Market Basket Analysis: A Profit Based Product Promotion Forecasting

HYS Samarasinghe#1, WJ Samaraweera#2 CP Waduge#2
# Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka
#1 yasarasam929@gmail.com

Abstract- Data mining is the area that helping extracting the useful information by finding patterns or rules from the existing dataset. By using the extracted information then used to predict future tendencies and behavior patterns. Association mining is a branch of data mining which used to identify itemsets that take place frequently in a specific dataset and to determine rules. Association mining can find out the rules that predict the occurrence of an item with regard to the similar occurrences of other in a particular transaction. Eclat algorithm is kind of a frequent itemset mining which is a sub section of the association mining based on the mining frequent patterns by exploring the vertical data format. Eclat algorithm was actually developed for Market Basket Analysis which is an effective technique in retail industry that helps the shop owner to increase the sales distribution techniques. Market Basket Analysis is completely done by the association rule mining in which analyses the customer buying behavior against the purchasing item from the shop. Eclat algorithm is the one of the most effective ways to mining of large data set since it follows the depth in search. When it comes to the real world, the main objective of market basket analysis is to gain maximized profit at all with the help of operational research theories. In this approach, the condensed data is used for mine the frequent itemset using the Eclat algorithm. After all, one of the operational research theories which are termed linear programming will use to maximize the profits.

Support value and the Confidence value are the foremost factors in generating the Eclat. Eclat algorithm abandons Apriori's breadth-first search for a recursive depth-first search. Moreover, consideration of frequent items as well as non-frequent items, considerably impact the profit maximization. Because if the retail owner identified the non-frequent itemset; can provide the promotions to the customers. It will enhance the profit maximization. Therefore, this research was mainly focused to identify frequent itemset as well as the non-frequent itemset in a market basket analysis alone with the profit maximization using linear programming. This developed approach is applied to a real world dataset and results were compared considering Eclat algorithm and Eclat algorithm alone with the linear programming separately. Finally, the results conclude that proposed approach significantly increase the profit.

Key Words: Eclat Algorithm, Support Value, Confidence Value, Market Basket Analysis, Depth-First Search

I.INTRODUCTION

Today many modern trading or retail shops are challenged with the mining of vital consumer data set from the huge data collection and item featured databases to achieve more profits for the organization (S.O. Abdulsalam, 2014). With the aid of the data mining, analysts can solve business consequences with high complexity. Market basket analysis determines the information about the related sales on combination of products (Mamtora, 2014). So, modern trading retailers are used to group the related products nearly, in order to remind the customers about the related products and lead those to purchase without any influence in mind but in a logical manner and also it can be a good situation to most selling product promotions though.

Market basket analysis is the most effective way to identify market owners to increase the market sales, profits and consumer buying behavior patterns (Pooja Pandey, September 2016). Market based analysis is a vital module when analyzing the placement of retail products, sale discounts, customer satisfaction and customer retention while increasing the profits of the modern trades (Bogdan Hoanca, 2011). In accordance with the promotion products that go well with the real products is totally combined with the association rule mining which is analyzing the consumer buying behavior with the specific product that consumer purchasing from the retails. By analyzing the consumer buying behavior patterns; which reveals the frequent sales and non-frequent sales in products, association mining which is a sub part of data mining helps to produce the frequent itemsets.

There are many algorithms to produce association mining rules such as Eclat algorithm, FP-Growth algorithm, Apriori algorithm, K-means, K-nearest Neighbor Classification, Naïve Bayes, K-Apriori etc. This research based on the Eclat algorithm which is the most efficient way to generate the frequent itemset. By generating the frequent itemset it is easier to identify the most necessary itemset and the most annoying itemset for the consumers. Also, it will enhance the market strategies, convenience of the consumer and increasing level in sales of goods and any one can do profitable business. To maximize the profit using operational research theories will enhance the outcome of the effectiveness. Combing the éclat algorithm alone with the linear programming is the main core of this research article. Using the Market Basket Analysis; it can be used to reveal cross-sell and related products. The prediction will be done by the aid of mined dataset from the sales department of the retail shop.

