Profitable Association Rule Mining using Weights

T.lakshmi Surekha[#1], P.Ramadevi[*2], J.Malathi[#3]

[#1]*Assistant professors & Department of Information Technology VRSEC,Kanuru,Vijayawada,Andhra Pradesh,Krishna Dist*
[#2]*Assistant professors & Department of Information Technology, VRSEC,Kanuru,Vijayawada*
[#3]*Assistant professors & Department of Information Technology , Sir C R Reddy College of Engineering, Vatluru, Eluru*

Abstract *---In recent years, a number of association rule mining algorithms like Apriori were developed, they are purely binary in nature. It doesn't consider quantity and profit(profit per unit). In these algorithms, two important measures viz., support count and confidence were used to generate the frequent item sets and their corresponding association rules . But in reality, these two measures are not sufficient for decision making in terms of profitability. In this a weighted frame work has been discussed by taking into account the profit(intensity of the item) and the quantity of each item in each transaction of the given database. FP Growth algorithm is one of the best algorithm to generate frequent item sets, but it does not consider the profit as well as the quantity of items in the transactions of the database. Here we propose an algorithm FP-WQ, which eliminates the disadvantages of frequent database scanning and it also considers quantity and profit per unit.. In this by incorporating the profit per unit and quantity measures we generates Weighted Frequent Itemsets (FP-WFI) and corresponding Weighted Association Rules (FP-WAR).*

Keywords— *FP-WQ, Weighted frequent item sets, Minimum Weight Threshold.*

I. INTRODUCTION

Data Mining is a process in the discovery of Knowledge from a very large database (VLDB). Association rule is one of the important techniques in data mining for extracting knowledge from a VLDB. Association rule finds the association or correlation between two sets of items. A typical example of association rule mining is market basket analysis. A market basket database is a transactional database containing Transaction Identifiers (TID) and a set of items bought by the customer. Association rule mining helps to find the buying pattern of customers, which will be very much useful to the sales managers in designing the catalog, target marketing, customer segmentation, planning the shelves and so on.

A. Basic Concepts

Let I = {i1, i2, i3, . . . , in} be a set of 'n' items. Let D be a set of transactions, where each transaction T is a set of items such that T ⊆ I. Associated with each transaction is a unique identifier called Transaction identifier (TID). An itemset is defined as any nonempty subset of I. The support count or support of an itemset X, is the number of transactions in D that contain all the items in X. An itemset X is called a frequent (or) large itemset (FIS), if support (X)≥ α where α is a user specified minimum support threshold . An association rule is an implication of the form A ⇒ B [s , c], where A and B are two non-empty disjoint itemsets (i.e.,) A ≠ φ , B ≠ φ, A⊂ I, B ⊂ I, and A ∩ B= φ.
The support 's' is the percentage of containing A that also contain B.

An association rule A ⇒ B [s , c] is said to be strong if s ≥ min_sup, and c ≥ min_conf, where min_sup and min_conf are the user specified minimum support threshold (MST) and minimum confidence threshold(MCT) respectively.
Consider the following two transactions:

T1: {20 Books, 5 pens}

T2: {1 Book, 1 pen}

In the support-confidence frame work the above two transactions are considered to be the same, since the quantity of an item is not taken into account. But in reality, it is quite clear that the transaction T1 gives more profit than the transaction T2. Thus to make efficient marketing we take in to account the quantity of each item in each transaction. In addition we also consider the intensity of each item, which is represented using profit per item p.
Consider the following two transactions:

T3: {10 Milk packets, 1 ice cream}

T4: {2 Milk packets, 3 ice creams}

In reality the quantity sold in transaction T3 is greater than transaction T4, but the amount of profit gained by selling an ice cream is 10 times that of a milk packet. So, the profit is also given priority represented by p, 'p' may represent the retail price / profit per unit of an item.

