ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Content Clustering Analysis Using Modern Map Reduce Model

Saranya.P

M.E Student, Dept. of C.S.E., Arunai Engineering College, Thiruvannamalai, India

ABSTRACT: In order to resolve the scalability and load balancing challenges in the existing parallel mining algorithms for common itemsets. We can apply the map reduce programming model, by using this model we can achieve reduction storage and avoid conditional pattern bases. It can be executed with Hadoop cluster which it is aware to data distribution and dimension since item sets with different lengths. We need to look up the performance so we built up the work load balance across the cluster node. It will integrate Hadoop with data placement mechanism on heterogeneous cluster. So data placement scheme is used to balance the amount of data stored in each heterogeneous cluster node in order to improve data processing performance. The energy saved and thermal management will be of future generation to ensure our life. So this project will propose various approaches to improving energy efficiency of Hadoop running on clusters.

KEYWORDS:Hadoop, Map Reduce, Frequent item sets, Parallel mining

I. INTRODUCTION

Text mining is a growing new field that tries to collect important information from natural language text. It may be slackly described as the text used to examine and to remove information that it is used for a particular reason. It can be evaluated with the type of data stored in the file is unstructured form, but it's hard to deal with algorithmically. But in current society, the text is mainly the general vehicle for the proper replacing of information. The field of text mining generally deals with texts whose purpose is the communication of exact information or view and the quick for trying to taken out information from the text. It is generally used to denote any system that analyzes great amount of natural language text and spot the lexical and linguistic procedure patterns in order to extract most useful information.

Frequent itemsets mining is a main problem in association rule mining and sequence mining. Speeding up the growth of FIM is sensitive and significant because FIM utilizations are important for the part of mining time, which it is suitable for its high calculation of input and output strength.. It can be run on a single machine that can be affected from performance failure. To report this difficulty, we check how to take out FIM uses MapReduce model. Frequent itemsets mining can be classified into two types namely:

1. Apriori
2. Frequent Pattern growth

Apriori algorithm: It is an algorithm for frequent itemsets mining and association rule over transactional databases. While considering the large data sets, apriori attempts to fail the reliability and efficiency. It aggressively prunes the set of potential candidates of amount m by using the following observation: a candidate of amount m can be frequent only if all of its subsets also meet the least amount of threshold support. Even with the pruning, the task of discovering all association rules needs a lot of computation of power and memory.

Frequent Pattern growth algorithm: It uses the frequent items ultrametric tree in the design of our parallel frequent itemsets mining technique. While using a FP growth algorithm we are achieving 4 things they are:

1. Falling input and output overhead
2. Dense storage
3. Good way of partitioning dataset
4. Avoiding recursively traverse

Most prominently, in the existing parallel mining algorithms has some problem. In order to solve this problem, we implement MapReduce on parallel computing. It offers a potential solution to the calculation needed for this task, if the efficient and scalable algorithms can be designed.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

II. RELATED WORK

Map reduce is a programming model which it could be designed for large scale processing. While using this model we save time and energy for large database.

Sun, Dawei , Lee, Frada , Haghighi, Pari [1]: In this paper used to take the large database but they are using the algorithms with small data size Vertical apriori MapReduce algorithm is used only for analysing the large database.

S. Hong, Z. Huaxuan, C. Shiping, and H. Chunyan, [2]:In this paper uses the frequent pattern growth algorithm ,it reduces the performance and leads to poor performance. So we propose a improved frequent pattern growth algorithm.

J. Choi, C. Choi, K. Yim, J. Kim, and P. Kim, [3]:In this method uses the cloud computing technology. It is able to manage the large amount of data and to increases their efficency and performance.

ZhuoboRong,Chongqing,Dawen Xia,Zili Zhang,[3]: In this paper used apriori and FP growth algorithm, so it take more time to do and it need high memory .So we implement MapReduce technique to mine large datasets.

Jiawel Han,Jian Pei,Yiwen Yin,Runying Mao[4]:In this paper proposed an algorithm called frequent pattern growth which it is achieving a compressed storage. It takes more cost and time.

III. PROPOSED WORK

A. Load Balance:

The *decompose* task of the MapReduce work accomplishes the disintegration process. If the extent of an item set is *m*, the point convolution of the item-set is $O(2^m)$. In another way, when the item sets time taken is not going to higher, so the decomposition transparency will be increases for the data sets In which the datasets to be presented on Hadoop. The action towards comparison of load between data nodes of a Hadoop group is to quantitatively calculate the total workout load of limited item sets. We need to compute load balance between the data nodes.

