AUTO RECOMMENDATION ENGINE USING HADOOP

[1]Mandar Puranik, [2]Sunetra Kanade, [3]Kalyani Raut, [4]Nikita Dange,
[5]Prof. Nishigandha Wankhade
[1,2,3,4,5]Department of Computer Engineering, Marathwada Mitra Mandal Institute of Technology, Pune, India.
Email: [1]mandarmpuranik@gmail.com, [2]sunetra.kanade13@gmail.com,
[3]rautkalyani1992@gmail.com, [4]nikita.dange45@gmail.com, [5]n.wankhade@mmit.edu.in .

Abstract— The Prototypical application of Hadoop is using Apriori algorithm. However, Apriori is efficient to generate and mine association rules form larger database or repository. Apriori algorithm generates interesting frequent or infrequent candidate itemsets with respect to support and confidence count. Apriori algorithm requires to produce vast number of candidate sets and to generate these, it needs several scans over database. While it requires more memory space and scans for candidate generation process we will improve Pruning strategy as it will reduce the scans required to generate frequent item sets and accordingly and a valence or weightage to strong association rule. And hence Apriori will get more effective and efficient. Previously various improved versions of apriori algorithm were used. This paper actually presents an idea of improved version of Apriori algorithm along with Association Rules and their performance comparison with Apriori algorithm..

***Index Terms*— Apriori algorithm, frequent itemsets, data mining, hadoop, candidate itemset, support and confidence.**

I. INTRODUCTION

THE Apriori algorithm is a seminal algorithm proposed by R. Agarwal and R. Shrikant. Many researches were done in order to make mining frequently used itemsets scalable and efficient.

Basically the name of apriori based on the algorithm uses prior knowledge to find frequent itemset. Any subset of a large itemset must be large is its motto. Two major subsets of apriori algorithm are join and prune. The itemsets are being compared with a fixed minimum support and they are less than the minimum support are being deleted from L_k(set of k-itemset) and C_k,(Candidate k-itemset) a superset is formed. Apriori algorithm due to too many scans over database and generation of huge number of candidate set suffers many deficiencies. To overcome these deficiencies we use techniques like mapping, reducing in hadoop, transaction reduction, improved version of apriori algorithm is proposed.

Association Rule Mining (ARM) is technique used for data mining. ARM extracts interesting correlation, frequent patterns, association or casual structures among set of database or other repositories. Support and Confidence are two pillar criteria used by association rule. ARM needs to satisfy both a user specified min support and confidence at the same time.

II. PROBLEM STATEMENT

Compare different itemsets in Apriori algorithm and find how much efficiency they improve. Efficiency depends upon various factors like minimum support value accuracy, number of records.

III. UNITS

• BIG DATA(HADOOP)

As data sets continue to grow, the potential of a failure impacting a customer grows with it. As more business groups, applications, and jobs leverage the same infrastructure, a failure can have a dramatic impact on overall system performance. This concern falls in line with ESG research, which shows that reliability is top of mind for businesses considering a business intelligence, analytics, and big data solution. When it comes to Hadoop, customers want a highly available solution that has no single point of failure and can sustain multiple failures while self-healing where able to. Traditional Hadoop distributions cannot reliably support these demands, especially comprehensively across the entire Hadoop ecosystem.

With a no single point of failure architecture, Map Reduce can reliably and comprehensively minimize data loss during minor or major cluster failures, including support for projects ranging from Impala, Hive, and Oozie, to Storm, MySQL, and Kafka. When selecting a Hadoop platform to store, manage, and analyze big data, it is crucial to select a platform that can meet the scalability, high availability, reliability, and performance needs of a dynamic organization. Though numerous vendors and solutions can meet most of those requirements, almost all of them rely on add-on, customized point solutions, minimizing deployment flexibility and adding additional management tasks for an already busy IT administrator ground up, bringing together the innovative Map Reduce architecture with the comprehensive projects in the Hadoop ecosystem to deliver enterprise-grade features and functions on low-cost commodity hardware. The distributed Map Reduce architecture offers high levels of availability and reliability with no single points of failure, helping to drastically reduce the mean time to data loss .

