Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Apriori Algorithm for Mining Frequent Patterns using Parallel Computing: Survey

Rajdeep Kaur, Puneet, Deepika
Chandigargh University, Gharuan,
Punjab, India

Abstract-: ***Apriori algorithm is useful for mining frequent pattern from large databases. Number of the techniques is used for the frequent pattern mining which associates the dataset with each other and most useful algorithms are Apriori & FP-growth algorithms. This paper presents the survey of Apriori algorithm for frequent pattern mining used to calculate the association in different data sets and apply the parallel computing to increase the execution speed and to reduce the cost parameters. The analysis of literature survey would give the information about what has been done previously in frequent pattern mining, what is the current trend and what the other related areas are and presents efficient scalable Multi-core processor parallel computing that reduce the execution time and increase performance. For the multi core utilization, Java concurrency libraries package are used which execute the independent functions on multiple cores of the processor to improve the speed. Java concurrency libraries create the threads equal to the number of the cores in processor.***

Keywords- Parallel processing, Frequent pattern mining, FP growth, Apriori, Java concurrency Library

I. INTRODUCTION

Frequent pattern mining [1] plays a significant role in research and it is a part of data mining. The main focus of the survey is mining of frequent patterns by using apriori algorithm which is suitable for calculating the association rules for the dataset and generate the rules for the dataset. Apriori algorithm is basically used for mining the frequent patterns and associated datasets from the large databases. It is also used for the market basket analysis that discovers the relationship between the different datasets. It find the association rules by analyzes the transactions of the different databases and on the basis of the analysis[2], it find the association rules which discover the interesting relations between the different transactions of database. Association rules include the confidence and support.

Apriori algorithm: Candidate generation approach

Apriori algorithm is basically divided into two steps[3,4]. It generates the association rules for the frequent itemsets of transactional databases. It is based on the apriori principle that all the nonempty subsets of a frequent itemset must be frequent. It is a two step process.

Step 1: Find frequent itemsets from the database

The whole transactional database is scanned to find the count of candidate set Ck where Ck represents the frequency of item occur in the database. The frequency of the each itemset Ck is compared with the threshold value of support which is predefined, if the count of itemset is greater than the minimum support then itemset placed in frequent k-itemset Lk , otherwise discarded from the candidate set.

Step 2: Assocation rule formulation

Lk is joined with itself to find the next candidate k+1 itemset Ck+1. This step is iterated for every candidate step .When all the items in the database are combined with each other and if that set satisfy the minsup criteria then association rules are formed for the frequent item set. As in the following example rules are formed for itemset {2 3 5}.

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	Support
{1 3}*	2
{1 4}	1
{3 4}	1
{2 3}*	2
{2 5}*	3
{3 5}*	2
{1 2}	1
{1 5}	1

C_3

Itemset	Support
{1 3 4}	1
{2 3 5}*	2
{1 3 5}	1

Association rules are formed for the set of items that have the support greater than the minimum support and confidence greater than the minimum required confidence and lift of the rule is defined as the ratio of the support of frequent item to the expected if both X and Y are independent

$$Rule: X \Rightarrow Y \quad Support = \frac{frq(X,Y)}{N} \quad Confidence = \frac{frq(X,Y)}{frq(X)} \quad Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

1.1 Scope for Parallel Processing in Data Mining Algorithms

Data mining plays significant role in market for the analysis of large databases. Different algorithms used for the analysis of large databases take more cost and computation time. To reduce the computation cost and time data mining algorithms are parallelized. Data mining algorithm parallelization can be perform because most of the data mining tasks involve subtasks are like to finding of maximum, minimum, count, average etc[5,6]. These tasks can be performed on independently and on different partitions of a dataset. Then the sub solutions can be integrated for the final results.

Parallelism in Association Rule Mining: The Apriori algorithm is basically used for finding frequent patterns and association rule mining from the large databases. There are two steps in apriori to find the frequent pattern when the size of dataset is k+1. In the first step it finds a set of candidate frequent item sets by using join operation and in the second step entire dataset is searched to count the frequency of items appear together in all the transactions of the database. Now parallelism can be performed by distributing the dataset to the multiple cores of the processor simultaneously. Parallelism can also be performed in the first step to increase the speed to obtain frequent items from the transactions of the database. Global frequent item is obtained by integrating local frequent items and sometimes it must be a local frequent item for some data partition to achieve this there is need of communication for exchanging support counts. Time taken by the different cores of the processor also added in computing time. So parallelism of apriori is useful only for the large databases.

Parallel computing on multiple cores of the processor reduce the execution time and increase performance. Communication is required in between the multi cores of the processor and implemented by using shared memory or message passing. Independent functions of the application executed by the multiple threads and threads are assigned to the multiple cores of the processor but improvement of parallelism depends upon the functions of the program that can be parallelized and can be executed on the multiple cores of the processor[7,8]. Multi threading feature of the java enables to write different concurrent applications where different threads can be executed simultaneously. The Java Concurrency framework is a library that is basically designed to building blocks and to create concurrent applications. java.util.concurrent is a package defines the executor framework which provides scheduling scheme and thread management of different application independent functions.

