth" . Accessed on 24-12-2013 at http://www.forbes.com/sites/louiscolumbus/2013/09/04/predicting-enterprise-cloud-computing-growth/

[2] Gerald Salton and Christoper Buckley, "Term-Weighing Approaches in Automatic Text Retrieval". Information processing and Management Vol.24, No.5 , pp513-523, 1988

[3] Yogita, Durga Toshniwal "Clustering Techniques for straming Data- A Survey" 2013 3rd IEEE International Advance Computing Coference(IACC)

[4] Ahamad Shafeeq B M and Hareesha K S "Dynamic Clustering of Data with modified K-Means Algorithm" 2012 International conference on Information and Computer Networks(ICICN 2012)

[5] Dong-Moon-Kim, Kun-su Kim, Kyo-Hyum-Park,Jee-Hyon-Lee, Keon Myung Lee" A Music Recommendation System with a Dynamic K-Means Clustering Algorithm" IEEE Sixth International Conference on Machine Learning and applications2007.

[6] Hongyang Liu, Jia He "The Application of Dynamic K-Means Clustering Algorithm in the Center Selection of RBF Neural Networks" 2009 IEEE Third International conference on Genetic and Evolutionary Computing.

[7] Jonathan de Andrade Silva, Edurado Raul Hruschka " Extending K-Means-Based Algorithmsfor evolving Data Streams with Variable Number of clusters" 2011 IEEE 10th International Conference on Machine Learning and Applications.

[8] Nir Ailon, Ragesh Jaiswal, Claire Montelenoi" Streaming K-Means approximation" accessed on 24-12-2013 at http://machinelearning.wustl.edu/mlpaper/paperfiles/NIPS2009_1085.pdf

[9] Yixin Chen. Li Tu " Density Based Clustering for Real-Time Stream Data" IEEE

[10] Luis Filipe Da Cruz Nassif and Enduardo Raul Hrushchka " Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection" IEEE transaction in Information Forensics and Security Vol 8 No. 1, January 2013