© 2019 IJSRSET | Volume 6 | Issue 1 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : 10.32628/IJSRSET196172

Efficient Algorithm for Frequent Item Set Generation in Big Data

Priyanka Wankhede[1], Prof. Vijaya Kamble[2]

[1]Department of Computer Science and Engineering, Gurunanak Institute of Engineering and Technology, Nagpur, India

[2]Assistant Professor, Department of Computer Science and Engineering, Gurunanak Institute of Engineering and Technology, Nagpur, India

ABSTRACT

Data mining faces a lot of challenges in the big data era. Association rule mining is an important area of research in the field of data mining. Association rule mining algorithm is not sufficient to process large data sets. Apriori algorithm has limitations like the high I/O load and low performance. The FP-Growth algorithm also has certain limitations like less internal memory. Mining the frequent itemset in the dynamic scenarios is a challenging task. To overcome these issues a parallelized approach using the mapreduce framework has been used. The mining algorithm has been implemented using the Hadoop.

Keywords : Incremental FP-Growth Algorithm, Big Data, Data Mining, Frequent Itemset Mining.

I. INTRODUCTION

With the fast development of networking, data storage, and the data collection capacity, the size of databases is rapidly growing in all domains.

Big data typically includes masses of unstructured data that requires more real-time analysis. Big data analytic by machine learning and data mining techniques has become an important research problem. Mining with big data is very difficult problem when the current data mining methodologies tools with a single personal computer are used to deal with very large datasets due to their large size and complexity.

Big Data has special characteristics that make it an extreme challenge for discovering useful knowledge. These characteristics including, large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data.

Data mining is an emerging concept which is powerful and promising and advances in data collection and storage technologies have led organizations to store vast amount of data pertaining to their businesses and extracting useful information from such vast amount of data is an important activity and is referred to as data mining or knowledge discovery in databases.

Applying frequent itemset mining to large databases is problematic. First of all, very large databases do not fit into main memory. In such cases, one solution is to use levelwise breadth first search based algorithms, such as the well-known Apriori algorithm. where frequency counting is achieved by reading the dataset over and over again for each size of candidate itemsets. Unfortunately, the memory requirements for handling the complete set of candidates itemsets blows up fast and renders Apriori based schemes very inefficient to use on single machines. Secondly, current approaches tend to keep the output and runtime under control by increasing the minimum frequency threshold, automatically reducing the number of candidate and frequent itemsets. However, studies in recommendation systems have shown that itemsets with lower frequencies are more interesting.

Parallel programming is becoming a necessity to deal with the massive amounts of data, which is produced and consumed more and more everyday. Parallel programming architectures, and hence the algorithms, can be grouped into two major subcategories: shared memory and distributed (share nothing). On shared memory systems, all processing units can concurrently access a shared memory area. On the other hand, distributed systems are composed of processors that have their own internal memories and communicate with each other by passing messages.

II. Objective

The association rule mining (ARM) contains two phase

- Mining the Frequet Itemset (FIs)
- Association Rule (AR) extraction

Apriori Algorithm is considered as one of the most influential data mining algorithm to extract knowledge in the forms of ARs or FIs

Horizontal parallel-Apriory (HP-Apriory)which mines the FIs from big data in a parallel way.

This algorithm partitions the dataset both horizontally and vertically into four subsets.

Also,it devides the itemset mining task into three sub-task, including

- candidate generation
- support count calculation
- results combination.

