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. 3, Issue 7, July 2015

An Adaptive Approach for Association Rule Mining in Cloud Computing Using Hadoop Technology

D.KeranaHanirex, Dr.K.P.Kaliyamurthie, Sundararajan.M, Arulselvi S

Assistant Professor, Dept. of CSE, Bharath University, Chennai, Tamil Nadu, India

HOD, Dept. of CSE, Bharath University, Chennai, Tamil Nadu, India

Director, Research Center for Computing and Communication, Bharath University, Chennai, Tamil Nadu, India

Co-Director, Research Center for Computing and Communication, Bharath University, Tamil Nadu, India

ABSTRACT: An association rule mining is one of the research data mining technique.Association rule mining in cloud computing is one of the an emerging area of research.This paper proposes the recent technology Hadoop that are used for mining association rules.The association rules are developed on the basis of the frequent item set generated from the data items. The frequent item sets which are generated from traditional algorithm such as Apriori,FP_Growth algorithm requires lots of space and memory. The recent technology Hadoop is used to provide parallel, scalable, robust framework in the distributed environment.

I. INTRODUCTION

Data mining is the process of analyzing data and getting useful information from the database. The data mining tasks are prediction ,classification ,association rule finding correlations, patterns from the data set. The motivation for searching association rules is to analyze large amounts of super market basket data. Association rule specifies how often items are purchased together.The discovery of association rules can be divided into 2 phases: First discover all frequent itemsets and then association rules using the frequent itemsets.There are 2 interesting measures in association rule mining support and confidence.Support determines how often the rules occur in the database.Support of an association rule X =>Y is the ratio of the number of occurrences of {x,y} to the total number of transaction of D. Confidence measure of an association rule X =>Y is the ratio of the total occurrences for item X and Y to the total number of occurance for item X.In our earlier research work various association mining algorithms[10,11,12,13,14] are implemented for different datasets.

Association rules are widely used in various areas such as telecommunication networks, marketing and inventory control etc. Association rules can also be mined for the field of bioinformatics, medical diagnosis, web mining and scientific data analysis. The traditional algorithms may generate an extremely large number of association rules in many cases and sometimes the association rules are very large. It becomes almost impossible for the users to validate such large number of complex association rules.The rules are generated by applying the criteria such as confidence, coverage, leverage, lift or strength.

II. RELATED WORK

The parallel association rule mining algorithm is proposed in paper[2] .Here either the data are the task is divided among the nodes.It uses the concept of parallelism since the data input size high and distributed in nature clouds can be used for generating rules.In association rule mining the frequent itemsets can be mined using hashing techniques which further increases the efficiency[3,4,5].The efficiency of finding the frequent itemsets can be further increased by applying distributed

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. 3, Issue 7, July 2015

technique[8,9].Mining the maximal frequent itemsets for finding the frequent itemsets as well as mining the frequent pattern with out the generation of candidate itemsets also proposed in paper[6,7].We need some techniques to improve the efficiency of finding the frequent itemsets.

Cloud computing is one of the emerging technique which allows consumers and businesses to use applications without any installation and access their files at any computer with internet access. We can do efficiently by centralizing storage, memory, processing and bandwidth. Since we have to pay only for service and its usage it is a reliable and inexpensive option for association rule mining.

Hadoop is a technology which provide scalability and reliability option for to a distributed system. Hadoop provides tools for analyzing both structured and unstructured data. MapReduce and HDFS of Hadoop use simple, robust techniques on inexpensive computer systems to deliver very high data

availability and to analyze enormous amounts of information quickly[3]. The reliable part of mining is taken care by the Hadoop framework. Thus the association rule mining is done on Hadoop and cloud so that outcome is a reliable and efficient. Mining association rules may require iterative scanning of large transaction or relational distributed databases, which is quite costly in processing. It aims to show that Hadoop along with cloud computing can be considered as an option for association rule mining. [3]As with increase of data set and increase of items in the candidate set all the data cannot be kept and managed on a single computer. It also shows that the scalability of Hadoop system with respect to increase in number of candidate sets and input data.