Eclat algorithm is kind of a frequent itemset mining which is a branch of a data mining and the sub section of the association mining. Eclat algorithm was actually developed for the Market Basket Analysis (Borgelt, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012). Finding the similarities in the shopping basket is the main aim of the mining the frequent itemset. Also, this can get to know about the set or couple of products that are frequently bought together. These types of mined itemsets of associated products may be used to the giving promotions to the customers and optimize the retail shop of the offered products on the shelves when there is low sales in different kind of associate products. They may give the hints that some products may conveniently be together or tend to buy other products.

II.LITERATURE REVIEW

Association rule mining connects with the one/ more attributes in a one dataset with another attributes. It helps to find out hidden patterns and relationships among the attributes. There are many algorithms for creating association rules. Eclat (Mohammed Javeed Zaki, 1997), Apriori (Rakesh Agrawal, 1994), FP-Growth (Jiawei Han, 2000), Naïve Bayes (Rahman, 2010) are the bit of the association mining algorithms. There are two main measures that association rules use. They are Support and Confidence. By analyzing data for frequently used these two values identifies the relationships and then the patterns of a particular dataset. Association rules are usually needed to satisfy a user-specified minimum support and a user -specified minimum confidence at the same time of same dataset. Also, there is 2-step approach in generation of the association mining rules. Frequent itemset generation where generate all itemset that Support ≥ minsup (Support Threshold) and Rule generation where generate the high confidence rule from each frequent itemset. So, frequent itemset generation is computationally expensive. The Eclat algorithm was first introduced by Mohammed Javeed Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, and Wei Li (1997) which generates the frequent itemset based on the 'Support Threshold' or minimum support.

A. *Useful concepts of Eclat algorithm*

- Support - In simply, support value of an itemset is the percentage of transactions in the Database that enclose the itemset.

The support value of a rule, *A* → *B*, is the percentage of transactions in T that contains *A* ∪ *B*, and can be seen as an estimate of the probability, (*A* ∪ *B*) (WJ Samaraweera, 2016). It can calculate as follows,

$$Support = \mathrm{P}(A \cup B)$$

- Confidence – In simply, confidence value, *A* → *B*, is the support value of the itemset of all items that appear. It can calculate as follows,

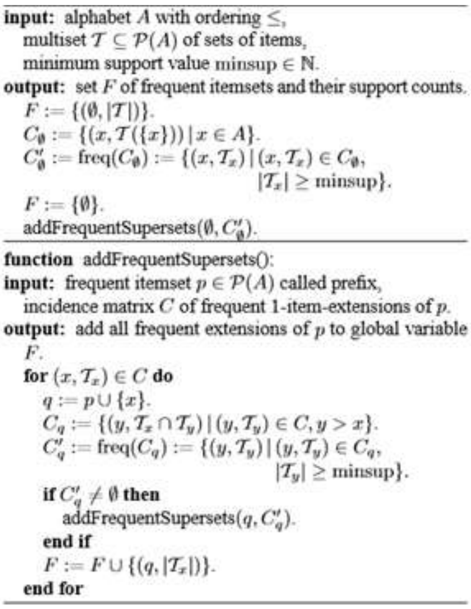
$$Confidence = \frac{supp(A \cup B)}{supp(A)}$$

- Support Threshold - Value that helps to remove non-frequent items from a database. (Borgelt, Eclat:Find Frequent Item Sets with the Eclat Algorithm, 2012)

B. *Eclat Algorithm*

Eclat, as defined by Goethals (2003) is presented as an Algorithm. Eclat is the acronym for Equivalent CLass Transformation. It is used for mostly in itemset mining. Eclat mainly improves the efficiency of Apriori algorithm. This algorithm based on Depth-first search manner (Heaton, 30 March-3 April 2016) which is an initial elegant algorithm suitable for both sequential as well as parallel execution with locality and enhancing the properties. Because of the expensiveness of the frequent itemset generation Eclat algorithm was born. By using the Eclat algorithm it is efficient in generating the frequent itemset manner of Depth-first search. Eclat is an algorithm for discovering itemsets (group of items) occurring frequently in a transaction database (frequent itemsets). Because of the Depth-first search manner, it is reduces the memory requirements. Engage with Eclat algorithm, it is no need to scan the database to find the support of ($K^{+}1$) itemsets, for K>1. But transactional ID (TID) sets can be quiet long.

