Integrated Intelligent Research (IIR)
International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.110-113
ISSN: 2278-2419

Market Basket Analysis using Improved FP-tree

Abhishek Priyadarshi ,Chirag Gupta, G Poornalatha
Manipal Institute of Technology Manipal University, Manipal, Karnataka, India 576104
Email:abhi101993@gmail.com,911chirag@gmail.com,poornalatha.g@manipal.edu

Abstract — The Market Basket Analysis helps in identifying the purchasing patterns of customers such as, which products are purchased more and which products are purchased together. This helps in decision making process. For example, if two or more products are frequently purchased together then they can be kept at the same place so as to facilitate the customer, to further increase their sale. The price of products that are not frequently purchased can be reduced in order to enhance their purchase. Additionally the promotion of one product will also increase the sales of other products which are purchased together with the product being promoted. The traditional Apriori algorithm based on candidate generation cannot be used in Market Basket Analysis because it generates candidate sets and scans database regularly for the generation of frequent itemsets. The FP-growth algorithm cannot be used despite of the fact that it does not generate candidate sets and scans the database only twice because, it generates a lot of conditional trees recursively. Therefore, an efficient algorithm needs to be used. In this paper an efficient algorithm is used for development of market basket analysis application. This efficient algorithm neither generates candidate sets nor conditional FP- tree; like FP-growth scans the database twice.

Keywords - Association rule; FP-tree; Frequent pattern data; Market Basket Analysis; Data Mining

I. INTRODUCTION

Data Mining is a technique which is used to find the hidden patterns and useful knowledge in a large data. Market Basket Analysis makes use of this. It identifies the purchasing patterns [7] of customers such as which products are purchased more and which products are purchased together. This helps in business making decision such as which products should be sold more and with what combinations they should be sold.But the purchasing patterns of customers changes every season. For example, if we consider clothes as a product category, in summer season customer tend to buy more T-shirts while in winter season sweaters are purchased more. Therefore, generation of frequent itemsets has to be done on a regular basis and, so an efficient algorithm has to be used which not only takes less computational time but also less space such that it works on even larger growing database.

The traditional Apriori Algorithm cannot be used for this as it generates a large amount of candidate sets and scans database each time it has to generate frequent items. The FP-growth algorithm also cannot be used despite of the fact that it scans database twice and does not generate candidate sets. This is because it generates a lot of conditional trees recursively.Here, an application on market basket analysis has been developed which helps in finding the correlation among the items [8] purchased by customers by making use of improvedFP-tree algorithm which neither generates candidate sets nor conditional FP-tree. It takes less computational time and less storage space.

II. LITERATURE SURVEY

Data Mining is a process where hidden patterns and useful knowledge are extracted from large chunks of data and is used for decision making process. Association rules is a process which checks for relationship of an item with other items. Association rules [1] are ”if” and ”then” such as ”if X then Y” indicating if X occurs then probability of occurrence of Y will be high. Support and confidence are the two measures needed for developing association rules. Support is the indication of how frequently an item appears in the transaction where as confidence indicates the number of times ”if and then” statements found to be true. The frequent itemsets are those whose support is equal to or greater than the minimum threshold support.The basic algorithm for finding frequent itemset is apriori algorithm [3]. The algorithm works in two steps which are join and prune. Each Itemset is considered as candidate 1- itemset. The frequent itemsets that satisfy the support are combined to obtain the candidate set. The candidate set is pruned which is guided by the Apriori principle called downward closure property which states that, “if an itemset is frequent, then all of its subsets must also be frequent. The database is scanned for all itemset to find whether it is frequent or not. Here, algorithm extends candidate generation procedure of Apriori to add pruning using interest measure. Finally, the frequent itemset are obtained and the algorithm terminates when the entire frequent items are combined.

