|| Volume 5 || Issue 7 || July 2020 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

SURVEY ON MINING FREQUENT PATTERN ON BIG DATA USING MAP REDUCING TECHNIQUE

Komal Ramchandra Jadhav[1], Prof. Pravin. P. Nimbalkar[2]

PG Student, Dept.of Computer Engineering, ICOER , Wagholi, Pune[1]
Assistant Professor, Dept.of Computer Engineering, ICOER , Wagholi, Pune[2]
jadhavkomal56@gmail.com[1], ppnimbalkar1@gmail.com[2]

Abstract: Recently, There incorporates a fast development of internet and as fast growing cluster users, several corporations have to manage higher amount of data every day. Acquiring important information quickly from this continuously growing data is vital issue. Frequent pattern mining is a good approach to get correlation in dataset. The foremost well-liked data mining Apriori algorithm that mines frequent item set has downside that computation time will increase once data size will increase. The planned models are supported the well-known Apriori algorithmic program and also the MapReduce framework. The planned algorithms are divided into three main groups. Two algorithms are properly designed to extract patterns in giant datasets. These algorithms extract any existing item-set in data regardless their frequency. Pruning the search space by suggests that of the antimonotone property. Two additional algorithms space pruning are planned with the aim of discovering any frequent pattern available in data. Maximal frequent patterns. A last algorithm is also proposed for mining condensed representations of frequent patterns, i.e., frequent patterns with no frequent supersets.

Keywords: *Big Data, Hadoop, Data Mining.*

I INTRODUCTION

Data Mining is a vital facet of each organizations growth. Every company has lots of data to be accessed and processed. Those data should be handled in such the way that there is no any vital data loss. Data mining handles tasks such as data classification, data clustering, text classification, frequent pattern mining, semantic web mining, regression, summarization, prediction, combinations, sequential patterns etc. Our project involves the task of determinative frequent pattern from a given data set . This approach can realize the frequent patterns within which user is interested[7], the growing interest in data has caused the performance of existing pattern mining techniques to be born [1]. The goal is to propose new efficient pattern mining algorithms to figure in big data. To the present aim, a series of algorithms supported the MapReduce framework and the Hadoop open-source implementation have been proposed [9]. Pattern mining is one of the most important tasks to extract meaningful and useful information from raw data. This task aims to extract item-sets that represent any type of homogeneity and regularity in data. MapReduce is an emerging paradigm that has become very popular for intensive computing. Pruning the search space by means of the antimonotone property [6]. Two additional algorithms space pruning AprioriMR (SPAprioriMR) and top AprioriMR (TopAprioriMR)] are planned with the aim of discovering any frequent pattern available in data.

The Apriori Algorithm based mostly frequent mechanical phenomenon pattern mining algorithmic program to with efficiency and effectively handle the trajectory information dealings. Advantage of this technique is later iterations are much faster than the initial iterations of the algorithm. The results obtained by this method are more accurate and reliable. Generation of candidate coordinate sets is dear (in each space and time). Since generation and pruning steps are in square measure in memory resident, it desires additional RAM. Another disadvantage is it needs n+1 database scans, n is the length of the coordinates within the longest pattern.[2]

Mining class association rules (CARs) with the item set constraint is bothered with the invention of rules, that contain a collection of specific items within the rule antecedent and a class label in the rule subsequent. This task is often encountered in mining medical data. There are two naive strategies to solve this drawback, namely pre-processing and post-processing. The post-processing methods have to generate and consider a huge number of candidate CARs

while the performance of the pre-processing methods depend on the number of records filtered out. efficient technique for mining CARs with the itemset constraint supported a lattice structure and also the distinction between two sets of object identifiers (diffset)[1]

The most existing algorithms mine frequent patterns from traditional transaction databases that contain precise data. However, there are many real-life situations in which one needs to deal with uncertain data. The uncertainty of such suspicion can be expressed in terms of existential probability. An efficient algorithms for mining uncertain data are in demand. Two algorithms for mining frequent patterns from uncertain data. The algorithms follow the horizontal data representation. of mining frequent itemsets from existential uncertain data using the Tidset vertical data representation.UF-growth algorithm is conducted and showed that the algorithm outperforms the UF-growth.[3]