III. Literature Review

1) “The AIS algorithm“was the first algorithm proposed by Agrawal, Imielinski, and Swami for mining association rule”. AIS algorithm depends on scanning the databases many times to get the frequent itemsets . “The support count of each individual item was accumulated during the first pass over the database. Based on the threshold of support count those items whose count is less than its minimum value are eliminated from the list of items. Candidate 2-itemsets are generated by extending frequent 1-itemsets with other items in list. “During the second pass over the database, the support count of those candidate 2-itemsets are accumulated and checked against the support threshold”. The candidate itemsets generation and frequent itemsets generation process iterates until any one of them becomes empty.“AIS Algorithm has efficiency problems so some modifications have been introduced to give an estimation for candidate itemsets that have no hope to be large, consequently the unnecessary effort of counting those itemsets can be avoided” .“Also since all the candidate itemsets and frequent itemsets are assumed to be stored in the main memory, memory management is also proposed for AIS when memory is not enough”.

2) Original Apriori Algorithm is one of the well-known algorithms for mining frequent itemsets. It was introduced in (R. Agrawal, 1993). “The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemsets properties” . “Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k + 1)-itemsets. First, the frequent 1-itemset is found, this is denoted by L1, which is used to find the frequent 2-itemset L2 and so on”. “To improve the efficiency of the level-wise generation of frequent itemsets, a property called Apriori property is used to reduce the search space”. “This property states that all nonempty subsets of a frequent itemset must also be frequent”.

3) “FP-Growth Algorithm is the most popular frequent itemset mining algorithm that was introduced in” (J. Han, 2000). “The main aim of this algorithm was to remove the bottlenecks of the Apriori algorithm in generating and testing candidate sets”. “The problem of the Apriori algorithm was dealt with by introducing a novel, compact data structure called frequent pattern tree or FP-tree”. “Then based on this structure an FP-tree based pattern fragment growth method was developed” . “FP-Growth uses a combination of the vertical and horizontal database

layout to store the database in main memory". "Instead of storing the ID for every transaction in the database, it stores the actual transactions from the database in a tree structure and every item has a linked list going through all transactions that contain that item". "This new data structure is denoted by FP-tree (Frequent Pattern tree)"

4) CARMA Algorithm (C. Hidber et al, 1999) Continuous Association Rule Mining Algorithm. The algorithm mainly aims to process large online datasets . CARMA algorithm comprises two phases during its operation: in the first phase CARMA builds "lattice of all potential large itemsets with respect to the scanned part of the data. "For each set on the lattice CARMA determines the lower and upper limits of its support". The deduced association rules are displayed to the user after processing transactions and the user is free to adjust the support count . In the second phase after getting the user's feedback CARMA fully scans all the dataset to determine exactly the occurrences of each itemset and removes all the itemsets below the user's specified threshold.

IV. Research Methodology

One of the mostly used algorithms in association rule mining is the Apriori algorithm. In this algorithm the high dimensional frequent item sets are obtained using the low dimensional frequent item sets.

The process is iterative. But the basic Apriori algorithm faces some major drawbacks such as the number of database scans required is more. This results in the I/O overburdened; the process of obtaining frequent item sets at the higher level from the lower level needs more time as the number of itemsets are more at the lower level. ; the basic Apriori algorithm should not consider the transactions that need not scan. Most of the researchers have proposed the various improved versions of the Apriori algorithm.

In the DHP algorithm proposed by Park et al. the algorithm is based on dynamic hash hashing algorithms and pruning algorithm. Here the transactions that are not involved in generating frequent item sets are not considered while traversing the database. Thus the efficiency for frequent itemset mining is improved. The dynamic itemset counting algorithm proposed by Brin requires a smaller number of database scans. In this algorithm the transaction database is divided in to data blocks of same size.

These information squares are gotten to for producing the I-incessant itemsets then the hopeful 2-successive item sets independent from anyone else join, finally it consolidates the I-continuous itemsets and competitor 2-reguJar item sets that every piece created. The procedure is rehashed until there is no new itemsets or achieve the breaking point.

These algorithms are able to perform well when the size of the database is small or the dimensionality of the data is not so high. But while considering the big data, these algorithms cannot perform well. MapReduce framework produced by Googlein 2004 is able to handle the big data. In the similar way, Hadoop based on MapReduce, cluster based parallel data mining are some of works for big data mining. Various classical data mining algorithms are based on Hadoop. Various parallel algorithms using Apriori such as SPC, FPC and DPC are proposed by Lin et al. These are based on MapReduce.

