

# Association Rule Mining for Large Dataset Using Map Reduce

Rahul S. Roy

Student, CSE

SVP CET

Nagpur, India

e-mail: srroy.oo7@gmail.com

**Abstract**— Traditional Association Rules rule has computing power shortage in coping with massive datasets. So as to beat these issues, a distributed association rules rule supported Map-Reduce programming .Mining association rules is a crucial task. Past dealings information will be analyzed to get client buying behaviours such the standard of business call will be improved. The association rules describe the associations among things within the massive info of client transactions. However, the scale of the info will be terribly massive. it's terribly time overwhelming to seek out all the association rules from an oversized info and users is also solely curious about the associations among some things. Hence, an information mining has to be provided such users will question solely attention-grabbing knowledge to them from an oversized info of client transactions. During this paper, a knowledge mining language is conferred. From the information mining language, users will specify the interested things and also the criteria of the principles to be discovered. Also, Associate in nursing economical data processing technique is projected to extract the association rules per the user's requests.

**Keywords**- data mining, association rule mining, Map-reduce, Large Dataset.

\*\*\*\*\*

## I. INTRODUCTION

Information mining has high materialness in retail industry. The powerful administration of business is altogether subject to the nature of its choice making. It is along these lines vital to break down past exchange information to find client buying practices and enhance the nature of business choice. Since the measure of these exchange information is expansive, a productive calculation needs to be concocted for finding helpful data inserted in the exchange information. Affiliation standard mining is a method to distinguish the shrouded realities in extensive dataset and draw obstructions on how subsets of things impact the vicinity of different subsets. Affiliation standard mining plans to discover solid connection between qualities. All regular summed up examples are not exceptionally effective in light of the fact that a bit of the incessant examples is repetitive in the affiliation standard mining. This is the reason this calculation creates some excess manage alongside the fascinating principle. This disadvantage can be overcome with the assistance of min-max calculation. Since a large portion of the information mining methodologies utilizes the avaricious calculation rather than min-max calculation. Min-max calculation is to a degree best as contrast with the eager calculation on the grounds that it performs a worldwide inquiry and adapts better to the trait connection. In min-max calculation populace development is reproduced. Min-max calculation is an organic system which utilizes chromosome as a component on which arrangements (people) are controlled. Related work

## II. EASE OF USE

Following method plays an important role in our project work to analyse data from different dataset to improve association rule.

### 2.1. Efficient Data Mining Algorithm

In this section, we describe how to process a user's request. We develop an efficient data mining (EDM) algorithm to

generate the interesting association rules according to the user's request. For a user's request, if both the two keywords **Antecedent** and **Consequent** are specified in the **With** clause and there is no notation "\*" specified, then the antecedent and the consequent of the discovered rule will contain only the items specified in < Items >'s after the keywords **Antecedent** and **Consequent**, respectively. We call this type of users' requests the *Type I* request. If the user likes to extract association rules whose antecedent or consequent can contain other items except the items specified in < Items >, then the notation "\*" has to be specified in the **With** clause. We call this type of users' requests the *Type II* request. The request in which only one of the two keywords **Antecedent** and **Consequent** is specified also belongs to the *Type II* request. In this phase, EDM algorithm scans the database to record related information for each *interested item* and find large items. The interested items for the *Type I* request are the items specified in the **With** clause. The interested items for the *Type II* and *Type II1* requests are all items in the database.

The second phase is the *association graph construction* phase which constructs an association graph to indicate the associations between every two large items generated in the first phase. The third phase is the *interesting large itemset generation phase* which generates all interesting large itemsets by traversing the constructed association graph according to the user's request. The final phase is the *interesting association rule generation phase* which generates all interesting association rules according to the discovered interesting large itemsets, the items specified after the two keywords **Antecedent** and **Consequent**, and the user-specified minimum confidence in the user's request.

## III. RELATED WORK

Previously, numerous authors have planned completely different algorithmic program and techniques to supply economical association rule mining. and conjointly scale back multi scan downside with completely different

datasets. Association rule mining aims to extract attention-grabbing correlations, frequent patterns, associations or casual structures among sets of things within the dealings databases or other knowledge repositories. Frequent absence and presence itemset for negative association rule mining deals with Pattern from negative association rules thought-about to be distinctive and surprising compared to positive rules.