Efficient pattern mining algorithms to work in big data. All the models are based on the well-known Apriori algorithm. This algorithm has been also for mixing condensed representations of frequent patterns. Pruning the search space by means of anti-monotone property. Two additional algorithms have been with the aim of discovering any frequent pattern available in data. Ranking functions are evaluated by a variety of means; one of the simplest is determining the precision of the first k top-ranked results for some fixed k; Frequently, computation of ranking functions can be simplified by taking advantage of the observation that only the relative order of scores matters, not their absolute value; hence terms or factors that are independent of the features may be removed, and terms or factors that are independent of the feature may be pre-computed and stored with the dataset.[4]

II PATTERN MINING APPROACHES

Collections of items which appear in a data set at an important frequency and that can thus support association rules and describes relations between variables is called as Frequent patterns. a day to reduce and compare the candidate patterns.

Frequent patterns are required to be identified to know the hidden facts in the dataset. Frequent patterns can easily adapt to the data mining tasks. Identifying the frequent pattern consumes less time. From a frequent pattern, It is easy to find the frequent items in the data sets and to represent the relationship between the datasets. The frequent pattern mining is an active method used now

a. Market Basket Analysis

Frequent patterns area unit patterns that seem oft among a dataset (surprised?). A frequent itemset is one that is created from one in all these patterns, that is why frequent pattern mining is commonly alternately brought up as frequent itemset mining.

Frequent pattern mining is most simply explained by introducing market basket analysis (or affinity analysis), a typical usage that it's well-known. Market basket analysis tries to spot associations, or patterns, between the varied things that are chosen by a specific shopper and placed in their market basket, be it real or virtual, and assigns support and confidence measures for comparison. the worth of this lies in cross-marketing and client behavior analysis.

The generalization of market basket analysis is frequent pattern mining, and is truly quite just like classification except that any attribute, or combination of attributes (and not simply the class), may be foreseen in association. As association doesn't need the pre-labeling of categories, it's a sort of unsupervised learning.

b. Apriori Algorithm

.The rule for frequent item set mining and association rule learning over dealings databases. It followed by characteristic the frequent individual things within the info and increasing them to larger and bigger item sets as long as those item sets seem sufficiently typically within the info. The frequent item sets verified by Apriori may be wont to determine association rules that highlight general trends within the info.

ASSOCIATION RULE MINING

A. Association rule mining is defined as:

Let **I**= { …} be a set of '*n'* binary attributes called items.

Let **D**= { ….} be set of transaction called database. Every transaction in **D** has a distinctive transaction ID and contains a subset of the items in **I**. a rule is defined as implication of the form **X→Y** where **X,**

Y⊆ I and **X∩Y=Φ.** The set of items X and Y are called antecedent and consequent of the rule respectively.

B. **Useful Terms**

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best known constraints are minimum thresholds on support and confidence.

a) Support:

The support **supp(X)** of an item set **X** can be defined as proportion of transactions in the data set which contain the item set.

Supp(X) = no. of transactions which contain the item set **'X'** / total no. of transactions

b) Confidence:

The confidence of a rule is defined as:

Conf (**X→Y**) = supp (**XUY**)/supp(**X**)

1. MAP-REDUCE

MapReduce could be a process technique and a program model for distributed computing supported java. The MapReduce algorithmic rule contains two vital tasks, specifically Map and cut back. Map takes a group of information and converts it into another set of information, wherever individual components are countermined into tuples (key/value pairs). Secondly, cut back task, that takes the output from a map as associate degree input and combines those knowledge tuples into a smaller set of tuples. because the sequence of the name MapReduce implies, the cut back task is usually performed when the map job.