B. Problem Definition

In this paper taking into account the profit / intensity of the item and the quantity of each item in each transaction of the given database, we propose an algorithm FP-WQ, which eliminates the disadvantages of frequent database scanning and it also considers quantity and profits. In this by incorporating the profit per item and quantity we generates Weighted Frequent Itemsets (FP-WFI) and corresponding Weighted Association Rules (FP-WAR).

In this paper, the notations corresponding to the new structure is given below. Universal Itemset = I

I = { i1 :q1*p1 , i2 : q2*p2 , . . . , im :qm* pm } , { i1 , i2 , . . . , im } represents the items purchased, {q1,q2,q3.....qm} represents quantity of purchase, {p1,p2,p3.....pm} represents their respective profits.

The ith transaction of the database D is of the form

$Ti=\{wi_r,....\}$

r represents item number,

w = q*p.

The remainder of this paper is organized as follows. In Section II we made a review of the earlier work and in Section III we propose our new algorithm FP-WQ. Section IV describes the implementation of FP-WQ and the results of experiments on a market basket database. Finally, conclusions are drawn in Section V where we also indicate possible directions of future work.

II. BACKGROUND WORK

Frequent Pattern - Growth algorithm influential algorithm for mining frequent item sets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties. The FP-Growth Algorithm, proposed by Han in [3], is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for storing compressed and crucial information about frequent patterns named frequent-pattern tree (FP-tree).

FP-GROWTH ALGORITHM FOR DISCOVERING FREQUENT ITEMSETS FOR MINING BOOLEAN ASSOCIATION RULE

I. TABLE

FP-Growth: allows frequent itemset discovery without candidate itemset generation.
Two step approach:

Step 1: Build a compact data structure called the FP-tree

Step 2: Extracts frequent itemsets directly from the FP-tree
Step 1:

- Scan data and find support for each item.
- Discard infrequent items.
- Sort frequent items in decreasing order based on their support.

Use this order when building the FP-Tree, so common prefixes can be shared.
Step 2:

Nodes correspond to items and have a counter

1. FP-Growth reads 1 transaction at a time and maps it to a path
2. Fixed order is used, so paths can overlap when transactions share items (when they have the same prfix).
   - In this case, counters are incremented
3. Pointers are maintained between nodes containing the same item, creating singly linked lists (dotted lines)
   - The more paths that overlap, the higher the compression. FP-tree may fit in memory.
4. Frequent itemsets extracted from the FP-Tree.

III. PROPOSED WORK

Weighted Frequent Itemset (FP-WFI)[4]

Input:
1. D, a Transactional database(table1) that also includes quantity of items purchased.
2. min_sup, the minimum support count threshold.
3. Profit table that displays profit earned by each item(table 2)

Output:
The complete set of profitable frequent patterns.(table 7)
Method:
Step1:

1. Scan the database and develop a new table by multiplying profit of each item with their corresponding quantity , represented as weight(w=profit of an item * quantity of an item Table 3).
2. Calculate the frequency for each item i.e, sum of weights of each item in all the transactions(support Table 4).
3. Discard infrequent items.

4. Sort frequent items in decreasing order(table 5) based on their support and a new database table is generated(table 6).

Use this order when building the FP-Tree, so common prefixes can be shared.

Step2:

Nodes correspond to items and have a counter

1. FP-Growth reads the first transaction at a time and maps it to a path
2. Fixed order is used, so paths can overlap when transactions share items (when they have the same prefix).
   - In this case, counters are incremented by their corresponding weights
3. Pointers are maintained between nodes containing the same item, creating singly linked lists (dotted lines)
   - The more paths that overlap, the higher the compression. FP-tree may fit in memory.
4. Frequent itemsets extracted from the FP-Tree.