III. ASSOCIATION RULE MINING

Apriori algorithm is the traditional algorithm to find the frequent itemsets.
Input: D, (Data set),
min_sup (minimum support threshold),
Output: L, (frequent itemset)
Method:
(1) L1 =find_frequt_1-itemset (D);
(2) for (k=2; Lk-1 ≠ φ ; k++) {
(3) Ck = Apriori_gen(Lk-1,min_sup);
(4) for each transaction t € D {
(5) Ct = subset (Ck, t); t
(6) for each candidate c€ Ct
(7) c.count++;
(8) }
(9) LK = {c € Ck | c.count>min_sup}
(10) }
(11) return L= L ∪Lk ;
In the join algorithm Lk-1 is joined with Lk-1.

IV. MAP/REDUCE

A MapReduce technique splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The map, maps the job into key and value. The framework sorts the outputs of the maps, which are then input to the reduce tasks. The input of reduce and output of map must have same type. Typically both the input and the output of the job are stored in a file-system. The output from the 'map' is stored in the temporary file in the HDFS, after completion of the all the map reduce task the file is converted into the permanent one. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

The MapReduce [3] framework operates exclusively on <key, value> pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job. It is not necessary that the output of map and output of reduce both be of the same type. They can be of different data types.Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. Hadoop is designed to run on a large number of machines that don't share any memory or disks. It creates clusters of machines and coordinates work among them. Hadoop consists of two key

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. 3, Issue 7, July 2015

services: reliable data storage using theHadoop Distributed File System (HDFS) and high-performance parallel data processing using a technique called Map/Reduce. It was designed for clusters of commodity, shared-nothing hardware.Even if a machine fails, Hadoop continues to operate the cluster by shifting work to the remaining machines. It automatically creates an additional copy of the data from one of the
replicas it manages.