• MAP REDUCE

The computation takes a set of input key/value pairs, and produces a set of output key/value pairs. The user of the MapReduce library expresses the computation as two functions: Map and Reduce. Map, written by the user, takes an input pair and produces a set of intermediate key/value pairs. The MapReduce library groups together all intermediate values associated with the same intermediate key I and passes them to the Reduce function.The Reduce function, also written by the user, accepts an intermediate key I and a set of values for that key. It merges together these values to form a possibly smaller set of values. Typically just zero or one output value is produced per Reduce invocation. The intermediate values are supplied to the user's reduce function via an iterator. This allows us to handle lists of values that are too large to fit in memory.

• ASSOCIATION RULES

The associative rule mining concept inherits the Map Reduce scalability to huge datasets and to thousands of processing nodes. For finding frequent item sets, it uses hybrid approach between miners that uses counting methods. Association rule mining finds interesting associations and correlation relationships among large set of data items. Association rules show attribute value conditions that occur frequently together in a given dataset. A typical and widely-used example of association rule mining is Market Basket Analysis. Association rules provide information in the form of "if-then" statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature. In addition to the antecedent (the "if" part) and the consequent (the "then" part), an association rule has two numbers that express the degree of uncertainty about the rule. In association analysis the antecedent and consequent are sets of items (called itemsets) that are disjoint. The first term is called the support for the rule. The Support is the number of transactions that include all items in the antecedent and consequent parts of the rule. The support is also expressed as a percentage of the total number of records in the database. The other term is known as the confidence of the rule. Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all items in the antecedent.

For example, if a supermarket database has 100,000 point-of-sale transactions, out of which 2,000 include both items A and B and 800 of these include item C. The association rule "If A and B are purchased then C is purchased on the same trip" has a support of 800 transactions (alternatively 0.8% = 800/100,000) and a confidence of 40%

(=800/2,000). One way to think of support is that it is the probability that a randomly selected transaction from the database will contain all items in the antecedent and the consequent, whereas the confidence is the conditional probability that a randomly selected transaction will include all the items in the consequent given that the transaction includes all the items in the antecedent.

Association rule mining has a wide range of applicability such market basket analysis, medical diagnosis/ research, website navigation analysis, homeland security, education, financial and business domain and so on. Association rule mining is easy to use and implement and can improve the profit of companies. The computational cost of association rule mining represents a disadvantage and future work will focus on reducing it.

• Steps for Association:

Association Rule Mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. Association rules are typically generated in stepwise process.

1. Minimum support is used to generate the set of all frequent itemsets for the data set. Frequent itemsets are itemsets which satisfy the minimum support constraint.

2. Then each frequent itemsets is used to generate all possible rules from it and all rules which do not satisfy the minimum confidence constraint are removed.

3. Analyzing this process, it is easy to see that in the worst case we will generate 2n - n - 1 frequent itemsets with more than two items from a database with n distinct items.

4. Each frequent itemset in the worst case generate at least two rules, we will end up with a set of rules in the order of O(2n).

5. Increasing minimum support is used to keep the number of association rules found at a manageable size. However, this also removes potentially interesting rules with less support. Therefore, the need to deal with large sets of association rules is unavoidable when applying association rule mining in a real setting.

IV. HELPFUL HINTS

A. Figures and Tables

• APRIORI ALGORITHM

Apriori algorithm is very famous algorithm used for finding the frequent items from the itemset. For finding the frequent itemset Apriori uses prior knowledge. It uses a property ‘Any subset of a large itemset must be large’ [2]. Data Processing is the very important area of research in the industry. There are many effective algorithms which are used for it. Apriori algorithm is used for the large amount of data. As the data set increases the apriori algorithm gives the more accurate results.

Apriori algorithm in carried out in two main steps: Join and Prune. In join the large itemsets sets are found and in prune these itemsets are compared with support values which is fixed and the less itemsets than that of minimum support are deleted. And then set candidate itemsets are formed.

