ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Data Mining Techniques for Cloud Technology

Kareemunnisa[1], D Swetha[2], P S Silpa[3], D Anuradha[4], Ravindra Changala[5]

Assistant Professor, Dept of IT, Guru Nanak Institutions Technical Campus[1,2,5]

Assistant Professor, Dept of CSE, Keshav Memorial Institute of Technology, Hyderabad[3]

Assistant Professor, Dept of CSE, CMR Technical Campus, Hyderabad[4]

Abstract: Over the past few years, Cloud Computing retrieves required data from the huge database with lesser amount of time becomes tedious. Hence without having a proper cache management framework the retrieval of required data from the cloud data servers becomes difficult. Here this paper suggest a probabilistic framework for efficient retrieval of data from huge database using combined approach of caching and frequent transaction analysis. A novel method of cache management is used to serve the cloud user effectively. A Derived Cluster which is created in online mode, acts as cache which holds the frequent transactions for downloading required data quickly. Frequent transactions are identified based on the run time statistics of the transaction. We have implemented the model in a cloud environment and the experimental results show that our approach is more efficient than the existing search technology and increases throughput by handling more number of transactions efficiently with reduced latency.

Keywords*:* Cloud Computing, Data mining, frequent Item sets, cloud storage, and clustering.

1. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Mining Association rule is a way to find interesting associations among large sets of data items. Using this we have determined the frequent item sets based on a predefined support [1].

During the last one decade, awful development and maturation have happened in the cloud computing system. As a key component, the large-scale cloud computing system has been attracting a great deal of attention from both industry and academia. The Cloud performance was improved by an improved scalability, availability and security. In order to improve the scalability and availability of the cloud computing system, metadata management plays a vital role [2]. Setting up of an external metadata server (MDS) for metadata service speeds up the file retrieval by reducing latency. However, due to the rapid growth of data in cloud, the growth of metadata also increases and by this searching of metadata becomes complex.

Although, research in this field of frequent pattern mining is not new, it is still emerging [3]. The reputation of the cloud and large volume of data have caused researchers to use such techniques for the exploration of unknown knowledge in data collections. This perhaps justifies the sudden increase of research in the general topic of mining the data in cloud computing environment. Among various data mining techniques employed, clustering is a popular technique. Although metadata in a cloud environment is employed to reduce the search space, yet the rapid growth of data in cloud causes a fast growth of metadata as well and hence there is a need to transform metadata into information too. In order to address this problem, we implement clustering technique in metadata which can further improve the availability of the metadata file. To achieve this, the entire metadata file is clustered based on the keyword of the original file and is further clustered dynamically. Although standard clustering techniques such as k-means can be applied for clustering the metadata, they usually do not satisfy the special requirements and are not robust enough to handle clustering the metadata of high dimensionality and huge volume of data [4]. For example, in some metadata servers, the cluster size increases from few hundreds to thousands of metadata files. This increase in cluster volume reduces the clustering efficiency for some of the state-of-the art algorithms. In this paper, we propose a novel approach called Mining Maximum Frequent Transaction set using bloom filter (MFT), for metadata clustering based on the idea of frequent transactions. The main idea of our clustering technique is that there exists some frequent transactions for each cluster based on the user and the file transaction and different clusters share few frequent transactions. A frequent transaction is a set of transactions that occur together which involves any number of cluster and also from various metadata servers. We have conducted the experiments on MFT algorithm and compared the traditional algorithms with proposed algorithm. Our results show that the MFT is scalable in terms of number of items and transactions and can also maintain the dynamic workload effectively as they are processed independently.

2. LITERATURE SURVEY

More research works were being pursued in analytics in cloud storage. Yu Hua et al. [5] have projected a scalable and adaptive metadata management in ultra lage scale file systems. Abhishek Verma et al. [6] have suggested metadata using Ring file system. In this scheme metadata for a file is stored based on hashing its parent location. Replicas are stored in its successor metadata server. Ashok

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Savasere et al. [7] proposed the partition algorithm, and that aims to further improve the efficiency, by reducing the number of database scans, however, considerable time is still wasted in scanning infrequent candidate item sets. Wuling Ren et al.[8] has proposed a fast Mining maximum frequent item using Interest Frequent Pattern Matrix (IFPM), which is based on user's interests, filters the transaction database according to the level of data item. The Bloom filter is a space-efficient probabilistic data structure that supports set membership queries. Chaohui Liu et al. [9] has proposed a technique of mining frequent item sets using Matrix reduces scanning cost and execution times, but the algorithm works only for few transactions. R. Anitha, Saswati Mukherjee [10] the authors has claimed that the data retrieval using metadata is less when compared to without using metadata due to reduction in latency.

