International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 7, July 2020 ||

Mining Frequent Pattern on Big Data Using Map Reducing Technique

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

PG Student, Department of Computer Engineering, ICOER, Wagholi, Pune, India[1]

Assistant Professor, Department of Computer Engineering, ICOER, Wagholi, Pune, India[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.

II.LITERATURE SURVEY

A. Title: Efficient Mining of Class Association Rules with the item set Constraint.
Author: Dang Nguyen, Bay Vo.
Year:2015
Mining class association rules (CARs) with the itemset constraint is concerned with the discovery of rules, which contain a set of specific items in the rule antecedent and a class label in the rule consequent. This task is commonly encountered in mining medical data. For example, when classifying which section of the population is at high risk for the HIV infection, epidemiologists often concentrate on rules which include demographic information such as gender, age, and marital status

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 7, July 2020 ||

in the rule antecedent, and HIV-Positive in the rule consequent. There are two naive strategies to solve this problem, 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. Therefore, such approaches are time consuming. This study proposes an efficient method for mining CARs with the itemset constraint based on a lattice structure and the difference between two sets of object identifiers (diffset)[1]

B. Frequent Pattern Mining of Trajectory Coordinates using Apriori Algorithm.
Author:Arthur.A.Shaw, N.P. Gopalan.
Year:2011
Proposes the Apriori Algorithm based frequent trajectory pattern mining algorithm to efficiently and effectively handle the trajectory database transaction. Prior to that the trajectory dataset is extracted from a text file and is imported to a Oracle database after doing the initial data cleaning process. Initial frequency count is done in Oracle database using its programming feature. Then the data is written in the operating system then further processing is done to find the frequent trajectory pattern. Advantage of this method is later iterations are much faster than the initial iterations of the algorithm. The results obtained by this method are more accurate and reliable. This algorithm uses large coordinate set property. Each iteration in this algorithm can be parallelized so that execution time can be reduced. More over this algorithm is easy to implement. Disadvantage of this method are, it uses a generate, prune and test approach generates candidate coordinate sets (1-coordinate, 2- coordinate, 3-coordinate,…), to check the generated sequence of coordinates are already generated or not, and tests if they are frequent by scanning the database and counting their support each time. Generation of candidate coordinate sets is expensive (in both space and time). Since generation and pruning steps are in memory resident, it needs more RAM. Another disadvantage is it needs n+1 database scans, n is the length of the coordinates in the longest pattern.[2]

C. Vertical Mining of Frequent Patterns from Uncertain Data
Author:Laila A. Abd-Elmegid, Mohamed E. El-Sharkawi, Laila M. El-Fangary & Yehia K. Helmy
Year:2010
Most existing algorithms mine frequent patterns from traditional transaction databases that contain precise data. In these databases, users definitely know whether an item (or an event) is present in, or is absent from, a transaction in the databases. However, there are many real-life situations in which one needs to deal with uncertain data. In such data users are uncertain about the presence or absence of some items or events. For example, a physician may highly suspect (but cannot guarantee) that a patient suffers from a specific disease. The uncertainty of such suspicion can be expressed in terms of existential probability. Since there are many real-life situations in which data are uncertain, efficient algorithms for mining uncertain data are in demand. Two algorithms have been proposed for mining frequent patterns from uncertain data. The previous two algorithms follow the horizontal data representation. In this paper we studied the problem of mining frequent itemsets from existential uncertain data using the Tidset vertical data representation. We introduced the U-Eclat algorithm, which is a modified version of the Eclat algorithm, to work on such datasets. A performance study is conducted to highlight the efficiency of the proposed algorithm also a comparative study between the proposed algorithm and the well known algorithm UF-growth is conducted and showed that the proposed algorithm outperforms the UF-growth.[3]

D. Frequent pattern mining on big data using Apriori algorithm
Author:Lakshminarayanan
Year:2018
New efficient pattern mining algorithms to work in big data. All the proposed models are based on the well-known Apriori algorithm. This algorithm has been also proposed for mixing condensed representations of frequent patterns. Pruning the search space by means of anti-monotone property. Two additional algorithms have been proposed with the aim of discovering any frequent pattern available in data. In Future, We will use the Top – K Ranking Algorithm to find the top k frequent patterns from the given dataset. 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]

