AJCSE
ASIAN JOURNAL OF COMPUTER SCIENCE ENGINEERING

www.ajcse.info

Asian Journal of Computer Science Engineering 2017; 2(6):10-13

RESEARCH ARTICLE

Implementation of Improved Apriori Algorithm on Large Dataset using Hadoop

[1]Deepak Mehta*, [2]Makrand Samvatsar

[1]Research Scholar, Patel College of Science and Technology, Indore, M.P, India
[2]Assistant Professor, Patel College of Science and Technology, Indore, M.P, India

Received on: 25/09/2017, Revised on: 30/10/2017, Accepted on: 26/11/2017

ABSTRACT

The association rule of data mining is an elementary topic in mining of data. Association rule mining discovery frequent patterns, associations, correlations, or fundamental structures along with sets of items or objects in transaction databases, relational databases, and other information repositories. The amount of data increasing significantly as the data generated by day-to-day activities. In data mining, Association rule mining becomes one of the important tasks of descriptive technique which can be defined as discovering meaningful patterns from large collection of data. Mining frequent itemset is very fundamental part of association rule mining. As in retailer industry many transactional databases contain same set of transactions many times, to apply this thought, in this thesis present an improved Apriori algorithm that guarantee the better performance than classical Apriori algorithm. Compare existing system and proposed system on the basis of execution time and memory. Found that proposed system taking less time and memory compare to existing system.

Keywords:-Hadoop, Map-Reduce, Apriori, Support and Confidence.

INTRODUCTION

Data mining is the main part of KDD. Data mining normally involves four classes of task; classification, clustering, regression, and association rule learning. Data mining refers to discover knowledge in enormous amounts of data. It is a precise discipline that is concerned with analyzing observational data sets with the objective of finding unsuspected relationships and produces a review of the data in novel ways that the owner can understand and use.

The incidence of data quality issues arises from the nature of the information supply chain [1], consumer of a data product may be several supply-chain steps removed from the people or groups who gathered the original datasets on which the data product is based. These consumers use data products to make decisions, often with financial and time budgeting implications. The separation of the statistics buyer from the data producer creates a situation where the consumer has little or no idea about the level of quality of the data [2], leading to the potential for poor decision-making and poorly allocated time and financial resources.

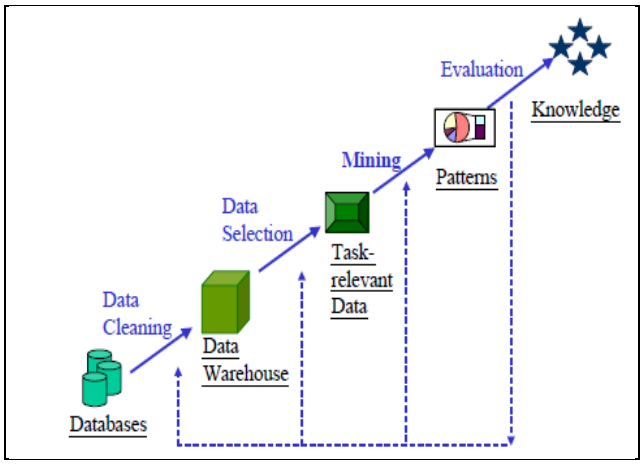

Figure 1: Process of Knowledge Discovery

Hadoop is an open source framework from Apache and is used to store process and analyze data, which are very huge in volume. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel with others. In short, Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data.

Hadoop Architecture At its core, Hadoop has two major layers namely:

***Corresponding Author:** Deepak Mehta**, Email**: deepak.mehta@meu.edu.in

- Processing/Computation layer (MapReduce),
- Storage layer (Hadoop Distributed File System)