Frequent Pattern Tree (FP-tree) [2] is the other basic algorithm developed by Han et.al. The FP-tree [2] structure is for storing compressed and vital information about frequent patterns and develops an efficient FP-growth mining algorithm for mining complete frequent patterns. This has an advantage over the Apriori algorithm as it reduces space and time complexity. The algorithm requires two database scan; one is for constructing the tree to order frequent patterns and second is for tree branches building. First the frequent item sets are obtained by comparing the support count with the user defined support value. The frequent items are sorted in decreasing order for construction of the FP-tree. The sub frequent conditional patterns are extracted from the FP-tree for each itemset and the mining is done. Effciency of mining is achieved as (a) The FP-tree construction reduces the repeated scanning of the database reducing time complexity.(b) Mining is done by FP-Growth which does not produce costly large number of candidate set generation. This reduces the space for storage of large candidate sets. However, the FP-tree constructed for large data items require large storage space. (c) A divide and conquer method is used to decompose the mining

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.110-113
ISSN: 2278-2419

task into a smaller one for mining limited patterns in conditional databases, which reduces the search space.

Vembandasamy et al. [9] proposed improved FP-growth to find frequent patterns considering uncertain data like medical data. Vidya [10] used FP tree to reduce the time complexity while mining weighted association rule. But like other algorithms even FP-growth has some dis- advantages [4] like it generates FP-trees recursively due to mining. But in improved FP-tree conditional trees are not generated as recursive elements are stored separately in a different table called spare table.

III. PROPOSED WORK AND METHODOLOGY

The Market Basket Analysis uses two algorithms for analy sis. The first algorithm constructs a FP-tree without duplicate nodes [4] and the second algorithm mines frequent item from the constructed FP-tree [4].

A. Construction of FP-tree without duplicate nodes

- Input : Transactional Data.
- Output : Improved FP-tree, Frequency of each item in FP-tree, and the Stable table.

1) The items are sorted in decreasing order of their occurrence in the transaction.
2) The items in each transaction are considered one by one.
3) If the considered item is not present in FP-tree then a node for that item is constructed and its frequency is initialized to 1.
4) If the considered item is present in FP-tree and it belongs to the same sequence as required then the frequency of the item is incremented.
5) If the considered item is present in FP-tree but it does not belong to the current sequence then it is added to a table called spare table and its frequency in spare table is initialized to 1.
6) The total frequency of each item in stable table is added

B. Mining Frequent Itemset

- Input: Constructed FP-Tree, Stable Table, User defined Support.
- Output: Frequent Itemsets.

1) The items in constructed FP-tree and spare table are considered one by one.
2) If the frequency of item in the FP-tree is equal to user defined support then frequency of the itemsets for this item will be FP-tree frequency and all possible combi- nations of the item and nodes with higher frequency in FP-tree are generated.
3) If the frequency of item in the FP-tree is less than the user defined support then frequency of the itemsets for this item will be sum of the FP-tree frequency and frequency of this item in stable table and all possible combination of item and all intermediate nodes up to most frequent item node in FP-tree are generated.
4) If the frequency of item in the FP-tree is greater than user defined support then frequency of the itemsets for this item will be FP-tree frequency and all possible combinations of item and its parent node in FP-tree are generated.

The modified FP-tree can be explained by taking an example given in table I [4]. Let the minimum support be 1. The

Table I: Sample Transactions

ID	Transactions
1	Bread,Butter,Jam
2	Butter,Cheese
3	Butter,Sugar
4	Bread,Butter,Cheese
5	Bread,Sugar

Number of occurrence of each item is calculated by scanning the transactions. The table II gives the number of occurrence of each items. As all items satisfy the minimum support,

Table II: Transaction Items and Their Occurrence in Transactions

Item	Number of occurrence
Butter	4
Bread	3
Cheese	2
Sugar	2

all the items are considered and sorted in descending order based on number of occurrence of each items as given in table III. Now each of the transaction are considered one by