3. PROPOSED SYSTEM MODEL

The architecture diagram of the proposed system model is shown in Figure 1. Each block in the architecture explains about the dynamic creation of derived cluster based on frequent transaction-set of metadata from user perspective and also from the temporal perspective.

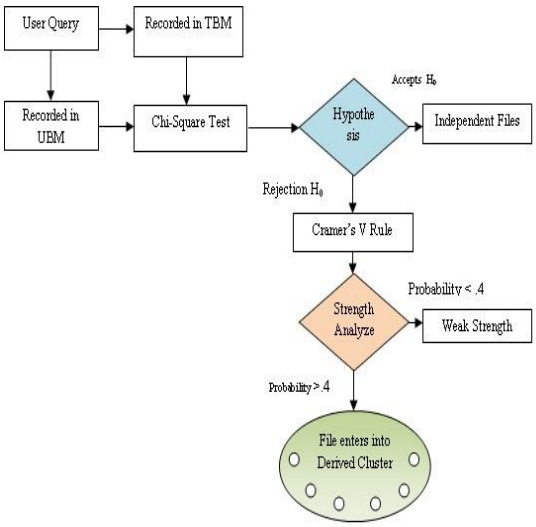

Fig.1:- Architecture Diagram

Clustering is a process of grouping a collection of data into different category groups so that documents in the same category group describe the same feature. The proposed approach consists of the following major steps: Recording the transactions, Construction of Derived cluster using statistical analysis, mining of frequent transaction-set using Bloomier Matrix Filter. When a user transaction happens the corresponding transaction is recorded in the hash based bloomier matrix filter.

In order to make the retrieving process further efficient, a dynamic cluster is formed based on the history of transactions from the user perspective and time based transaction. The transactions are recorded and are used in the Bloom filter and analyzed. The Bloomier matrix filter is constructed based on the transactions recorded. The transactions with respect to time slot are recorded in Time Based Transactions BMF (TBT). The transactions with respect to user are recorded in User Based Transactions BMF (UBT) The analysis is done based on the similarity measure between the transactions and by using this run time statistical analysis a new cluster is formed. Thus knowledge has been extracted based on run time analysis which further improves the efficiency of the proposed model.

4. BLOOMIER MATRIX FILTER

All paragraphs a special data structure called Bloomier Matrix Filter derived from a basic bloom filter, extracts the history of transaction based on temporal behavior and the user behavior. The structure of bloomier matrix filter is as shown in figure 2.

	f1	***f2***	***f3***		***fn***
x1	0	0	0		0
X2	0	0	0		0
X3	0	0	0	…	0
….					
X_n	***0***	***0***	***0***		***0***

Fig.2 Structure of Bloomier Matrix Filter

Bloomier matrix filter represents and work on transaction set efficiently as it is a probabilistic data structure which has independent look up. The method works efficient for large databases and leverages statistical analysis on the recorded transactions and thus establishes accurate links between the records. Work is to develop matrix based manipulations in order to estimate similarity between user centric transactions and time-slot centric transactions. The structure takes up a matrix format where the column represents the files and row represents the time or user. Initially all the values are set to zero and whenever transaction happens, the counter value of corresponding rows and columns gets incremented. The name bloomier matrix filter, because it filters out the unwanted files occupying the search space, thus reduces the database scanning in large databases. In spite of scanning the database every time the usage of bloomier matrix minimizes the overhead by means of look up confidence values and compare them to the other set of recorded values based on transactions recorded with respect to a specific query. The Bloomier Matrix manipulations were originally developed to go for fast join operations in large distributed database. The similarity between the values are carried out and based on the similarity values, the files are stored in the derived cluster.