REFERENCES

[1] R. Agrawal and R. Srikant.Fast algorithms for mining association rules.In Proc. 1994 Int. Conf.Very Large Data Bases, pages 487-499, Santiago, Chile, September 1994.
[2] Sundararajan M., "Optical instrument for correlative analysis of human ECG and breathing signal", International Journal of Biomedical Engineering and Technology, ISSN : 0976 - 2965, 6(4) (2011) pp.350-362.
[3] Ashrafi, M. Z., Taniar, D., & Smith, K. (2004). ODAM: An optimized distributed association rule mining algorithm. Distributed Systems Online, IEEE, 5(3)
[4] Rekha C.V., Aranganna P., Shahed H., "Oral health status of children with autistic disorder in Chennai", European Archives of Paediatric Dentistry, ISSN : 1818-6300, 13(3) (2012) pp.126-131.
[5]PallaviRoy ,"A Thesis on Mining Association Rules in Cloud ",August 2012.
[6] Sharmila D., Muthusamy P., "Removal of heavy metal from industrial effluent using bio adsorbents (Camellia sinensis)", Journal of Chemical and Pharmaceutical Research, ISSN : 0975 – 7384, 5(2) (2013) pp.10-13.
[7] Jong Soo Park, Ming-Syan Chen, and Philip S. Yu. 1995. An effective hash-based algorithm for mining association rules. In Proceedings of the 1995 ACM SIGMOD international conference on management of data (SIGMOD '95)
[8] Kulanthaivel L., Srinivasan P., Shanmugam V., Periyasamy B.M., "Therapeutic efficacy of kaempferol against AFB1 induced experimental hepatocarcinogenesis with reference to lipid peroxidation, antioxidants and biotransformation enzymes", Biomedicine and Preventive Nutrition, ISSN : 2210-5239, 2(4) (2012) pp.252-259.
[9] Ozel, S. A. and Guvenir, H. A. 2001. An algorithm for mining association rules using perfect hashing and database pruning. In 10th Turkish Symposium on Artificial Intelligence and Neural Networks,257--264.
[10] Langeswaran K., Revathy R., Kumar S.G., Vijayaprakash S., Balasubramanian M.P., "Kaempferol ameliorates aflatoxin B1 (AFB1) induced hepatocellular carcinoma through modifying metabolizing enzymes, membrane bound ATPases and mitochondrial TCA cycle enzymes", Asian Pacific Journal of Tropical Biomedicine, ISSN : 2221-1691, 2(S3)(2012) pp.S1653-S1659.
[11] KaramGouda , Mohammed JaveedZaki, Efficiently Mining Maximal Frequent Itemsets, Proceedings of the 2001 IEEE International Conference on Data Mining, p.163-170, November 29-December 02, 2001.
[12] J. Han, J. Pei, and Y. Yin.Mining frequent patterns without candidate generation. In ACM-SIGMOD, Dallas, 2000
[13] D.W. Cheung , et al., "A Fast Distributed Algorithm for Mining Association Rules", Proc.Parallel and Distributed Information Systems, IEEE CS Press, 1996,pp. 31- 42;
[14] Ansari E, Dastghaibifard G, Keshtkaran M, Kaabi H. Distributed frequent itemset mining using trie data structure. IAENG International Journal of Computer Science. 2008;35(3):377-381.
[15]D.KeranaHanirex.,Dr.K.P.Kaliyamurthie. An Adaptive Approach For Mining Frequent Itemsets: A Comparative Study On Dengue Virus Type 1, IEEE International Conference on Human Computer Interaction , (2013)
[16]. D.KeranaHanirex. Dr.K.P.Kaliyamurthie "Mining Frequent Itemsets Using Genetic Algorithm", Middle-East Journal of Scientific Research ,19 (6): 807-810,(2014).
[17]. D.KeranaHanirex.,Dr.K.P.Kaliyamurthie. Finding the Dominating Amino Acids in Dengue Virus (Type-1) Study on mining frequent itemsets ,Int. Journal of Pharama and Bio Sciences, July; 4(3): (B) 880 – 889;(2013)
[18]. D.KeranaHanirex. An Efficient TDTR Algorithm for Mining Frequent Itemsets, International Journal of Electronics and Computer Science Engineering, V2(N1):251-256;(2012).
[19]. D.KeranaHanirex.,Dr.K.P.Kaliyamurthie, Multi-Classification Approach for Detecting Thyroid Attacks,IJPBS, 4(3), (B) 1246 – 1251,July (2013).
[20] Jemima Daniel,Language Teaching in the Digital Age,International Journal of Innovative Research in Science, Engineering and Technology,ISSN: 2319-8753,pp 11029-11031, Vol. 3, Issue 4, April 2014.
[21] Jemima Daniel,Importance of Group Discussions,International Journal of Innovative Research in Science, Engineering and Technology,ISSN: 2319-8753,pp 9081-9084,Vol. 3, Issue 2, February 2014.
[22] Jemima Daniel,The Playboy of the Western World' As a Tragi-Comedy, International Journal of Innovative Research in Science,Engineering and Technology,ISSN: 2319-8753,pp 10379-10381,Vol. 3, Issue 3, March 2014.
[23] Jemima Daniel,Techniques Used in Teaching English, International Journal of Innovative Research in Science, Engineering and Technology,ISSN: 2319-8753,pp 8791-8793,Vol. 3, Issue 1, January 2014.
[24] M. Santhi & Dr. A. Mukunthan,A Detailed Study of Different Stages of Sleep and Its Disorders – Medical Physics,International Journal of Innovative Research in Science, Engineering and Technology ,ISSN: 2319-8753, pg 5205-5212 , Vol. 2, Issue 10, October 2013.
[25] M.NAGESHWARI, Dr.A.MUKUNTHAN , C.RATHIKA THAYA KUMARI,A Study of Surface Ozone Measurement at Vadasery, Kanyakumari District,International Journal of Computer & Organization Trends (IJCOT),ISSN: 2319-8753, pp 160-165,Vol. 1, Issue 2, December 2012.