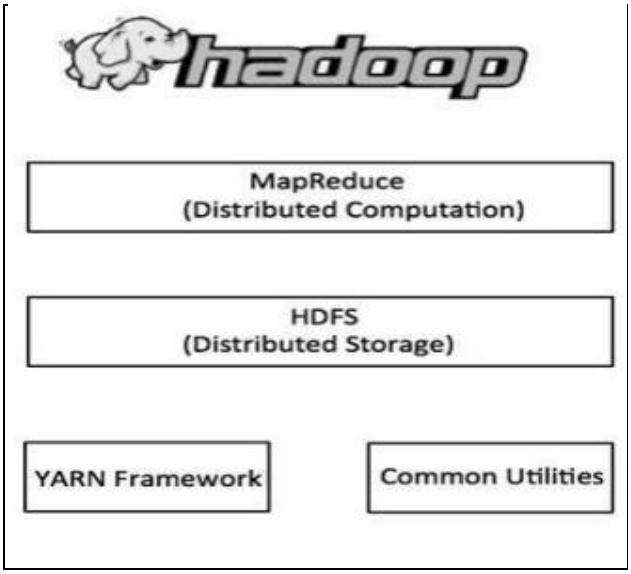

Figure2: Hadoop Architechure

EXISTING WORK

Apriori employs an iterative approach known as a level-wise search [5], where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found. This set is denoted L_1.$L_1$is used to find L_2, the set of frequent 2-itemsets, which is used to find L_3, and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the database. In order to find all the frequent itemsets, the algorithm adopted the recursive method. The main idea is as follows [6]:

Apriori Algorithm (Itemset [])
{

L_1 = {large 1-itemsets};
for (k=2; $L_{k-1} \neq \Phi$; k++) do
{
C_k=Apriori-gen (L_{k-1});
{
C_t=subset (C_k, t);
// get the subsets of t that are candidates
for each candidates c$\in$ C_t do
c.count++;
}
L_k={c$\in C_k$ |c.count$\geq$minsup}
}
Return=$\cup_k L_k$;
}

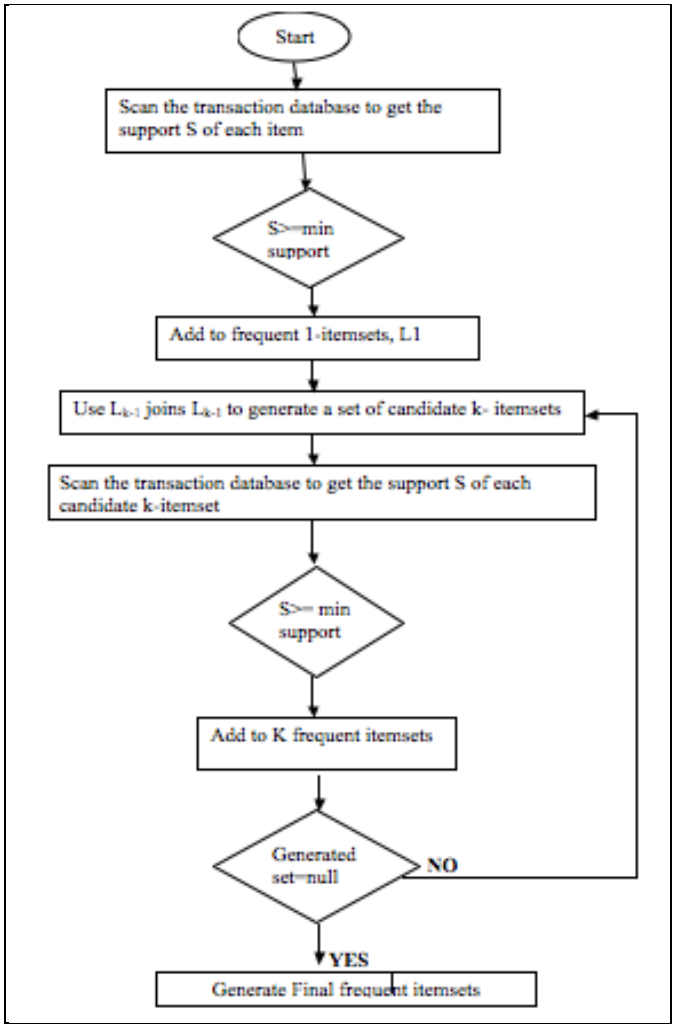

Figure 3: Flowchart of Existing System

PROPOSED SYSTEM

It is necessary to research on Apriori algorithm utilizing MAP-REDUCE (HADOOP) approach. The improved Apriori algorithm is generally used MAP-REDUCE (HADOOP) approach.