The Bloomier Matrix filter can be viewed as a matrix, which consists of parallel sub matrices in order to represent transactions with respect to files. A sub matrix is composed of parallel arrays and can be used to represent transactions. An array consists of counters and its related transaction frequency. The array values are processed parallel to calculate the frequent item set. Assume that the query is with respect to i^{th} file, the frequent transaction file

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

with respect to the respective file is found. The process uses various time slots ranging from Ti , where i=1 to n. Thus, each sub matrix are used for processing frequent files. By using BMF the values are processed parellely and the results are used for further dependency tests. Figure 2 presents the algorithm of adding values in the proposed structure.

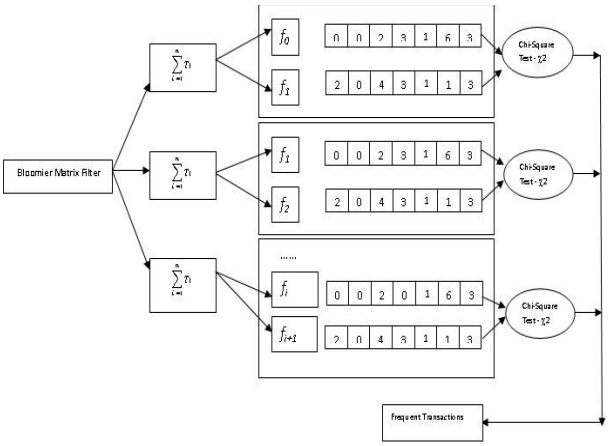

Fig 3. Structure of proposed Bloomier Matrix Filter

5. MINING OF MAXIMUM FREQUENT TRANSACTION-SET ALGORITHM

The MFT algorithm is used to effectively discover the frequent transaction set based on the concept of support approximation. Two important measures, the chi-square measure and cramer's V rule are adopted in MFT to approximate the support of item sets. Chi-square analyses the support of the transactions which results in the dependency of the transactions. Cramer's V Rule is used to determine the strength of association. The files with higher strength are physically copied at the derived cluster. The approximation quality, the quality of clustering obtained depends on the number of levels required to create the cluster. From this perspective it is required using a space efficient algorithm in order to create a cluster with approximate quality. The space efficient data structure results in handling large number of transactions. In cloud scenario as the growth of file reaches extra byte's, increases the count of transaction. Hence to handle this scenario perfectly the algorithm uses a space efficient data structure called bloomier matrix filter. The bloomier matrix has an advantage of independent look up hence reduces the time complexity for lookup. The advantage of the algorithm is that it maintains categorical data so that the space used is less. The algorithm uses two step processes to deduce the files, which fits into the derived cluster. Algorithm: Creation of Derived Cluster using Maximum Frequent Transaction-set MFT algorithm.

Input: Record of time centric transactions, user centric transactions

Output: Building a derived cluster

Step.1 Create Bloomier Matrix1-TBM

Scan the DB based on time slot

for (all transactions in time slot interval){

do{

if (F_{ID} exists in Transaction = 1) then

F_{ID}= F_{ID}++; /* Transaction taken place for same F_{ID} */

Else

F_{ID} th Column=1

} until transaction = END }

Step 2 : Create Bloomier Matrix 2-UBM

Scan the DB based on User's interest

for (all transactions on users interest)

do {

if (F_{ID} exists in Transaction = 1) then

F_{ID}= F_{ID}++; /* Transaction taken place for same F_{ID} */

else

F_{ID} th Column=1

} Until transaction = END }

Step 3: Analysis of Dependence for TBM and UBM for (i=0 ; i< Tn; i++)

for (j=0 ; j α(r-1)(c-1))

for (k=0 ; k α(r-1)(c-1))

{

Chi-squared $\chi 2 = \sum((\text{observed-expected})^2 / \text{expected})$

/* determines whether there exists an association between two files*/

if ($\chi 2 >$ α(r-1)(c-1)) then

{

Call Step 4

}

}

End for

End for

Step 4: Analysis for Strength

Read the output from step 3

for (j=0 ; j<F_j;j++) /*Apply Carmer's V Rule*/

for(k=0;k< F_{j+1};k++)

{

Carmer's V Rule /*varies between 0 to 1*/

If(V>.4) then

{

Enter into Derived Cluster

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

}

}

}

}

/*Close to 0 it shows little association between variables. Close to 1, it indicates a strong association.*/ Maximum value close to 1 and values greater than 0.4 enters derived cluster.