Table III: Descending Ordered Transaction Data

ID	Transactions
1	Butter,Bread,Jam
2	Butter,Cheese
3	Butter,Sugar
4	Butter,Bread,Cheese

one. The first transaction is (Butter, Bread, Jam) so Root will be Butter with its child as Bread and Jam each with count 1i.e. (Butter:1, Bread:1, Jam:1). Now the next transaction is considered which is (Butter, Cheese). Since Butter is already a root so its frequency is increased by 1 and Cheese is added as its child with frequency 1 i.e. (Butter:2, Cheese:3). Now moving to third transaction which is (Butter, Sugar, it will be similar like the second transaction. The next transaction is (Butter, Bread, Cheese). As Butter, Bread are in one branch and cheese is in another so instead of repeating the node for cheese it is added to a table called spare table as shown in table IV. Now the last transaction is (Bread, Sugar). Since this transaction has no frequent items so it is also added to the spare table. Now after the construction of modified FP-tree,

Table IV: Spare Table

ItemName	Frequency
Cheese	1
Bread	1
Sugar	1

mining of Frequent item will be done. For the minimum support threshold is to be considered. If the frequency of the item in the FP tree is equal to or greater than minimum support threshold then the frequency of the item will be its

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.110-113
ISSN: 2278-2419

frequency in the FP-tree but if the frequency of item in FP-tree is less than the minimum threshold then its frequency will be equal to the sum of its frequency in FP-tree and that in the spare table. After this the frequent itemsets are generated for each of the items and the association rule are discovered between different items with the help of the frequency of frequent itemsets. if the minimum confidence threshold is assumed to be 60% then:

R1 :Jam and Bread $\Rightarrow$ *Butter*$\frac{Frequency(Jam,Bread,Butter)}{Frequency(Jam,Bread)}$

which is 100% and so R1 will be selected.

R2 :Butter and Bread $\Rightarrow$ *Jam*

$\frac{Frequency(Butter,Bread,Jam)}{Frequency(Butter,Bread)}$

which is 50% and so R1 will be rejected.

R3 :Butter $\Rightarrow$ *Cheese*

$\frac{Frequency(Butter,Cheese)}{Frequency(Butter}$

which is 50% and so R3 will be rejected.

In this the strong association rules are discovered which will help in analyzing the data and taking the appropriate decision.

IV. IMPLEMENTATION

The application for market basket analysis has been developed. The backend has been implemented using Java language. The Front end are implemented using jsp, CSS, HTML. Mysql is used for database. Software used are Netbeans,xampp,windows 8. Dataset used is T10I4d100K which is available at Frequent Itemset Mining Repository [6]. The application works as follows. The products for sale are added by the administrator. Every new customer has to register before buying any product. Once the customer registers, he/she logins using his/her credentials. After that they select the product of their choice and add to the cart. There is a proviosion to customer to cancel the product anytime else once they save, it will be added to the transactions. Now Apart from adding products the administrator can view the generated FP-tree from the transactions and can also view the frequent items and generate the association rules. This helps in decision making process. Further the super-administrator will be able to view the modified FP-tree, generated frequent itemset for even larger dataset. The Figure 1 shows the overall flowchart of the program.

Fig. 1: Flowchart of Program

V. RESULT

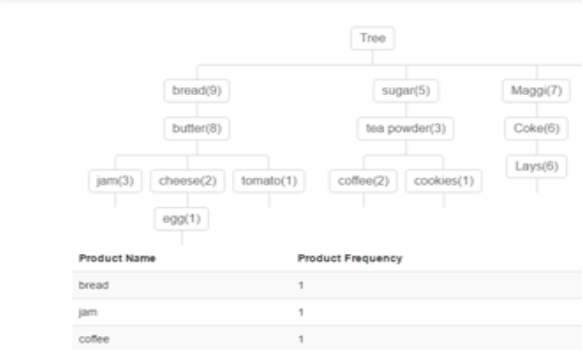

Fig. 2: Modified FP-tree

As shown in Figure 2 the modified FP-tree with distinct nodes are constructed. Also the items that are repeated are added into the spare table.Figure 3 depicts the time comparison graph of old FP- tree represented in black versus new FP-tree represented in blue. The X-axis is the Frequent itemsets while Y-axis is the execution time in milliseconds. As can be seen from the graph the new FP-tree takes less computational time than the old FP- tree.Figure 4 depicts the memory comparison graph of old FP- tree represented in black versus new FP-tree represented in blue. The X-axis is the Frequent itemsets while Y-axis is the memory usage in Bytes. As can be seen from the graph the new FP-tree occupies lesser memory than the old FP-tree.