IV IMPLEMENTATION OF PROPOSED WORK

TABLE1: ORIGINAL DATABASE

T100 - 2I1,3I2,4I5

T200 - 3I2,4I4,1I5

T300 - 3I2,4I3

T400 -4I1,8I3,2I4,1I5

T500 -3I1,2I2,3I5

T600 4I4,3I3,2I2,5I1

Note: Where I Represents The Item Number And Prefix of I represents quantity purchased

TABLE 2: PROFIT TABLE

Item	Profit	Profit/item
I1	P1	3
I2	P2	2
I3	P3	1
I4	P4	5
I5	P5	6

TABLE 3: MULTIPLY QUANTITY WITH WEIGHTS

Transaction id	Item with weights
T100	6I1,6I2,24I5
T200	6I2,20I4,6I5
T300	6I2,4I3
T400	12I1,8I3,10I4,6I5
T500	9I1,6I2,18I5
T600	15I1,4I2,3I3,20I4

TABLE 4 : CALCULATE THE FREQUENCY OF EACH ITEM

	I1	I2	I3	I4	I5
T1	6	0	0	0	24
T2		6		20	6
T3		6	4		
T4	12		8	16	6
T5	9	6			18
T6	15	4	3	20	
Total	42	22	15	56	54

TABLE 5: ARRANGE THE ITEMS ACCORDING TO THE DESCENDING ORDER OF THEIR FREQUENCY

Items	I4	I5	I1	I2	I3

TABLE 6: DATA SET AFTER CALCULATING SUPPORT

T100 - 24I5,6I1,6I2

T200 - 20I4,6I5, 6I2

T300 - 6I2,4I3

T400 -10I4,6I5,12I1,8I3

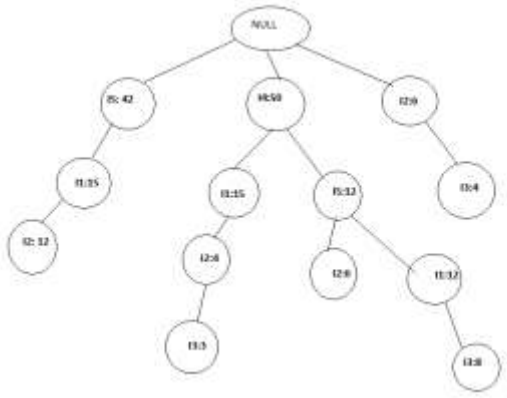
T500 -18I5, 9I1,6I2

T600 -20I4, 15I1,4I2,3I3

V. CONCLUSIONS & RESULTS

In this paper the following issues are highlighted.

- The profit and quantity factor given to each item in each transaction is used to find the weight of each frequent set that gives the importance of each frequent set. Two frequent sets may have same support count but the weight (or) profit of a frequent set decides which is more important. The marketers are more interested in the profit than the frequency of a set.
- Frequent item sets and corresponding association rules were prioritized based on their weights.



Frequent profitable itemset assuming min_sup count is 11 as shown in the below table

Table 7:

<I4, I3:12>
<I5, I1, I2:12>
<I5, I2:12>
<I1, I2:12>
<I5, I1:15>
<I4, I1:15>
<I4, I5, I1:12>

REFERENCES

[1] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", *Proc. 1993 ACM SIGMOD*,Washington, DC, pp. 207–216, May 1993

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", *Proc. 20th International Conference on Very Large Databases* (VLDB'94), Santiago, Chile, pp. 487–499, Sept. 1994

[3] J. Han and Micheline Kamber, "Data Mining – Concepts and Techniques", Morgan Kaufmann Publishers, 2001

[4] Robert J. Hilderman, Colin L. Carter, Howard J. Hamilton, And Nick Cercone, "Mining Association Rules From Market Basket Data Using Share Measures and characterized Itemsets"

[5] Feng Tao, Fionn Murtagh, Mohsen Farid, "Weighted Association Rule Mining using Weighted Support and Significance framework"

[6] Wei Wang, Jiong Yang, Philip S.Yu, "Efficient Mining of Weighted Association Rules (WAR) "

[7] Kantardzic, Mehmed, "Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley & Sons, 2003

.