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 7, July 2020 ||

III.EXISTING SYSTEM

Conventional cluster ensemble approaches have several limitations: They do not consider how to make use of prior knowledge given by experts, which is represented by pair wise constraints. Pair shrewd requirements are frequently characterized as the must-interface imperatives and the can't connect limitations. The must-connect requirement implies that two component vectors ought to be doled out to a similar bunch, while the can't interface limitations implies that two element vectors can't be allotted to a similar group. The majority of the bunch outfit strategies can't accomplish acceptable outcomes on high dimensional datasets. Not all the gathering individuals add to the outcome.

IV.PROBLEM STATEMENT

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.PROPOSED SYSTEM

The process starts with new efficient pattern mining algorithms to work in big data. All of them rely on the MapReduce framework and the Hadoop open-source implementation[2]. Two of these algorithms (AprioriMR and IAprioriMR) enable any existing pattern to be discovered. Two additional algorithms (SPAprioriMR and TopAprioriMR) use a pruning strategy for mining frequent patterns. Finally, an algorithm for mining MaxAprioriMR is also proposed.

V.SYSTEM ARCHITECTURE

The process starts with new efficient pattern mining algorithms to work in big data. All of them rely on the MapReduce framework and the Hadoop open-source implementation[2]. Two of these algorithms (AprioriMR and IAprioriMR) enable any existing pattern to be discovered. Two additional algorithms (SPAprioriMR and TopAprioriMR) use a pruning strategy for mining frequent patterns. Finally, an algorithm for mining MaxAprioriMR is also proposed.

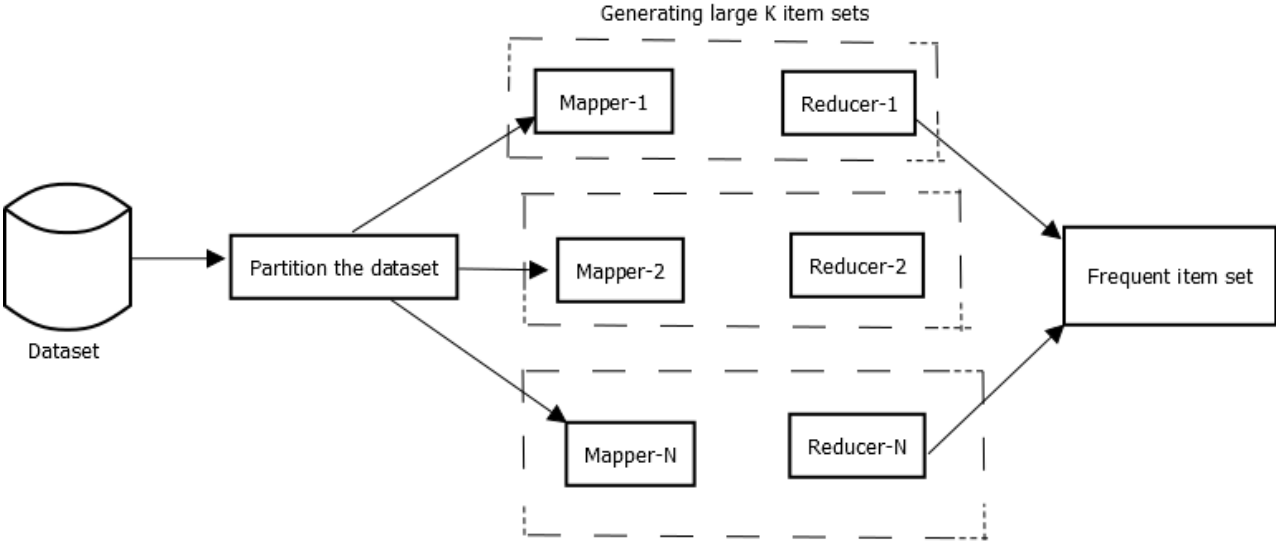

Fig. 1: System architecture

VI.METHODOLOGY

1. Pre-Evaluation of Min_Util: Though TKU provides a way to mine top-k HUIs, min_util is set to 0 before the construction of the UP-Tree. This results in the construction of a full UP-Tree in memory, which degrades the performance