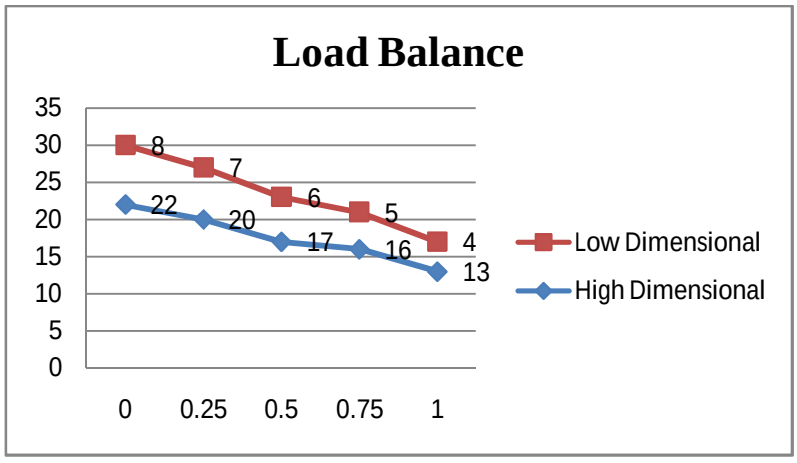

Fig1. Load Balance

B. Speedup:

It is used to evaluate the performance by achieve two things they are:

1. Decreases the quantity of itemsets which it is created by every node.
2. Improving communications between mappers and reducers functions.

In this method attain overall computing capacity by improving the number of nodes and communication among data nodes.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

C. Scalability:

It is the ability of the system to measure the amount of work. It is generally difficult to determine the particular needs of requirement. It is used as the most important issues in database, routers, networking, etc. A system whose performance has been improved by adding hardware. It is said to be scalable system.
Example: When the data size is large, so it more cost to read and write the file in HDFS.

IV. SYSTEM ARCHITECTURE

It used to identify the text document for investigation and then it using some statistical technique for identify the meaning of the sentence. In that sentence it used to extract some interesting pattern and concepts by using MapReduce model for content categorization and then it produced the meaningful outcome to the end users.
Text document: It contains only the text such as program code and batch files. Text based document are in readable form.
Semantic analysis: It is the structure and meaning of speech. It is used to discover grammatical rules, the meanings and to uncover specific meanings to words in foreign languages. The analyst compares the grammatical structure and meanings of different words to those who used in the native language. As the analyst discovers the differences in the sentence, it can help to understand the unfamiliar grammatical structure.
Extract Patterns and concepts: In that semantic analysis, it is used extract some meaningful sentence and interesting patterns for analysis of large text.
Content categorisation: Content may be classified according to the subjects or attributes such as document type, author, printing year etc. In the another article is considered only subject classification. There are two main categorisation of t classification of documents: the document based approach and the application based approach.
Content reduction: It is used to remove the common words and to preserve the keywords. It can classified into two ways : one is automatic text reduction and another one is automatically reduce the content or content summarisation.
Performance metric evaluation: It is used to measure the activities and performance of organisation which is used to identify the specific and quantifiable output.
Recommendations to end users: It is used to ensure the high quality of proces which it satisfy the needs of customer.

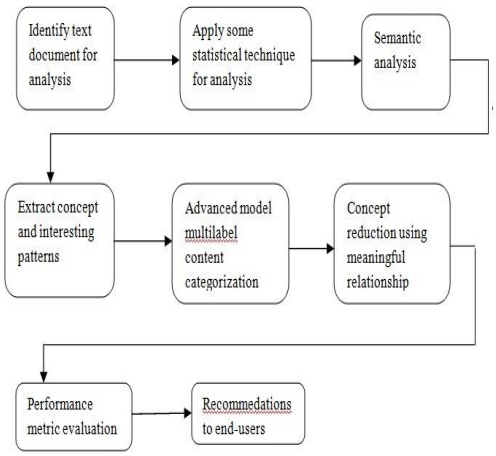

Fig 2. System Architecture

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

V.OVERVIEW OF MAP REDUCE

MapReduce is a programming model and it processed for generating large datasets on a cluster. It can be composed of two task :

- Map task
- Reduce task

Map task :

It takes a set of data and converted it into another set of data. Single elements are broken down into number of tuples. Each actor node uses the "map()" function to the local data, and writes the result to a temporary storage. *Shuffle step:* Actor nodes to understand the information which is based on the outcome keys are being obtain by the "map()" function, such that all data being located to one key is placed on the same actor node.

