Heuristic Approach to simplify Apriori Algorithm

Aman Jain[1], Poorvi Patni[2], Neeraj Kumar Verma[3]

[1]*Student, College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad*

[3]*Asistant Professor, College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad*

[1]*amanj0318@gmail.com*

[2]1998pj.cs@gmail.com

[3]*neerajk.computers@tmu.ac.in*

Abstract-**The word DATA means a set of values, fact and quantitative variables. Secondly the word mining is extraction of valuable material from the row data. So both word constitutes the term DATA MINING. It is all about mining of noble information from massive dataset. To do so APRIORI ALGORITHM is the simplest and most popular algorithm in data mining. Apriori is applied on frequent itemset using relevant association rule. It has a huge application in the market basket analysis.It is operated on the databases with simultaneous updation and transaction. This algorithm has many drawback like it is computationally very expensive. This paper deals with the apriori algorithm, and various techniques that were proposed to improve the apriori algorithm. HEURISTIC APPROACH is what we will be using to resolve this drawback.**

***Keywords*-Data Mining, Apriori Algorithm, Heuristic Approach, Association Rule.**

I. INTRODUCTION

Data mining is the process of extracting information from a large database. It is also treat as **knowledge discovery from data** (KDD).

Basically, it is the process of discovering the interesting patterns from database by using statistic, machine learning and database system approach for further use. Data mining is used as to make it feasible. In Data mining task, different type of data mining functionalities are used to specify the kind of patterns that can be frequent or interesting. Frequent Pattern, are the patterns that occurs frequently in data. It include frequent itemset, frequent structure and frequent sequences while interesting patterns validates the hypothesis that the user sought to confirm. These are easily understandable, useful and novel. For data mining different sources such as database, data warehouse, the web or data that are streamed into system dynamically.

In data mining, Apriori algorithm can be used to mine the frequent itemsets for Boolean association rules over transactional database. It was proposed by Agrawal and Srikant in 1994 to operate on database that contain transaction. It follow the bottom-up approach. To optimize space for storing immediate candidate various data structure have been designed but Hash table, Hash tree and Trie (prefix tree) are most eminent among them. In the sequential implementation of Apriori, trie performs better than hash tree. But hash table trie does not perform faster than trie. This algorithm have some drawback like if the database is small then it can find many false association that can be happened by chance. To address these issues we can evaluate obtained rules on the held out test data for the

support, lift and conviction values. In the process of Market Basket Analysis it computes some extra steps which can be reduced by using Heuristic Avenue.

II. RELATED WORK

We apply Apriori algorithm to uncover the hidden information and to analyze frequent itemsets. Through this paper we are highlighting use of association rule in extracting patterns in large dataset of data warehouse. Data set is taken from the central repository.

A. Association Rule

In data mining finding frequent pattern from a large dataset is called association. These Patterns can include itemsets, sequences and subsequences. A set of items that appear in a transactional dataset is called frequent itemset. for e.g. Pen and paper. It basically depicts interdependency of the data and related rules to be applied between those items. A rule is defined as an implication of the form X=>Y, where $X \cap Y \neq \infty$. It has two rules: antecedent (Left hand side rule) and consequent (Right hand side rule).

Support-It is the showcase for items appearing frequently in the data. The support for itemset $\mathbf{X_i}$ with respect to transaction T is given by:

$$\text{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

OR

$\text{supp}(X_i)$=count of X_i found together in data set

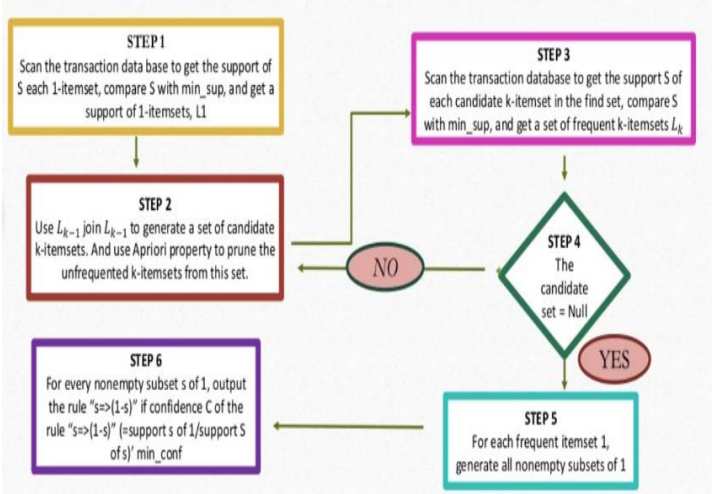
total number of transaction tuples

where i=1, 2,3,4,……N.

III. Apriori Algorithm

It is initiated by collecting frequent individual items as long as they appear often in database. It has application in many domains one such is **Market basket analysis.**

Steps to solve Apriori Algorithm



Algorithm:

K=total items in item set.