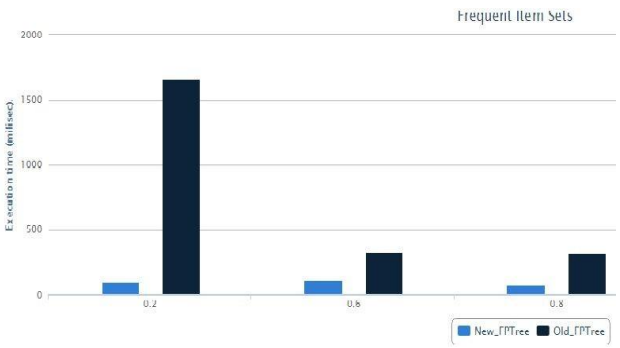

Fig. 3: Time Comparison

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.110-113
ISSN: 2278-2419

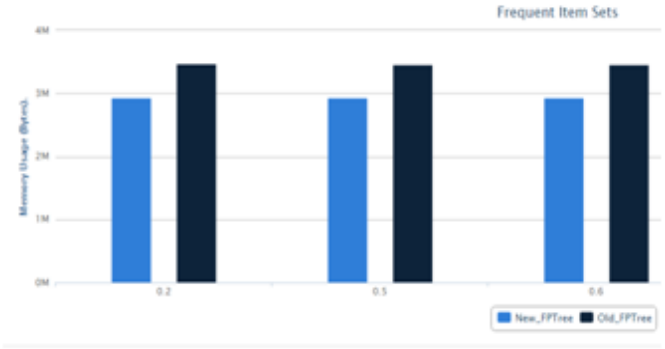

Fig. 4: Memory Comparison

VI. CONCLUSION

The modified FP-tree contains distinct nodes and also takes less computational time and lesser space than the traditional FP-growth algorithm and apriori algorithm. Therefore it is used to develop an application that helps in discovering purchasing patterns of customers which in turn helps in taking market decisions such as which product should be sold more, which products should be discarded or their price must be cut down so as to enhance their purchase. It will further help in other market strategies such as designing coupons etc. In future we will apply a new mining technique on graph based structure which will have distinct nodes just like the improved FP-tree.

References

[1] S. K. Solanki and J. T. Patel, Survey on association rule mining in 2015 Fifth International Conference on Advanced Computing Communication Technologies, Feb 2015, pp. 212-216
[2] H. P. J. Han and Y. Yin., Mining frequent patterns without candidate generation:A frequent pattern tree approach.
[3] A. K. Pujari,Data Mining Techniques.University Press (India) Pvt.Ltd., 2001
[4] T.S. C. A.B.M.Rezbaul Islam. An improved frequent pattern tree based association rule mining technique
[5] Y. Liu and Y. Guan, Fp-growth algorithm for application in research of market basket analysis, in Computational Cybernetics, 2008. ICCC 2008. IEEE International Conference on, Nov 2008, pp.269-272.
[6] http://fimi.ua.ac.be/data/
[7] Q.S.X. YONG QIU, YONGJIE LAN.An improved algorithm of mining from fp-tree.
[8] F. Gao and C. Wu, Mining frequent itemset from uncertain data in Elec- trical and Control Engineering (ICECE), 2011 International Conference on, Sept 2011, pp. 2329-2333.
[9] K. Vembandasamy and T. Karthikeyan, An Improved FP-Growth with Hyb-rid MGSO-IRVM Classifier Approach used for Type-2 Diabetes Mellitus Dagnosis, International Journal of Innovations in Engineering and Technology (IJIET), vol. 7, no. 6, pp.2293-2303.
[10] V. Vodya, Mining weighted association rule using FP-tree, International Journal on Computer Science and Engineering (IJCSE), vol. 5, no. 8, pp. 741-752.