In the SPC algorithm the dataset is mapped to all Map nodes and mining is performed parallelly. Afterwards the combining operation is executed in the reduce phase. SPC algorithm required to start the map and reduce phase only once.

The DPC and FPC algorithm needs to repeatedly start the Map and Reduce phase. The process is defined by the number of dimensions of the frequent item sets mining; The parallel frequent item set mining algorithm proposed by Li et al scans the transactions database for counting the frequent item sets in the Map stage, and the statistical operations are performed for obtaining frequent item sets in the Reduce phase, but But this algorithm also required to start MapReduce tasks repeatedly.

S. Hammoud proposed a parallel in each round of the iterative procedure, where cut datasets are alloted to every Map hub and measurable competitor incessant item sets, then converging to get continuous itemsets in the Reduce stage. As a result of the MapReduce system's high dormancy and absence of emphasis, Apriori calculation doesn't fit the MapReduce structure well. Spark is a memory-based parallel registering system, and it can incredibly enhance the ongoing information handling and guarantee the group's high adaptation to internal failure and high versatility in enormous information situations.

A parallel Apriori algorithm was introduced by Qiu H et al which is related to Spark YAFIM. The results obtained using SPARt are more promising than that using Hadoop.

V. EXPECTED OUTCOME& FUTURE SCOPE

We regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. With Big Data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real-time. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of Big Data has arrived.

Several issues can be considered in future work.

- First, more large scale datasets containing various amount of data samples, different numbers of features (i.e.dimensionalities), and different feature types including categorical, numerical, and mixeddata types can be used for further comparisons.
- Second, in addition to constructing the Apriori classifier, the performances of using other classification techniques under the different procedures can be examined. Last but not least, it would be useful to investigate the effect of using different computing hardware environments on the different procedures.

VI. CONCLUSION

In this paper, we proposed a new technique for mining Frequent Item Sets that uses Horizontal as well as vertical Association rule mining in parallel using the multithreading concept in java. Our approach reduces the total step of the process and also takes less time and less memory.

VII. REFERENCES

[1]. "Horizontal Format Data Mining with Extended Bitmaps", Buddhika De Alwis1, Supun Malinga2, Kathiravelu Pradeeban3, Denis Weerasiri4, Shehan Perera, International Journal of Computer Information Systems and Industrial Management Applications. ISSN 2150-7988 Volume 4 (2012) pp. 514-521

[2]. J. Han and M. Kamber. "Data mining: Concepts and Techniques", Morgan Kaufman, San Francisco, CA,2001.

[3]. Arun K Pujari "Data Mining Techniques" UniversityPress (India) Pvt. Ltd 2001

[4]. C. F. Tsai, W. C. Lin, and S. W. Ke, "Big data mining with parallel computing: a comparison of distributed and MapReduce methodologies", Journal of Systems and Software, vol. 122, pp. 83-92, 2016.

[5]. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Proc. VLDB, pages 487-499, 1994

[6]. B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalization based on association rule discovery from web usage data. In Proc. WIDM, pages 9-15. ACM, 2001

[7]. X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data", IEEE Transactions on Knowledge and Data Engineering, vol. 26, no.1, pp. 97-107, 2014.

[8]. R. Agrawal, and J. C. Shafer, "Parallel mining of ARs", IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 962-969, 1996

[9]. W. Fan, and A. Bifet, "Mining big data: current status, and forecast to the future", ACM SIGKDD Exploration, vol. 14, no. 2, pp. 1-5, 2012..

[10]. M. J. Zaki, M. Ogihara, S. Parthasarathy, and W. Li, "Parallel data mining for ARs on shared-memory multi-processors", In Conference of Supercomputing, pp. 43-43, 1996.