6. CONCLUSION

Due to the exponential increase in the volume of data stored at cloud data servers, the need for analyzing the metadata created is required. This paper has proposed an effective approach for clustering metadata in accordance with the frequent item sets that provides significant search space reduction. We have obtained a set of Base clusters using the attribute keyword which is extracted using TFIDF algorithm.

The Derived cluster is a dynamic cluster which is generated at the time of information retrieval using frequent item sets based on User Transaction and Time based Transaction. Our MFT method allows the user to retrieve the data with less latency and increase the throughput.

REFERENCES

[1] Brin, S., Motwani, R. and Silverstein, C. "Beyond market basket: Generalizing association rules to correlations", in the Proceedings of the 1997 ACM-SIGMOD International Conference on the Management of Data (SIGMOD'97), Tucson, AZ, 1997, pp. 265–276.

[2] Michael Cammert, Jurgen Kramer, and Bernhard Seeger, "Dynamic Metadata Management for Scalable Stream Processing Systems", Proc. of IEEE International Conference on Data Engineering Workshop, pp.644-653, 2007.

[3] NVB Gangadhara Rao, Sirisha Aguru, ‖A Hash based Mining Algorithm for Maximal Frequent Item Sets using Double Hashing‖ Journal of Advances in Computational Research, Vol. 1 No.1-2, 2012, pp.1-6.

[4] Tan Zhipeng, Zhou Wei, Sun Jianliang, Zhan Tian, and Cui Jie, ―An Improvement of Static Subtree Partitioning in Metadata Server Cluster,‖ International Journal of Distributed Sensor Networks, pp. 1-10, 2012.

[5] Yu Hua, Yifeng, Hong Jiang, Dan Feng, and Lei Tian,"Supporting Scalable and Metadata Management in Ultra Large Scale File Systems", IEEE Transactions on Parellel and Distributed Systems, Vol.22, No.4, 2011.

[6] Abhishek Verma, Shivaram Venkataraman, Matthew Caesar, and Roy Campbell, "Efficient Metadata Management for Cloud Computing Applications", Proc of International Conference on Communication Software and Networks, 2010.

[7] Ashok Savasere, Edward Omiecinski, and Shamkanth Navathe, ―An Efficient Algorithm for Mining Association Rules in Large Databases‖ In VLDB 1994, Zurich, Switzerland, pp.432-443.

[8] Wuling Ren, Guoxin Jiang, ― A Fast Algorithm for Maximum Frequent Itemsets Based on the User' Interest,‖ International Colloquium on Computing, Communication, Control, and Management, 2009.

[9] Chaohui Liu, Jiancheng an, The Software Engineering School, China ―Fast Mining and Updating Frequent Itemsets‖, 2008 ISECS International Colloquium on Computing, Communication, Control and Management, Vol.1, pp. 365-368.

[10] R. Anitha, Saswati Mukherjee, "A Dynamic Metadata Model in Cloud Computing", Proc. of Springer CCIS, Vol.2, pp.13–21, 2012.

BIOGRAPHIES

Ms. Kareemunnisa received M.Tech from JNTU Kakinada. Her research interest includes Wireless Sensor Networks and Data Mining. Presently she is working as Assistant Professor in IT Dept, Guru Nanak Institutions Technical Campus, Hyderabad.

Ms. D Swetha received M.Tech from JNTU Hyderabad. Her research interest includes Data Mining. Presently she is working as Assistant Professor in IT Dept, Guru Nanak Institutions Technical Campus, Hyderabad.

Ms. P S Silpa received M.Tech from JNTU Hyderabad. Her research interest includes Data Mining and Cloud Computing. Presently she is working as Assistant Professor in CSE Dept, Krishana Murthy Institute of Technology, Hyderabad.

Ms. D Anuradha received M.Tech from JNTU Hyderabad. Her research interest includes Data Mining and Computer Networks. Presently she is working as Assistant Professor in CSE Dept, Dept of CSE, CMR Technical Campus, Hyderabad.

Mr. Ravindra Changala received M.Tech from JNTU Hyderabad. Her research interest includes Data Mining. Presently she is working as Assistant Professor in IT Dept, Guru Nanak Institutions Technical Campus, Hyderabad.