T.Anu Radha ,P.Lavanya / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 1, Issue 4, pp.1820-1822

# RECENT TRENDS IN PARALLEL AND DISTRIBUTED APRIORI ALGORITHM

**[1]T.Anu Radha[*] , [2]P.Lavanya**

[1]Asst.professor, Department of ECM, K.L University, A.P.

[2]Student, Department of ECM, KL University, A.P.

**Abstract: Society produces massive amounts of data from different sources like business, science, medicine, economics, sports, web data etc. Tremendous amounts of data is stored in databases, data warehouses and other information repositories. The availability of large datasets and increasing importance of data analysis for scientific discovery is creating a new class of high-end applications. This class of applications includes data mining and scientific data analysis. This paper trace out the recent trends in parallel and distributed apriori algorithm which is one of the top 10 algorithms in data mining.**

**Key Words:** ***Apriori, datasets, data mining.***

## I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. The large size and dimensionality of many databases makes data mining tasks too slow and too big to be run on a single processor machine. It is therefore a growing need to develop efficient parallel data mining algorithms.

## II. PARALLEL AND DISTRIBUTED DATA MINING

To make systems scalable, it is important to develop mechanisms that distribute the work load among several sites in a flexible way. Parallel data mining requires dividing up the work so that processors can make useful progress toward a solution as fast as possible. In modern parallel computers, access to the data set is likely to be most costly, followed by communication, with computation being relatively cheap. Data Mining often requires huge amounts of resources in storage space and computation time. Distributed Data Mining explores techniques of how to apply Data Mining in a non-centralized way.

## III. ASSOCIATION RULE MINING

Association rule mining is one of the most important and well researched techniques of data mining . An association rule is an implication of the form A $\Rightarrow$ B where A and B are transactional item sets taken from a transctional database. A very influential association rule mining algorithm Apriori has been developed for rule mining in large transactional databases. Association rule discovery techniques have gradually been adapted to parallel systems in order to take advantage of the higher speed and greater storage capacity that they offer.

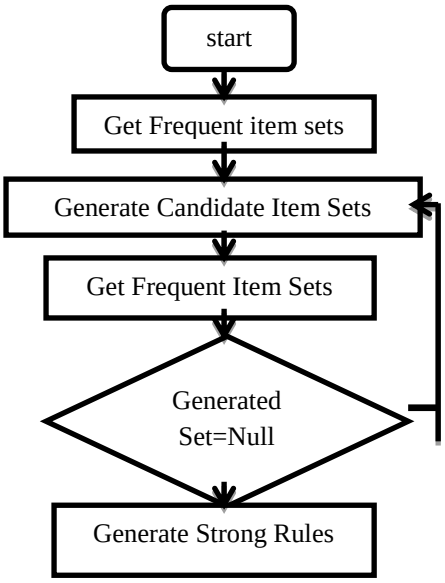
## IV. APRIORI

The Apriori algorithm by Rakesh Agrawal and colleagues has emerged as one of the best ARM algorithms. It also serves as the base algorithm fore most parallel algorithms. Apriori uses a complete, bottom-up search with a horizontal layout and enumerates all frequent item sets. An iterative algorithm, Apriori counts item sets of a specific length in a given database pass. The main property of apriori algorithm is that all non-empty subsets of a frequent item set must also be frequent.

A. Algorithm

$C_k$: Candidate item set of size k
$L_k$ : frequent item set of size k
$L_1$ = {frequent items};
**for** ($k$ = 1; $L_k$ !=∅; $k$++) **do begin**
$C_{k+1}$ = candidates generated from $L_k$;
**for each** transaction $t$ in database do
increment the count of all candidates in $C_{k+1}$ that are contained in $t$
$L_{k+1}$ = candidates in $C_{k+1}$ with min_support
**end**
**return** $\cup_k L_k$;

B. *Flow Chart*




Many parallel algorithms use Apriori as the base method, because of its success in the implementation.