For the generation of candidate set Apriori have to scan the database for too many times. To overcome this problem we are using map-reduce technique.

• Generation steps

Input items to the map reduce.

Map reduce will provide the large itemset required for the apriori.

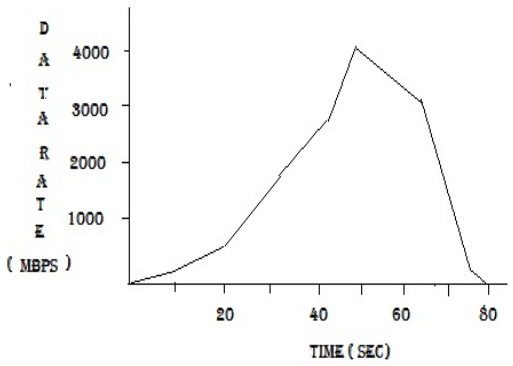
Apriori algorithm will compare each itemset with the support value.

If the value is greater than support then only it will be taken by the apriori and this process is performed till the last itemset.

The itemsets with value less than support value then it will be deleted.

Then the apriori will return the frequent itemset from the given input.

• Graph



B. Abbreviations and Acronyms

ARM means Association Rule Mining is the technique used for data mining.

L_k is the set of k-itemset and C_k, is the Candidate k-itemset.

sup stands for support and conf stands for confidence.

C. Equations

• Support

Support gives total number of transaction of any particular item are occurring in datasets. The rule holds with support sup in T (the transaction data set) if sup% of transactions contain X Y.

sup = Pr(X Y).

• Confidence Graph

Confidence gives strength of a data in a dataset. The rule holds in T with confidence conf if conf% of transactions that contain X also contain Y.

conf = Pr(Y | X)

An association rule is a pattern that states when X occurs, Y occurs with certain probability.

• Support Count:

The support count of an itemset X, denoted by X.count, in a data set T is the number of transactions in T that contain X. Assume T has n transactions. Then,

$$\text{support} = \frac{(X \cup Y) \,.\, \text{count}}{n}$$

$$\text{confidence} = \frac{(X \cup Y) \,.\, \text{count}}{X.\text{count}}$$

V. CONCLUSION

This Paper helps to understand the enormous variation of Apriori algorithm. Apriori algorithm is beneficial for finding the subsets of frequent itemset. Still Apriori has drawbacks of more time, space for candidate generation process .Apriori algorithm has a wide range of application like market and risk management, market basket analysis, inventory control, Telecommunication network, etc. Hence we conclude that the Apriori algorithm using map-reduce is the most efficient way to get the frequent data set from the large data set.

REFERENCES

[1] Margaret H. Dunham, Yongqiao Xiao Le Gruenwald, Zahid Hossain, "A Survey of Association Rules", Southern Methodist University, University of Oklahoma Dallas, Texas 75275-0122 Norman.

[2] Sana Irfan, "A Comparative Analysis of Improved Version of Apriori Algorithm of Data Mining," Meerut, Uttar Pradesh, India.

[3] Nilesh.S.Korde, Prof.Shailendra Shende, " Parallel Implementation of Apriori Algorithm" IOSR Journal of Computer Science (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727 PP 01-04.

[4] Jeffrey Dean and Sanjay Ghemawat "MapReduce: Simplified Data Processing on Large Clusters" .

[5] Bing Liu, Wynne Hsu, Yiming Ma, " Integrating Classification and Association Rule Mining ", KDD-98 Proceedings. Copyright © 1998, AAAI (www.aaai.org).

[6] Rakesh Agarwal, Ramkrishnan Srikant "Fast Algorithms for Mining Association rules", IEEE Trans. on Data Mining, vol. 4, pp. 570-578, July 2013.

[7] Rakesh Agrawal, Tomasz Imielinski, Arun Swami, " Mining Association Rules between Sets of Items in Large Databases".

[8] Irina Tudor, " Association Rule Mining as a Data Mining Technique", User's Guide and Reference Manual, Addison-Wesley 2008.

[9] Maur ice Houtsma, Arun Swami, " Set-oriented mining for association rules in relational databases".