This new proposed method use the large amount of item set and reduce the number of data base scan. This approach takes less time than Apriori algorithm. The MAP-REDUCE (HADOOP) Apriori algorithm which reduce unnecessary data base scan.

Pseudo Code of Proposed Method

Proposed Apriori Algorithm

{

Input: database (D), minimum support (min_sup).
Output: frequent item sets in D.
L1= frequent item set (D)
j=k; /* k is the maximum number of element in a transaction from the database*/
for k= maxlength to 1 {
for i=k to 2{
for each transaction Ti of order i
{
if (Ti has repeated)


```
{
  Ti.count++;
 }
 m=0;
 while (i<j-m)
{
 if (Ti is a subset of each transaction Tj-m of order j-m)
{
 Ti.count++; m++; }
 }

 If (Ti.count >=min_sup)
 {
Rule Ti generated
}
}
}
```

Steps in Map Reduce

- Map takes a data in the form of pairs and returns a list of <key, value> pairs. The keys will not be unique in this case.
- Using the output of Map, sort and shuffle are applied by the Hadoop architecture. This sort and shuffle acts on these list of <key, value> pairs and sends out unique keys and a list of values associated with this unique key <key, list(values)>.
- Output of sort and shuffle will be sent to reducer phase. Reducer will perform a defined function on list of values for unique keys and Final output will<key, value> will be stored/displayed.

RESULT ANALYSIS

For the estimation purpose we have conducted several experiments using the existing dataset. Those experiments performed on computer with Intel i7 2.00GHZ CPU, 8.00 GB memory and hard disk 500GB. This algorithm was developed by java language using Net Beans IDE 8.3.1 and for the unit of measuring the time and no of iteration.

As a result of the experimental study, revealed the performance of our improved Apriori with the Classical Apriori algorithm. The run time is the time to mine the frequent itemsets.

Table 1 Execution time with respect to number of transaction

S.No	No. of Transaction	Time in Milli Second	
		Existing System	Proposed System
1	15	0.6	0.42
2	30	0.53	0.5
3	35	0.55	0.49
4	40	0.65	0.43
5	45	0.6	0.47