Above mentioned all previous papers have approaches generate an oversized dealings result as they propose a further novel technique for improvement of association rule mining. Above mentioned all previous papers have techniques/algorithm/approaches generate an oversized dealing result as they propose an extra novel technique for optimisation of association rule mining.

With the explosive recent growth of the amount of data, we are moving from the terabytes era to petabytes era. This trend Creates new demand for advancement in data storing and analytical technology. Hence there is a growing need to run data mining algorithm on massive datasets. Cloud computing is the development of distributed computing, parallel processing, and grid computing, which represents an emerging business computing model. Computing tasks are distributed by the cloud computing in a resource pool composed of a large number of computers. Through the cloud computing, various application systems can obtain computing capability, storage space and a variety of software services as required. The novelty of cloud computing is that it can provide almost unlimited cheap storage and computing ability. With emerging trends in Cloud Computing, massive data storage and data mining is a research field with a very theoretical and practical value. We can use cloud computing techniques with data mining to reach high capacity and high efficiency[1].

Mining strong valid Association Rule form Frequent Pattern and Infrequent Pattern Based on Min-Max Sinc Constraints” deals with association rule mining based on min-max algorithm and MLMS formula. The process of rule optimization we used min-max algorithm and for evaluate algorithm conducted the real world dataset such as heart disease data and some standard database repository. Rule mining is very efficient technique for find relation of correlated data. The correlation of data gives meaning full extraction process. For the mining of rule mining a variety of algorithm are used such as Apriori algorithm and tree based algorithm. Some algorithm is wonder performance but generate negative association rule and also suffered from multi-scan problem. In this paper we proposed IMLMS-PANR-GA association rule mining based on min-max algorithm and MLMS formula. In this method we used a multilevel multiple support of data table as 0 and 1. The divided process reduces the scanning time of database. The proposed algorithm is a combination of MLMS and min-max algorithm. Support length key is a vector value given by the transaction data set. The process of rule optimization we used min-max algorithm and for evaluate algorithm conducted the real world dataset such as heart disease data and some standard data used from UCI machine learning repository[4].

A novel approach for efficient mining and hiding of sensitive association rule” proposed a system, the efficiency of mining association rules and confidentiality of association rule is becoming one of important area of knowledge discovery in database. Main goal of this paper is reduce unnecessary database scan for that author used novel approach to decrease support and confidence. Data mining is the process of analyzing large database to find useful patterns. The term pattern refers to the items which are frequently occurring in set of transaction. The frequent patterns are used to find association between sets of item. The efficiency of mining association rules and confidentiality of association rule is becoming one of important area of knowledge discovery in database. This paper is organized into two sections. In first part of paper an Improved Apriori algorithm is being presented that efficiently generates association rules. These reduces unnecessary database scan at time of forming frequent large itemsets. In second part of this paper we have tried to give contribution to improved apriori algorithm by hiding sensitive association rules which are generated by applying improved Apriori algorithm on supermarket database. In this paper we have used novel approach that strategically modifies few transactions in transaction database to decrease support and confidence of sensitive rule without producing any side effects. Thus in the paper we have efficiently generated frequent itemset sets by applying Improved Apriori algorithm and generated association rules by applying minimum support and minimum confidence and then we went one step further to identify sensitive rules and tried to hide them without any side effects to maintain integrity of data without generating spurious rules [3].

With the explosive recent growth of the amount of data, we are moving from the terabytes era to pet bytes era. This trend creates new demand for advancement in data storing and analytical technology. Hence there is a growing need to run data mining algorithm on massive datasets. Cloud computing is the development of distributed computing, parallel processing, and grid computing, which represents an emerging business computing model. Computing tasks are distributed by the cloud computing in a resource pool composed of a large number of computers. Through the cloud computing, various application systems can obtain computing capability, storage space and a variety of software services as required. The novelty of cloud computing is that it can provide almost unlimited cheap storage and computing ability. With emerging trends in Cloud Computing, massive data storage and data mining is a research field with a very theoretical and practical value. We can use cloud computing techniques with data mining to reach high capacity and high efficiency [5].