II. SUMMARY OF THE SURVEY

Multi core architecture and parallel computing is useful only for the applications in which number of the cores of the processor running the same algorithm independently. The partial support of all the candidate itemsets are generated from the local database partitioning parallel and at the end after each iteration global supports are calculated from partial support counts of all the processors[8]. The database is partitioned into the small parts and different partitioning are assigned to the different cores. The entire candidate set Lk-1 is computed by all the cores in parallel fashion. Each core of the processor independently compute the partial supports of the candidate set from its local database partition. Then global Ck counts is calculated by exchanges local counts Ck of all the other processors. Lk is computed from the Ck of each processor[9]. After computing global Lk , each processor computes Ck+1 in parallel and the whole process is repeated until all frequent itemsets are found.

Steps of Apriori Algorithm on multiple cores of processor:

1. Analyze the transactional database and generate first candidate set which is unique from a transactional dataset.
2. Create the number of threads equal to the number of cores in processor using executor object.
3. Calculate the partial support count for each itemset and submit itemsets in Completion Service object for the processing.
4. Store global support count value in a variable, calculated by multiple cores of processor.
5. Generate a frequent item set by comparing it with minimum support value.
6. Generate new candidate itemset by incrementing one item.
7. Repeat steps from 5 to 6 until frequent itemset is empty.
8. End.

III. CONCLUSION

In this paper, we are proposing apriori algorithm for mining frequent patterns using parallel computing. Apriori algorithm use java concurrency package for parallel computing by distributing the dataset to multiple cores of the processor simultaneously. Apriori algorithm run on a multi core processor with two datasets using various minSup and confidence to generate the association rules[10,11]. We analyze the execution of apriori algorithm when run sequentially and parallel

with respect to the parameters speed and cost. Apriori algorithm computed on multi core architecture takes the less time as compare to the simple Apriori algorithm. The other performance parameters also depends upon the processing time and the data communication cost. The data communication cost can be reduced by using client server architecture like Parallel Partitioning Algorithm and exchanging only the counts as in Count Distribution Algorithm. The processing time depends on the database layout, number of times the database is scanned and the size of the candidates generated[12]. But the apriori algorithm using parallel computing only useful to find the frequent patterns from large databases. For small database it may be take more time as compare to sequential execution because it also added the time required for the communication in between the multiple cores of the processor[13].

REFERENCES

[1] Chen, M. S., Han, J., & Yu, P. S. ―Data mining: An overview from a database perspective‖, IEEE Transactions on Knowledge and Data Engineering,‖ 8(6), 866–883,1996.

[2] Gosta Grahne, Member, IEEE, and Jianfei Zhu, Student Member, "Fast Algorithms for Frequent Itemset Mining Using FP-Trees," IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 10, October 2005 1347.

[3] Jianwei Li, Ying Liu, Wei-keng Liao, Alok Choudhary, ―Parallel Data Mining Algorithms for Association Rule and Clustering,‖ 2006 by CRC Press, LLC

[4] Li Liu2, 1, Eric Li1, Yimin Zhang1, Zhizhong Tang, "Optimization of Frequent Itemset Mining on Multiple-Core Processor" VLDB _07, Vienna, Austria, 2007.

[5] LIU Xiang. "An Agent-based Architecture for Supply Chain Finance Cooperative Context-aware Distributed Data Mining Systems". 3rd Intl. Conference on Internet and Web Applications and Services, IEEE-2008.

[6] Jinbiao Hou. "Research on Distributed Data Mining in E-commerce Environment Based on Web Services and Mobile Agent". Intl. Conf. on Computational Intelligence and Natural Computing, IEEE-2009.

[7] C. E. Leiserson and I. B. Mirman, "How to Survive the Multicore Revolution (or at Least Survive the Hype)," Journal of Advancing Technology. vol. 9, pp. 43-53, 2009.

[8] J. M. Bull, J. P. Enright, X. Guo, C. Maynard, and F. Reid, "Performance Evaluation of Mixed-Mode OpenMP/MPI Implementations," International Journal of Parallel Programming, vol. 38, no. 5-6, pp. 396-417, 2010

[9] V.Umarani, Dr.M.Punithavalli, "A Study On Effective Mining Of Association Rules From Huge Databases", International Journal of Computer Science and Research (IJCR), Vol. 1 Issue 1, 2010.

[10] Sunil Joshi et al: accepted research paper in The IEEE 2010 International Conference on Communication software and Netweorks (ICCSN 2010) on "A Dynamic Approach for Frequent Pattern Mining Using Transposition of Database" from 26 - 28 February 2010

[11] Ruowu Zhong, Huiping Wang, "Research of Commonly Used Association Rules Mining Algorithm in Data Mining", In Proceedings of International Conference on Internet Computing and Information Services(ICICIS), IEEE, September 2011

[12] Khadidja Belbachir, Hafida Belbachir, "The Parallelization of Algorithm Based on Partition Principle for Association Rules Discovery", In Proceedings of International Conference on Multimedia Computing and Systems(ICMCS), IEEE, May 2012.

[13] Mr.Kiran C. Kulkarni1, Mr.R.S.Jagale2, Prof.S.M.Rokade3, A Survey on Apriori algorithm using MapReduce Technique, International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008Certified Journal,, Volume 3, Special Issue 4, March 2013)

