OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

MINING FREQUENT PATTERN ON BIG DATA USING MAP REDUCING TECHNIQUE

Miss. Swathi Kalaskar[1] Prof. Vandana Navale[2]

PG Students Department of Computer Engineering, Dhole Patil College of Engineering, Wagholi, Near Eon IT park., Pune-412207, Maharashtra, India[1]

Assistant Professor, Department of Computer Engineering, Dhole Patil College of Engineering, Wagholi, Near Eon IT park., Pune-412207, Maharashtra, India[2]

swati03.kalas@gmail.com

Abstract: The proposed models are based on the well-known Apriori algorithm and the MapReduce framework. The proposed algorithms are divided into three main groups. Two algorithms Apriori MapReduce (AprioriMR) and iterative AprioriMR (IAprioriMR) are properly designed to extract patterns in large datasets. These algorithms extract any existing item-set in data regardless their frequency[2]. Pruning the search space by means of the antimonotone property. Two additional algorithms space pruning AprioriMR (SPAprioriMR) and top AprioriMR (TopAprioriMR) are proposed with the aim of discovering any frequent pattern available in data. Maximal frequent patterns[5]. A last algorithm maximal AprioriMR (MaxAprioriMR) is also proposed for mining condensed representations of frequent patterns [3], i.e., frequent patterns with no frequent supersets.

Keywords: Big Data, Hadoop, Data Mining

----------∴∴∴----------

I INTRODUCTION

The motivation of this project is many efficient algorithms have been developed in this regard, the growing interest in data has caused the performance of existing pattern mining techniques to be dropped[1]. The goal of this project is to propose new efficient pattern mining algorithms to work in big data. To this aim, a series of algorithms based on 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 proposed with the aim of discovering any frequent pattern available in data

II LITERATURE SURVEY

In this Paper, Mining class association rules (CARs) with the item set 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 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]

In this paper, 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]

In this paper 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]

In this paper, we have proposed 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 precomputed and stored with the dataset.[4]

III SYSTEM ARCHITECTURE

In this system propose 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.

Utility Pattern Mining:

The task of top-k high utility pattern mining was introduced by Chan et al. But the definition of high utility itemset used in their study is different from the one used in this project. Chan et al.'s study has considered utilities of various items, but quantitative values of items in transactions were not taken into consideration. We have defined the task of top-k high utility itemset mining by considering both quantities and profits of items.

