Effective Determination and Improved Performance of Frequent Item Sets by APRIORI and PAFI Algorithm

V. Ramya
Assistant professor, Department of computer science, SRM UNIVERSITY

Abstract: Text mining refers to the process of procuring high-quality information from text. There are different technologies for text mining. One of the most successful will be mining using the effective patterns. In this paper we have explained one of the most powerful and successful algorithms of association mining named as Apriori algorithm. Lots of algorithms for mining association rules and their variations are proposed on basis of Apriori algorithm, but traditional algorithms are not efficient. Proposed algorithm improves Apriori algorithm by the way of a decrease of pruning operations, which generates item sets by the apriori operation.

Keywords: Apriori, Text mining, Association mining.

1. INTRODUCTION

Association Rule:
Association rule mining is fundamentally focused on finding frequent coincide associations with a collection of items. It is sometimes quoted as “Market Basket Analysis”, since that was the original application area of association mining. The goal is to find associations of items that result together more repeatedly than you would expect from a causal sampling of all possibilities.

Association analysis is useful for identifying interesting patterns clouded in large data sets. The following rule can be Separated from the data set shown in table1.

Table 1

<table>
<thead>
<tr>
<th>TID</th>
<th>ITEM SET</th>
</tr>
</thead>
<tbody>
<tr>
<td>T100</td>
<td>{Banana, Cake}</td>
</tr>
<tr>
<td>T200</td>
<td>{Pencil, Biscuit, Banana, Book}</td>
</tr>
<tr>
<td>T300</td>
<td>{Cake, Biscuit, Book, Pencil}</td>
</tr>
<tr>
<td>T400</td>
<td>{Biscuit, Cake, Banana, Pencil}</td>
</tr>
<tr>
<td>T500</td>
<td>{Book, Banana, Cake, Pencil}</td>
</tr>
</tbody>
</table>

The rule prefers that a strong relationship occur between the sale of Biscuit and Banana because many customers who buy Biscuits also buy Banana. Business persons can use this type of rules to help them to identify new scopes for cross selling their products to the buyers.

Apriori Algorithm:
The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties. Apriori employs an iterative approach known as a level wise search, where k-item sets are used to explore (k+1) itemsets.
To improve the efficiency of the level wise generation of frequent itemsets, an important property called the Apriori property is used to reduce the search space. The two step process is followed, consisting of join and prune actions.

**Improved APRIORI Algorithm:-**

**Join Step:** - $C_k$ is generated by joining $L_{k-1}$ with itself.

**Prune step:** - Any $(k-1)$ -itemset that is not frequent cannot be a subset of a frequent $k-1$ itemset

Where, $C_k$: candidate itemset of size $k$

$L_k$: frequent itemset of size $k$

$L_1$={frequent items};

For (k=1; $L_k$! =Ø;k++)

do begin $C_{k-1}$=

candidates generated from $L_k$;

For each transaction $t$ in database do Increment the count of all candidates in $C_{k-1}$ that are contained in $t$

$L_{k-1}$=candidates in $C_{k-1}$ with min_support end Return $U_kL_k$ [3]

**Frequent item set generation**

Two basic components for identifying frequent itemsets are **Support** and **Confidence**.

**Support** is an indication of how frequently the items appear in the database.

**Confidence** indicates the number of times the if/then statements have been found to be true.

Determining frequent itemsets from table1

<table>
<thead>
<tr>
<th>ID</th>
<th>Banana</th>
<th>Cake</th>
<th>Pencil</th>
<th>Biscuit</th>
<th>Book</th>
</tr>
</thead>
<tbody>
<tr>
<td>T100</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T200</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T300</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T400</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T500</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Illustration of Apriori principle:**

[Diagram of Apriori principle]
2. METHODOLOGY

Improving the efficiency of determining frequent item sets by partitioning algorithm.

Any item set that is potentially frequent in database must be frequent in at least one of the partitions of database.

The partition algorithm is based on the attention that the frequent sets are normally very few in number compared to all set of item sets. By using partitioning algorithm can be easily created, where each partitioning could be handled by a separate machine.

PAFI (Partitioning algorithm for frequent item sets):

This algorithm separates large data sets into N partitions with T transactions in each partition.

Begin
Number of transactions in each partition (T)= Total transactions in D/N
Random number = N<cm
For each partition Ni DO begin
Take T transactions in Ni
Put each Ti in Ni
END
Return partitions with T transactions.

Illustration of PAFI algorithm

3. CONCLUSION

In this paper the improved Apriori algorithm and partitioning algorithm (PAFI) for frequent item sets is proposed with improved efficiency. In future, plan to apply classification methodology and pruning techniques for accuracy.
REFERENCES

[1]  A New Improvement on Apriori Algorithm by Lei Ji, Baowen Zhang, Jianhua Li. – June 2008