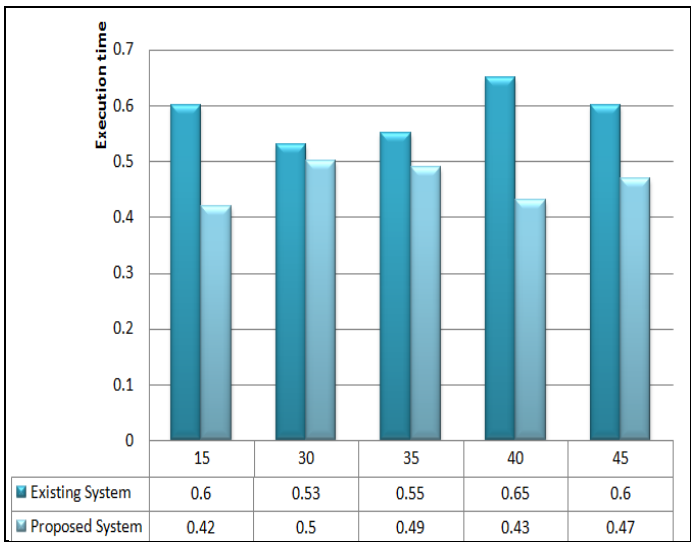

Figure 4: Execution time with respect to number of transaction

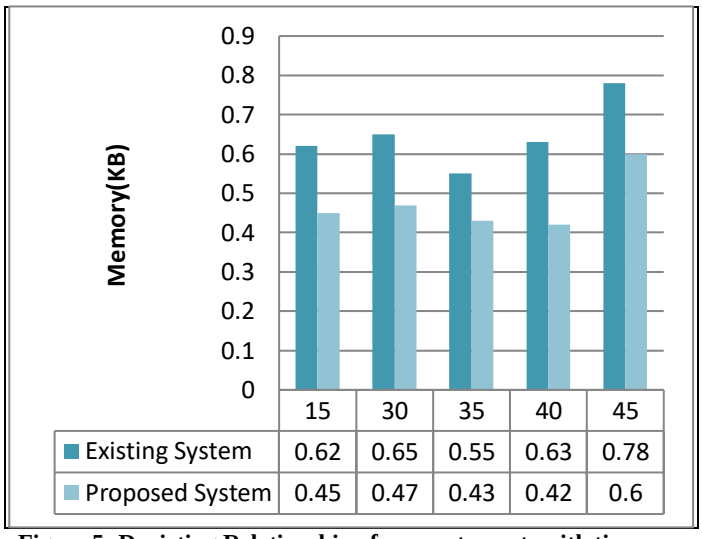

Figure 5: Depicting Relationship of support counts with time consumption

Table 2 Memory Comparison respect to number of transaction

S.No.	No. of Transaction	Memory in KB	
		Existing System	Proposed System
1	15	0.62	0.45
2	30	0.65	0.47
3	35	0.55	0.43
4	40	0.63	0.42
5	45	0.78	0.6

CONCLUSION

In this paper, we measured the following factors for creating our new idea, which are the time and the no of iteration, these factors, are affected by the approach for finding the frequent itemsets. Work has been done to develop an algorithm which is an improvement over Apriori with using an approach of improved Apriori algorithm for a transactional database. According to our clarification, the performances of the algorithms are strongly depends on the support levels and the features of the datasets(the nature and the size of

the datasets).There for we employed it in our scheme to guarantee the time saving and reduce the no of iteration Thus this algorithm produces frequent itemsets completely. Thus it saves much time and considered as an efficient method as proved from the results.

REFERENCES

1. Tan P. N., Steinbach M., and Kumar V: Introduction to Data Mining. Addison Wesley Publishers, 2006.
2. Han J. & Kamber M.: Data Mining Concepts and Techniques, First edition, Morgan Kaufmann publisher, USA 2001.
3. Ceglar, A., Roddick, JF: Association mining ACM Computing Surveys, volume 38(2) 2006.
4. Jiawei Han, Micheline Kamber, Morgan Kaufmann: Data mining Concepts and Techniques, 2006.
5. A. Savasere, E. Omiecinski and S. Navathe. : An efficient algorithm for mining Association rules in large databases, InProc. Int‟l Conf. VeryLarge DataBases (VLDB), Sept. 1995, p. p 432–443.
6. Agrawal. R and Srikant R.: Fast algorithms for mining association rules, InProc. Int‟l Conf. Very Large Data Bases (VLDB), Sept. 1994, p. p. 487–499.
7. Lei Guoping, DaiMinlu, Tan Zefu and Wang Yan: The Research of CMMB Wireless Network Analysis Based on Data Mining Association Rules, IEEE conference on Wireless Communications, Networking and Mobile Computing (WiCOM),ISSN :2161- 9646 Sept. 2011,p.p. 1-4.
8. Divya Bansal, Lekha Bhambhu : Execution of APRIORI Algorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, ISSN: 2277 128X September 2013 .
9. Shweta, Dr. Kanwal Garg: Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms International Journal of Advanced Research in Computer Science and Software Engineering 3(6), June – 2013, pp. 306-312.
10. Suraj P. Patil1, U. M.Patil2 and Sonali Borse: The novel approach for improving Apriori algorithm for mining association Rule,World Journal of Science and Technolog 2(3), ISSN: 2231 – 2587, 2012, p.p75- 78.
11. Toivonen. H.: Sampling large databases for association rules, In Proc. Int‟l Conf Very Large DataBases (VLDB), Bombay, India, Sept. 1996,p.p 134–145.
12. Yanfei Zhou, Wanggen Wan, Junwei Liu, Long Cai: Mining Association Rules Based on an Improved Apriori Algorithm 978-1-4244-585 8- 5/10/ IEEE 2010.
13. Luo Fang: The Study on the Application of Data Mining Based on Association Rules, International Conference on Communication Systems and Network Technologies (IEEE) ,may 2012,p.p 477 - 480 .