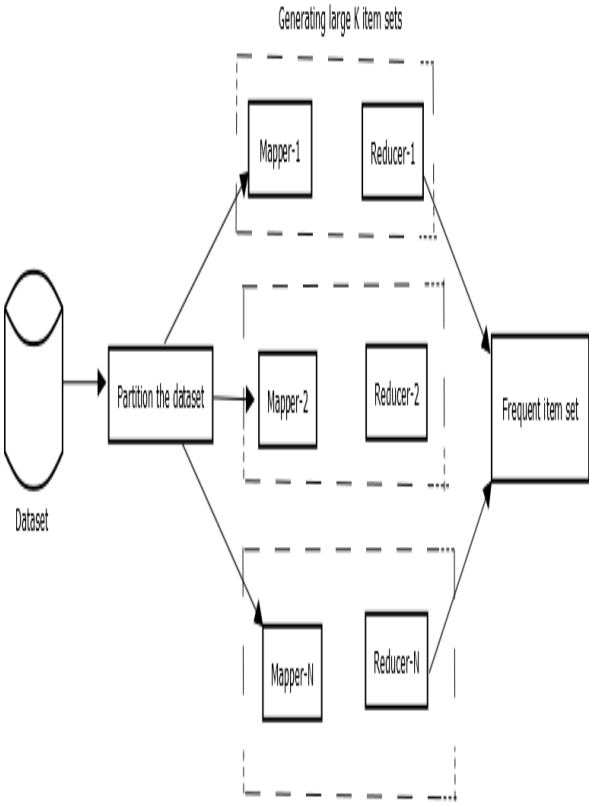

Figure 1: System architecture

IV RESULTS

Figure 2: Select Dataset

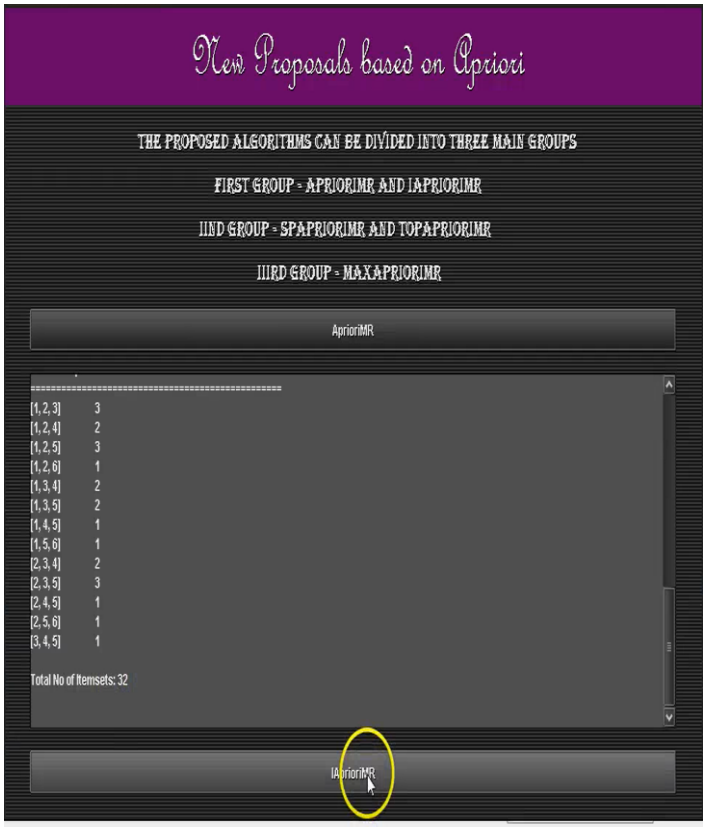

Figure 3: IaprioriMR

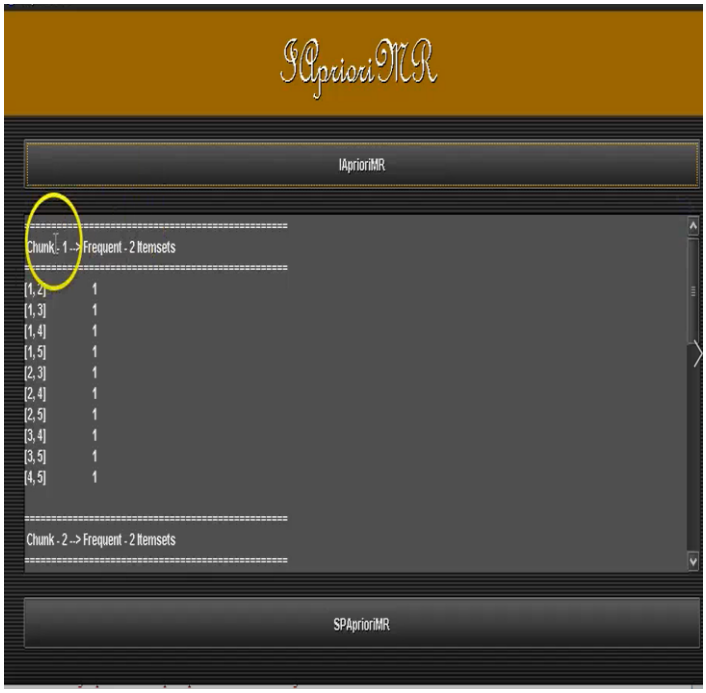

Figure 4: SPAprioriMR

Figure 5: Algorithm Vs Frequent Itemsets Count

V CONCLUSION AND FUTURE WORK

In this project, proposed new efficient pattern mining algorithms to work in big data. All the proposed models are based on the well-known Apriori algorithm and the MapReduce framework. The proposed algorithms are divided into three main groups[7].

1. No pruning strategy. Two algorithms (AprioriMR and IAprioriMR) for mining any existing pattern in data have been proposed.
2. Pruning the search space by means of anti-monotone property. Two additional algorithms (SPAprioriMR and TopAprioriMR) have been proposed with the aim of discovering any frequent pattern available in data.
3. Maximal frequent patterns. A last algorithm (MaxAprioriMR) has been also proposed for mining condensed representations of frequent patterns.

REFERENCES

1. Arthur.A.Shaw, " Frequent Pattern Mining of Trajectory Coordinates using Apriori Algorithm "
2. Laila A. Abd-Elmegid, " Vertical Mining of Frequent Patterns from Uncertain Data "
3. Lakshminarayanan, " Frequent pattern mining on big data using Apriori algorithm"
4. Dang Nguyen, Loan T.T, " Efficient Mining of Class Association Rules with the item set Constraint.
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