## V. APRIORI BASED ALGORITHMS

***A.*** *Dynamic Hashing and Pruning*
Association Rule Mining using Hash Based Algorithm to filter the unnecessary items can be found in an effective hash based algorithm for mining association rules in works by Jang et al., (1995), John and Soon (2002), and Han et. al. (2006).

***B.*** *Partition*
Partition algorithm proposed by Ashok Savasere and others is a fundamentally different algorithm that reads the database atmost two times to generate all significant association rules.

***C.*** *Eclat - Frequent item set mining*
A program to find frequent item sets (also closed and maximal as well as generators) with the eclat algorithm (Zaki et al. 1997), which carries out a depth first search on the subset lattice and determines the support of item sets by intersecting transaction lists.

***D.*** *Intelligent Data Distribution and Hybrid Distribution*
Eui-Hong Han and his colleagues have proposed two ARM methods based on Data Distribution using the platform of a 128-node Cray T3D DMM in which Data distribution uses an expensive all-to-all broadcast to send local database portions to every other processor.

***E.*** **N***on Partitioned, Simply Partitioned, and Hash-partitioned Apriori*
Takahiko Shintani and Masaru Kitsuregawa proposed these algorithms with their target machine as 64-nodeFujitsu AP1000DDV DMM. Non Partitioned Apriori is essentially the same as Count Distribution, except that the sum reduction occurs on one master processor. Hash-Partitioned Apriori is similar to Candidate Distribution. Each processor generates candidates from the previous level's frequent set and applies a hash function to determine a home processor for that candidate.

***F.*** *DMM*
For mining maximal frequent item sets from databases, an algorithm named Distributed Max-Miner (DMM) is proposed.

***G.*** *FDM*
Cheung et al. presented an algorithm called FDM. FDM is a parallelization of Apriori to shared nothing machines.

These algorithms have been implemented on single core processors so far. Implementation of these algorithms on multi core processors have been under research. By implementing on multi cores several factors of processors can be improved.

## VI. CONCLUSION

Association Rule mining is to gather the required information. Parallel data mining mainly refers to dividing the work among processors which is less scalable. Distributed Data Mining explores techniques of how to apply Data Mining in a non-centralized way. The base algorithm used for the development of several association rule mining algorithms is apriori which works on non empty subsets of a frequent itemset. This paper is a review of various parallel and distributed association rule mining algorithms which are developed based on the apriori algorithm.

## ACKNOWLEDGMENTS

The Authors like to express their thanks to the management of K L University and department

of ECM for their continuous encouragement and support during this work.

## REFERENCES

[1] Agrawal, R et.al. , (1994) Fast algorithms for mining association rules. In Proc. 20th Int. Conf. Very Large Data Bases.

[2]. Agrawal, R et al., (1996), "Fast Discovery of Association Rules," Advances in Knowledge Discovery and Data Mining.

[3]. Agrawal, R., Imielinski, T., and Swami et. al (1993). Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 207-216.

[4]. Ashok savasere et.al (1995) an efficient algorithm for mining association rules in large databases.21stVLDB conference.

[5]. Cheung, D., Han, J., Ng, V., Fu, A. and Fu, Y. (1996), A fast distributed algorithm for mining association rules, in `Proc. of 1996 Int'l. Conf. on Parallel and Distributed Information Systems.

[6]. Cheung D et.al (1998) "Asynchronous Parallel Algorithm for Mining Association Rules on Shared-Memory Multi-Processors

[7]. Congnon leo et.al. , (2006)journal of super computing.

[8]. Huang Darong et.al.,(2010)fuzzy systems and knowledge discovery.

[9]. John and soon data warehousing and knowledge discovery (2000).lecture notes in computer science.

[10]. J.S.Park et.al (1995) an effective hash based algorithm for mining association rules in ACM SIGMOD International Conference on Management of Data.

[11]. Sujni Paul; Saravanan, V.(2008) Computer Science and Information Technology, 2008. ICCSIT '08.

[12]. Zaki, M. J., et. al., (1999) Parallel and distributed association mining: A survey. IEEE Concurrency,Special Issue on Parallel Mechanisms for Data Mining.

[13]. zaki et. al (1997) the third International conference on knowledge discovery and data mining.