Reduce task : It is the combination of both shuffle and reduce step. It takes the output from map as input and combine those data tuples into smaller set of tuples. After processing it produce a new set of output whic will be stored in HDFS**.**

Combination of MapReduce:

MapReduce job splits the input data into independent pieces which are processed by the map task is a completely parallel manner. The framework which sorts the output of maps which are then input to the reduce tasks. Both input and output are stored in HDFS file system.

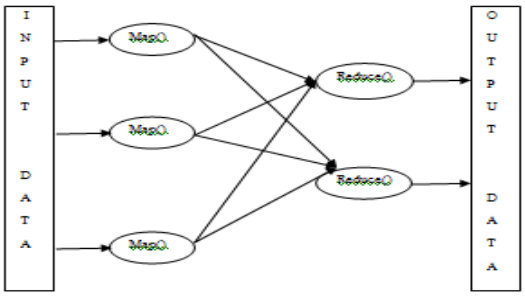

Fig3.Map Reduce

Advantages of MapReduce:

1. It is easy to scale data processing over multiple computing cluster.
2. It is designed to run on cluster of commodity hardware.
3. It improves the performance and efficency on the large cluster.

IV. EXPERIMENTAL RESULTS

To evaluate the recital of hadoop cluster equipped with 16 data nodes. Each node has an intel Pentium processor, 512MB main memory, and runs on the ubuntu operating system,on which java jdk1.6.0_37 and hadoop are installed. Here all the data nodes in the cluster have gigabit Ethernet network interface cards connected to gigabit ports on the switch. The nodes can communicate with one another using the secure shell protocol. It uses the default hadoop parameter to set the number of map and reduce tasks. It need to generate a set of synthetic datasets using the market-basket synthetic data generator which can be configured to create a wide range of datasets to meet the needs of diverse necessities.

Celestial and Sythetic Dataset are used to progress the Performance of the Proposed system.

1) *Synthetic Dataset:*

Using the IBM pursuit market-basket synthetic data initiator, which can be configured to create a wide range of datasets to meet the needs of various test requirements . We generated a series of synthetic datasets (i.e., the D1000W datasets) The number of items in each D1000W dataset is set to 1000 . while conducting the experiments, we vary the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

average transaction size and the number of transactions in the D1000W datasets. In the empirical study, we make use of the D1000W datasets to investigate the impacts of dimensions and the data size on the performance of the tested algorithms.

1) *Celestial Spectral Dataset:*

We apply to implement a parallel data mining application for celestial spectral data. We use the real-world celestial spectral dataset to evaluate speedup, load-balancing performance, as well as the impact of minimum support. The celestial spectral dataset used in our experiments has 6000 000 transactions and 54 dimensions. if not specified the min support is set as 0.0001 and the number of node is on five nodes. 10- and 60-dimension in the D1000W datasets are respective the representatives of low- and high-dimensional synthetic datasets in the following set of experiments.

