AJCSE ASIAN JOURNAL OF COMPUTER SCIENCE ENGINEERING

www.ajcse.info

Asian Journal of Computer Science Engineering 2017; 2(4):13-16

RESEARCH ARTICLE

Hadoop Map-Reduce To Generate Frequent Item Set on Large Datasets Using Improved Apriori Algorithm

[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: 30/04/2017, Revised on: 18/07/2017, Accepted on: 30/07/2017

ABSTRACT

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 item set is very fundamental part of association rule mining. Many algorithms have been proposed from last many decades including horizontal layout based techniques, vertical layout based techniques and projected layout based techniques. But most of the techniques suffer from repeated database scan, Candidate generation (Apriori Algorithms), memory consumption problem and many more for mining frequent patterns. 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.

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.

Data mining as a field of study involves the integration of ideas from many domains rather than a pure discipline. The four main disciplines [1], which are contributing to data mining include:

- Statistics: it can make available tools for measuring importance of the given data, estimating probabilities and many other tasks (e.g. linear regression).
- Machine learning: it provides algorithms for inducing knowledge from given data (e.g. SVM).
- Data management and databases: in view of the fact that data mining deals with huge size of data, an efficient way of accessing and maintaining data is needed.
- Artificial intelligence: it contributes to tasks involving knowledge encoding or search techniques (e.g. neural networks).

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.

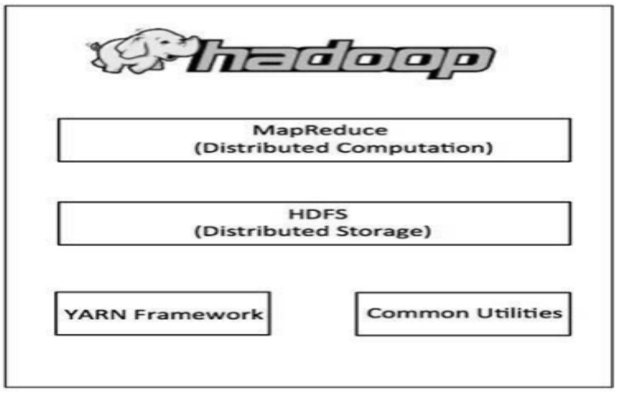

Figure1: Hadoop Architechure

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

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

AJCSE, July-Aug, 2017, Vol. 2, Issue 4

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

LITERATURE REVIEW

One of the most well known and popular data mining techniques is the Association rules or frequent item sets mining algorithm. The algorithm was originally proposed by Agrawal et al. [2] [4] for market basket analysis. Because of its important applicability, many revised algorithms have been introduced since then, and Association rule mining is still a widely researched area. Many variations done on the frequent pattern-mining algorithm of Apriori was discussed in this article.

AIS algorithm in [4] which generates candidate item sets on-the-fly during each pass of the database scan. Large item sets from preceding pass are checked if they were presented in the current transaction. Therefore extending existing item sets created new item sets. This algorithm turns out to be ineffective because it generates too many candidate item sets. It requires more space and at the same time this algorithm requires too many passes over the whole database and also it generates rules with one consequent item.

EXISTING WORK:

Apriori employs an iterative approach known as a level-wise search [15], 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++;
        }
$L_k=\{c \in C_k \mid c.count \geq minsup\}$
        }
          Return=$\cup_k L_k$;
}
```

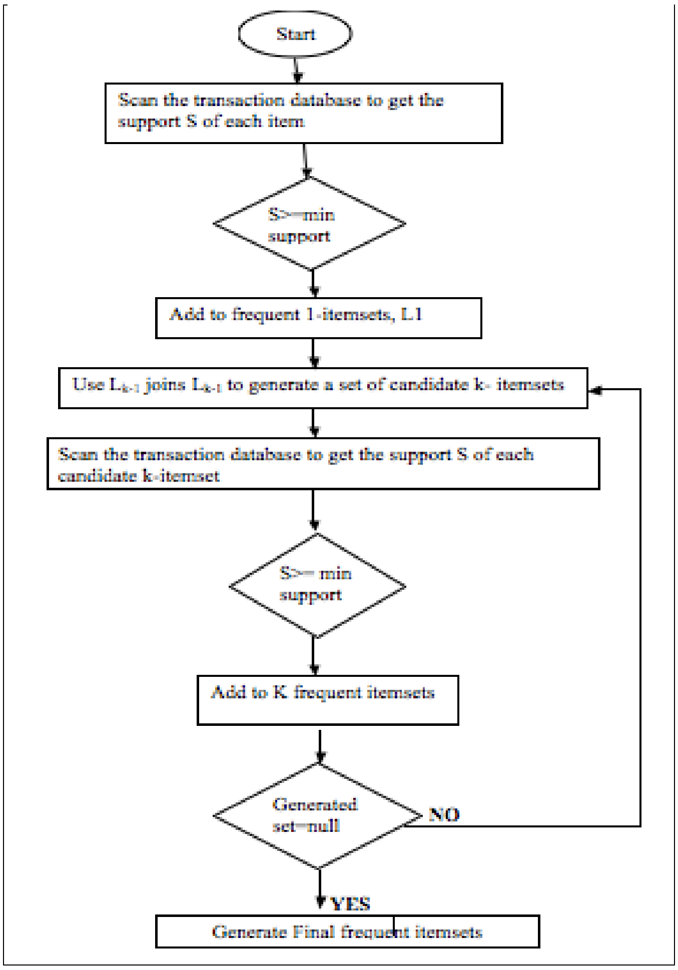
Start
Scan the transaction database to get the support S of each item
S>=min support
Add to frequent 1-itemsets, L1
Use Lk-1 joins Lk-1 to generate a set of candidate k- itemsets
Scan the transaction database to get the support S of each candidate k-itemset
S>= min support
Add to K frequent itemsets
Generated set=null
NO
YES
Generate Final frequent itemsets

Figure2: 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 reduces 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.

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 item sets. 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 data sets (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 item sets 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, J. F: 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‟lConf. VeryLarge Data Bases (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, Dai Minlu, 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. KanwalGarg: 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. SurajP .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 Data Bases (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 .