Relevant association rule mining from medical dataset using new irrelevant rule elimination technique” proposes a technique the n-cross validation system to reduce association rules which are irrelevant to the transaction of medical dataset. Association rule mining (ARM) is an emerging research in data mining. It extracts interesting association or correlation relationship in the large volume of transactions. Apriori based algorithms have two steps. First step is to find the frequent item set from the transactions. Second step is to construct the association rule. If ARM applied with medical dataset, it

produces huge quantity of rules; most of these rules are irrelevant to the transaction. These irrelevant rules consume more memory space and misguide the decision making. Here irrelevant rule reduction is important. This paper proposes the n-cross validation technique to reduce association rules which are irrelevant to the transaction set. The proposed approach used partition based approaches are supported to association rule validation. The proposed algorithm called as PVARM (Partition based Validation for Association Rule Mining). The proposed PVARM algorithm is tested with T40I10D100K and heart disease prediction. The performance analysis attempted with Apriori, most frequent rule mining algorithm and non redundant rule mining algorithm to study the efficiency of proposed PVARM. The proposed work reduces large number of irrelevant rules and produces new set of rules with high confidence. It is much use to mine medical relevant rule mining [5].

Frequent absence and presence item set for negative association rule mining” deals with Pattern from negative association rules are considered to be unique and unexpected compared to positive rules. Negative association rule (NAR) mining has created more attention recently due to the knowledge and discovery of the interestingness of the pattern of the negative association rules and the challenges during the mining process. Pattern from negative association rules are considered to be unique and unexpected compared to positive rules. Negative association rules are useful in analysis for decision making in identifying the items which conflict with each other or the items that complement each other. However, negative association rules mining still have their own issues such as mining space and good quality of negative association rules. In this paper, we provide the preliminaries of basic concepts of negative association rule. We proposed an enhancement in Apriori algorithm for mining negative association rule from frequent absence and presence (FAP) item set. Prominent literature will be discussed to further understand negative association rule mining employing a dataset which are collected from users within the real-world to evaluate approaches and to locate some attention-grabbing results [6].

#### IV. PROPOSED METHOD

In our system we tend to analyse completely different datasets with respects to making different sorts of new association rule.

in initial module we tend to square measure aiming to build dealing generation or building half to boot we tend to square measure providing index based mostly search. In second section we tend to square measure showing all dealing performed early at the moment we tend to get distinct things from all dealing so choice of things and realize completely different association rule and last realize connected items show results. One extra schema we tend to additional during this system like dynamic dataset analyse with various economical association rule.

#### V CONCLUSION

Author propose associate degree economical data processing Technique for locating fascinating association rules. economical methoding algorithmic rule (EDM) to process a user's request. The algorithmic rule EDM wants only 1 information scan and a few inner merchandise to get all fascinating association rules consistent with the user's request, that is extremely economical. within the future, we have a tendency to shall extend the information mining language to permit additional versatile question specifications, associate degree develop associate degree interactive data processing technique to find other forms of association rules consistent with the user's request, like generalized association rules and multiple-level association rules.

#### References

- [1]. S. P. Patil “A novel approach for efficient mining and hiding of sensitive association rule” 978-1-4673-1720-7 © 2012 IEEE Transaction
- [2]. Xueyan Lin "MR-Apriori: Association Rules Algorithm Based on MapReduce" 978-1-4799-3279-5 /14/\$31.00 ©2014 IEEE Transaction
- [3]. K. Rameshkumar “Relevant association rule mining from medical dataset using new irrelevant rule elimination technique” 978-1-4673-5786-9©2013 IEEE Transaction.
- [4]. Mr. A.S.A Kadir “Frequent absence and presence itemset for negative association rule mining” 978-1-4577-1676-8©2011 IEEE Transaction.
- [5]. Mr. Mukesh Poundekar et al “Mining strong valid Association Rule form Frequent Pattern and Infrequent Pattern Based on Min-Max Sinc Constraints” 978-1-4799-3070-8 © 2014 IEEE Transaction.