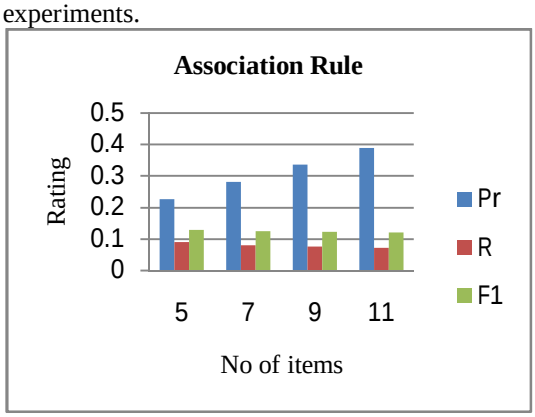

Fig.4 Association Rule

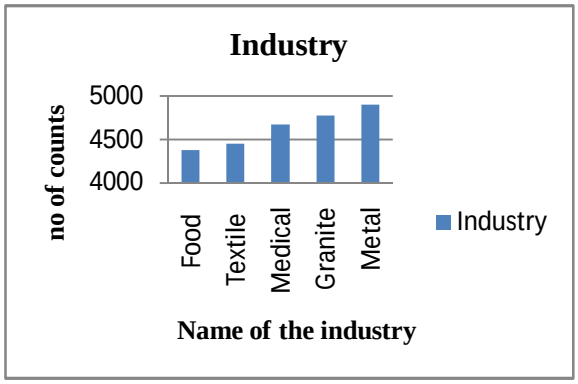

Fig.5.Result of map reduce

V. CONCLUSION AND FUTURE WORK

In the accessible system, the datasets undergo frequent item sets mining with Apriority Algorithm. But considering the large datasets, apriority attempts to fail the reliability and efficiency. To overcome this we go for Mapreduce technique is used for large cluster of data and to save the time ,energy and also to solve large computational problems. It improves the performance of workload balance across the large cluster node and also to improve the efficiency of data items which is it running on hadoop cluster. As a future study.we will apply this metric to examine sophisticated load balance approach in the framework of FiDoop. We plan to apply a data conscious load evaluation process to significantly get improved the load-balancing presentation. Our data assignment design is causal to evaluation the amount of data accumulate in all various node to attain better data processing performance. It fairly accurate to production concern power savings and thermal organization will be of our future investigate. We will suggest various steps to enhanced energy effectiveness which it running on Hadoop clusters.

REFERENCES

1. D. W. Cheung, S. D. Lee, and Y. Xiao, “Effect of data skewness and workload balance in parallel data mining,” IEEE Trans. Knowl. Data Eng., vol. 14, no. 3, pp. 498–514, May/Jun. 2002.
2. H. Li, Y. Wang, D. Zhang, M. Zhang, and E. Y. Chang, “PFP: Parallel FP-growth for query recommendation,” in Proc. ACM Conf. Recommend. Syst., Lausanne, Switzerland, 2008, pp. 107–114.
3. L. Cristofor. (2001). Artool Project[J]. [Online]. Available: http://www.cs.umb.edu/laur/ARtool/, accessed Oct. 19, 2012.
4. J. S. Park, M.-S. Chen, and P. S. Yu, “Using a hash-based method with transaction trimming for mining association rules,” IEEE Trans. Knowl. Data Eng., vol. 9, no. 5, pp. 813–825, Sep./Oct. 1997.
5. J. D. Holt and S. M. Chung, “Mining association rules using inverted hashing and pruning,” Inf. Process. Lett., vol. 83, no. 4, pp. 211–220, 2002.
6. F. Berzal, J.-C. Cubero, N. Marín, and J.-M. Serrano, “TBAR: An effi-cient method for association rule mining in relational databases,” Data Knowl. Eng., vol. 37, no. 1, pp. 47–64, 2001.
7. J. Zhang, X. Zhao, S. Zhang, S. Yin, and X. Qin, “Interrelation anal-ysis of celestial spectra data using constrained frequent pattern trees,” Knowl.-Based Syst., vol. 41, pp. 77–88, Mar. 2013.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

8. D. W. Cheung and Y. Xiao, "Effect of data skewness in parallel mining of association rules," in Research and Development in Knowledge Discovery and Data Mining. Berlin, Germany: Springer, 1998, pp. 48–60.
9. T. Shintani and M. Kitsuregawa, "Hash based parallel algorithms for mining association rules," in Proc. 4th Int. Conf. Parallel Distrib. Inf. Syst., Miami Beach, FL, USA, 1996, pp. 19–30.
10. E.-H. Han, G. Karypis, and V. Kumar, "Scalable parallel data mining for association rules," IEEE Trans. Knowl. Data Eng., vol. 12, no. 3, pp. 337–352, May/Jun. 2000.
11. I. Pramudiono and M. Kitsuregawa, "Parallel FP-growth on PC cluster," in Advances in Knowledge Discovery and Data Mining. Berlin, Germany: Springer, 2003, pp. 467–473.
12. P. Tang and M. P. Turkia, "Parallelizing frequent itemset mining with FP-trees," in Proc. 21st Int. Conf. Comput. Appl., Seattle, WA, USA, 2006, pp. 30–35.
13. K. Yu and J. Zhou, "Parallel TID-based frequent pattern mining algo-rithm on a PC cluster and grid computing system," Expert Syst. Appl., vol. 37, no. 3, pp. 2486–2494, 2010.
14. S. Cong, J. Han, J. Hoeflinger, and D. Padua, "A sampling-based framework for parallel data mining," in Proc. 10th ACM SIGPLAN Symp. Prin. Pract. Parallel Program., Chicago, IL, USA, 2005,255–265.
15. K.-M. Yu, J. Zhou, T.-P. Hong, and J.-L. Zhou, "A load-balanced dis-tributed parallel mining algorithm," Expert Syst. Appl., vol. 37, no. 3,2459–2464, 2010.
16. L. Zhou et al., "Balanced parallel FP-growth with MapReduce," in Proc. IEEE Youth Conf. Inf. Comput. Telecommun. (YC-ICT), Beijing, China, 2010, pp. 243–246.