The major advantage of MapReduce is that it's simple to scale processing over multiple computing nodes. below the MapReduce model, the information process primitives are known as mappers and reducers. Moldering a knowledge process application into mappers and reducers is typically nontrivial. But, once we tend to write associate degree application within the MapReduce kind, scaling the applying to run over tons of, thousands, or maybe tens of thousands of machines in an exceedingly cluster is just a configuration amendment. this easy quantifiability is what has attracted several programmers to use the MapReduce model.

The MapReduce framework views the input to job as a <key, value> pair and produces a intermediate set of <key,

value> pairs as shown in fig 4. These pairs are then shuffled across different reduce tasks based on {key, value} pairs. Each Reduce task accepts only one key at a time and process data for the key and outputs the results as {key, value} pairs. The job submitted by user is then received by Jobtracker and breaks it into number of map and reduce tasks. It then assigns task to Tasktracker, monitors the execution of job and when job is completed informs to the user. As in Hadoop all the jobs have to share commodity servers in cluster for processing the data, proper scheduling policy and algorithms are required.

III PROBLEM STATEMENT

Traditional pattern mining algorithms don't seem to be appropriate for truly big data, presenting two main challenges to be solved: computational complexity and main memory necessities. a series of algorithms supported the MapReduce framework and the Hadoop open-source implementation are projected.

The pattern mining is one of the important tasks to extract meaningful and useful information from raw data. This task aims to extract item-sets that represent any type of homogeneity and regularity in data. Traditional pattern mining algorithms are not suitable for truly big data, presenting two main challenges to be solved: computational complexity and main memory requirements. a series of algorithms based on the MapReduce framework and the Hadoop open-source implementation have been proposed.

IV. CONCLUSION

In this project, projected new efficient pattern mining algorithms to figure in big data. All the projected models are supported the well-known Apriori algorithm and also the MapReduce framework. The projected algorithms are divided into three main groups [5].

1. No pruning strategy. Two algorithms (AprioriMR and IAprioriMR) for mining any existing pattern in data have been projected.

2. Pruning the search space by suggests that of anti-monotone property. Two further algorithms (SPAprioriMR and TopAprioriMR) are projected with the aim of discovering any frequent pattern offered in data.

3. Maximal frequent patterns. A final algorithm (MaxAprioriMR) has been conjointly projected for mining condensed representations of frequent patterns.

REFERENCES

1. Dang Nguyen, Bay Vo, "Efficient Mining of Class Association Rules with the item set Constraint", publication at: https://www.researchgate.net/publication/260677503, January 2015.
2. Arthur.A.Shaw, N.P. Gopalan, "Frequent Pattern Mining of Trajectory Coordinates using Apriori Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 22– No.9, May 2011.
3. Laila A. Abd-Elmegid, Mohamed E. El-Sharkawi, Laila M. El-Fangary & Yehia K. Helmy "Vertical Mining of Frequent Patterns from Uncertain Data", Computer and Information Science, Vol. 3, No. 2; May 2010.
4. Lakshminarayanan, " Frequent pattern mining on big data using Apriori algorithm", International Journal of Advance Research and Development (Volume3, Issue5) Available online at: www.ijarnd.com
5. J. M. Luna, J. R. Romero, C. Romero, and S. Ventura, "On the use of genetic programming for mining comprehensible rules in subgroup discovery," *IEEE Trans. Cybern.*, vol. 44,

no. 12, pp. 2329–2341, Dec. 2014. [Online]. Available: http://dx.doi.org/10.1109/TCYB.2014.2306819

6. R. Agrawal, T. Imielinski, and A. Swami, "Database mining: A performance perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 5, no. 6, pp. 914–925, Dec. 1993.
7. J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *Data Min. Know. Disc.*, vol. 8, no. 1, pp. 53–87, 2004
8. S. Zhang, Z. Du, and J. T. L. Wang, "New techniques for mining frequent patterns in unordered trees," *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1113–1125, Jun. 2015. [Online]. Available: http://dx.doi.org/10.1109/TCYB.2014.2345579
9. Mrs. A. NANDHINI, "Apriori Versions Based on Map Reduce for Mining Frequent Patterns on Big Data" IJRREM Volume -2, Issue -6,June -2018.