of the mining task. If min_util could be raised before the construction of the UP-Tree and prune more unpromising items in transactions, the number of nodes maintained in memory could be reduced and the mining algorithm could achieve better performance. Based on this idea, we propose a strategy named PE (Preevaluation Step) to raise min_util during the first scan of the database.

2. Construction Of UP-Tree: A UP-Tree can be constructed by scanning the original database twice. In the first scan, the transaction utility of each transaction and TWU of each item are computed. During the second database scan, transactions are reorganized and then inserted into the UP-Tree.

3. Generating PTKHUIs: TKU algorithm uses UP tree to generate the potential top k high utility item sets. Parallelly, it raises the value of minimum utility threshold dynamically during the generation of PTKHUIs.

4. Identifying Top-K: HUIs From PTKHUIs After identifying PTKHUIs, TKU calculates the exact utility of PTKHUIs by scanning the original database once again, to identify the top-k HUIs.

5. Construction of Utility List Structure: In the TKO algorithm, each item (set) is associated with a utility-list. The utilitylists of items are called initial utility-lists, which can be constructed by scanning thedatabase twice. In the first database scan, the TWU and utility values of items are calculated. During the second database scan, items in each transaction are sorted in order of TWU values and the utility-list of each item is constructed.

6. Finding Top-K HUIs: In the TKO algorithm, initially, a list of top-k high utility item sets is generated .Parallelly,it raises the value of minimum utility threshold dynamically and updates the list of top-k high utility item sets.

VII.RESULTS

Fig. 2: Select Dataset

Fig. 3: IaprioriMR

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 7, July 2020 ||

Fig. 4: SPAprioriMR

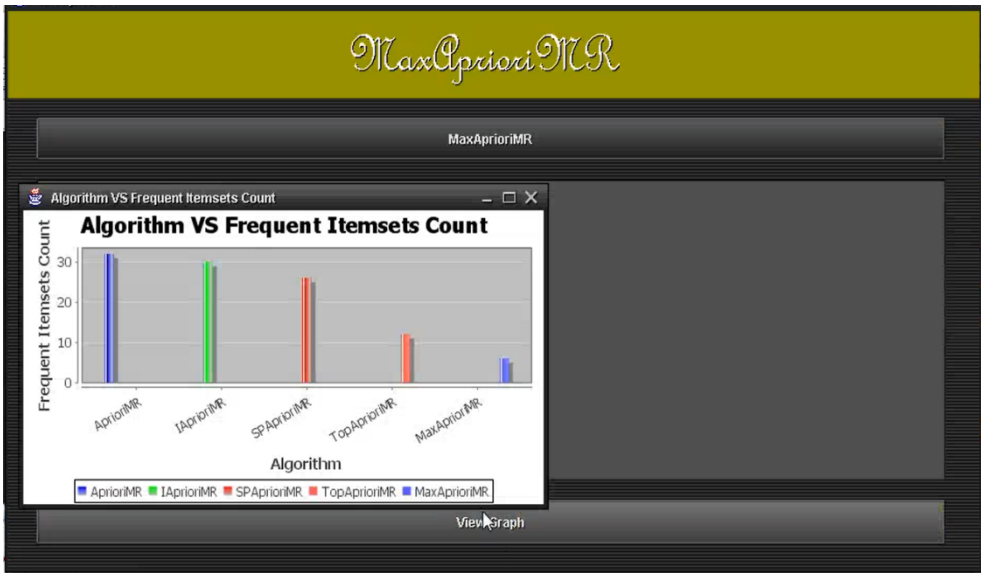

Fig. 5: Algorithm Vs Frequent Itemsets Count

VIII.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.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 7, July 2020 ||

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.