Input:D, a database of transaction; min_sup, the minimum support count threshold; **Output**: L, frequent item set in D.**Method**:
L1=find_frequent_1-itemsets(D);
for(k=2;$L_{k-1} \neq \infty$;k++){

C_k= apriori_gen (L_{k-1});
for each transaction t ϵ D{
C_t = subset(C_k,t);
for each candidate c ϵC_t
c.count++;
}
L_k= {cϵC_k|c.count$\geq$min_sup}
}
return L= $\cup_k L_k$; **procedure apriori_gen**(L_{k-1}:frequent(k-1)-itemsets)
for each itemset $l_1 \epsilon L_{k-1}$ for each itemset$l_2 \epsilon L_{k-1}$ { if($l_1[1]=l_2[1]$)$\wedge$($l_1[2]=l_2[2]$)$\wedge$ …

$\wedge(l_1[k-2]=l_2[k-2])\wedge(l_1[k-1]<l_2[k-1])$ then {

$c=l_1 \bowtie l_2$; // join step: generate candidates
if has_infrequent_subset(c,L_{k-1}) then
delete c;//prune step:
else add c to C_k;
}
return C_k;} **procedure has_infrequent_subset**(c: candidate k-itemset; L_{k-1}:frequent(k-1)-itemsets); //use prior knowledge
for each(k-1)-subset s of c
if(s $\notin L_{k-1}$) then
return **true;**
return **false;**

VI. Problem Statement

Let there be I_k elements in a data set where k is from 1,2,3….N. Apriori Algorithm checks the elements till N-1 items in itemset which increase the overhead of computation, to reduce this overhead we use maximum number of items in dataset(no of transaction) with respect to minimum support threshold of the transaction.

Methodology Used: (Heuristic Approach)

1. We will implement existing algorithm and find its weakness.
2. We will impose the enhancement in the existing algorithm.
3. Check the validity using real life dataset.

V.Heuristic Approach for modification of the existing classical Apriori Algorithm

It is the mental shortcut to reduce the complex problem using simple rules & make decision to choose the important aspects of problem and ignore the others .After applying heuristic approach we can optimize the problem of the Apriori algorithm and we can increase its efficiency to anoptimal level. Let's begin by putting up a question–What if we scan the itemset to the maximum value of all transaction?
This is something which is possibly supported by using heuristic approach– checking the elements till M items (maximum items involved in the transaction) rather checking it till N-1. We can modify the algorithm as-

K=total items in item set.

M=maximum no of items in a transaction of a dataset

N=total no of transaction

Input:
D, a database of transaction; min_sup= the minimum support count threshold;

Output:
L, frequent item set in D.**Method**:
L_1=find_frequent_1-itemsets(D);
M=1;

for(i=2;i<=N;i++){
z=no of item in itemset;
if(z>m)m=z;}

for(m=2;$L_{m-1}\neq\infty$;m++){

C_m= apriori_gen (L_{m-1});
for each transaction t ϵ D{
C_t = subset(C_m,t);
for each candidate c ϵC_t
c.count++;
}
L_m= {cϵC_m|c.count$\geq$min_sup}
}
return L= $\cup_m L_m$;
procedure apriori_gen(L_{m-1}:frequent(m-1)-itemsets)
for each itemset $l_1 \epsilon L_{m-1}$
for each

itemset$l_2 \epsilon L_{m-1}$ if($l_1[1]=l_2[1]$)∧($l_1[2]=l_2[2]$)∧ … ∧($l_1[m-2]=l_2[m-2]$)∧ ($l_1[m-1]<l_2[m-1]$) then {

$c=l_1 \bowtie l_2$; // join step: generate candidates
if has_infrequent_subset(c,L_{m-1}) then
delete c;//prune step:
else add c to C_m; }
return C_m;

procedure has_infrequent_subset(c: candidate k-itemset; L_{m-1}:frequent(m-1)-itemsets); //use prior knowledge
for each(m-1)-subset s of c
if(s ∉ L_{m-1}) then
return **true;**
return **false;**

IV. Conclusion

This paper is an attempt to work on the drawback of Apriori algorithm and improve it using heuristic approach. Data mining is used as a tool to find hidden patterns from large database and Apriori algorithm plays a very important role in finding the same. Apriori algorithm identifies customer behavior on the basis of frequently purchased item set using association rules .It has a wide application in market basket analysis .Heuristic approach is the way we had recover the problem of reduced efficiency and made the algorithm more efficient and fast.

VII.References

[1] https://en.wikipedia.org/wiki/Data_mining
[2] https://www.slideshare.net/INSOFE/apriori-algorithm-36045672
[3] https://ieeexplore.ieee.org/abstract/document
[4] https://en.wikipedia.org/wiki/Association_rule_learn
[5] https://www.google.com/search?client=firefox-b&ei=wkfgWpudMYjovgSO1rWQAg&q=support+rule+in+data+Alo&oq=support+rule+in+data+Alo&gs_l=psy-ab.3...410874.423300.0.423632.36.29.0.0.0.0.0.0..0.0....0...1c.1.64.psy-ab..36.0.0.0...0.D0ph7GM_QLg
[6] https://www.google.com/search?q=review+of+apriori+algorithm&ie=utf-8&oe=utf-8&client=firefox-b
[7] Jaiwei Han | Micheline Kamber | Jian Pei " DATA MINING Concepts and Techniques " 3rd Edition