C. *Pseudo Code for Eclat Algorithm*



```
input: alphabet $A$ with ordering $\leq$,
  multiset $\mathcal{T} \subseteq \mathcal{P}(A)$ of sets of items,
  minimum support value $\text{minsup} \in \mathbb{N}$.
output: set $F$ of frequent itemsets and their support counts.
  $F := \{(\emptyset, |\mathcal{T}|)\}$.
  $C_\emptyset := \{(x, \mathcal{T}(\{x\})) \,|\, x \in A\}$.
  $C'_\emptyset := \text{freq}(C_\emptyset) := \{(x, \mathcal{T}_x) \,|\, (x, \mathcal{T}_x) \in C_\emptyset,$
                                  $|\mathcal{T}_x| \geq \text{minsup}\}$.
  $F := \{\emptyset\}$.
  addFrequentSupersets($\emptyset, C'_\emptyset$).

function addFrequentSupersets():
input: frequent itemset $p \in \mathcal{P}(A)$ called prefix,
  incidence matrix $C$ of frequent 1-item-extensions of $p$.
output: add all frequent extensions of $p$ to global variable
  $F$.
  for $(x, \mathcal{T}_x) \in C$ do
    $q := p \cup \{x\}$.
    $C_q := \{(y, \mathcal{T}_x \cap \mathcal{T}_y) \,|\, (y, \mathcal{T}_y) \in C, y > x\}$.
    $C'_q := \text{freq}(C_q) := \{(y, \mathcal{T}_y) \,|\, (y, \mathcal{T}_y) \in C_q,$
                                  $|\mathcal{T}_y| \geq \text{minsup}\}$.
    if $C'_q \neq \emptyset$ then
      addFrequentSupersets($q, C'_q$).
    end if
    $F := F \cup \{(q, |\mathcal{T}_x|)\}$.
  end for
```

Figure 1: Pseudo Code for Eclat Algorithm

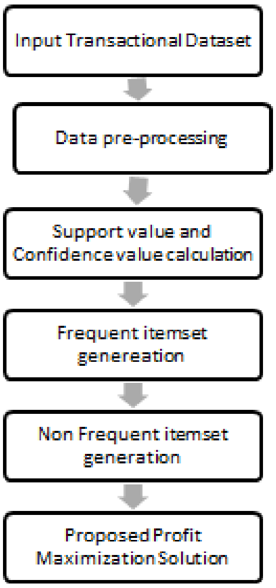
(Schmidt-Thieme, January 2004)

During the World War II the Linear programming was developed to increase the efficiency of resources was of highest importance. Actually, the term "programming" was a military term which referred to set of activities such as organizing the people and schedules that needed for the military work efficiently and optimally. Simplex method of optimization was first developed by the member of U.S. Air Force George Dantzing (Lewis, 2008) which can give an efficient way for solving linear structured problems. Linear programming uses a mathematical model which can describe the problem that you concerning about. The word *linear* means all the functions in the model should be in linear and the word *programming* define to not computer programming rather it to be planning (FREDERICK S. HILLIER, 2000). So, *linear programming* involves in planning collection of activities to get the optimal result as the output. In this research paper linear programming problem is defined as the problem of maximizing the function/ profit. In any kind of maximizing/ minimizing functions focus to the linear constraints; constraints may be equalities or inequalities (Ferguson).

III.METHODOLOGY

Proposed research work is mainly based on the Market Basket Analysis which is increase the profits by analyzing the shopping basket using the frequent itemset mining done by the Eclat algorithm. The proposed research work will give the results on frequent itemset, non-frequent itemset regarded to the promotional products and maximizing the profit using OR concepts which is the linear programming. Different data sets from supermarket have been used to calculate the values.

The workflow of the proposed work is shown below.



Input Transactional Dataset

Data pre-processing

Support value and Confidence value calculation

Frequent itemset genereation

Non Frequent itemset generation

Proposed Profit Maximization Solution

Figure 2: Work-flow of proposed methodology

Item	TID
A	1, 3, 5, 7, 9, 10
B	1, 2, 3, 4, 5, 6, 8, 9
C	3, 5, 7, 10
D	1, 2, 4, 7,10
E	2, 3, 4, 5, 6, 7, 8, 9, 10
F	3, 5, 9
G	1, 4, 6

There are number of association mining algorithms that can generate the frequent itemset and analyze the combination if the products. By analyzing the customer buying patterns will help the most frequently buying couple products and forecasting the promotion products by generation of non- frequent itemset. When maximizing the profits linear programming methods are using for the proposed research work. Analyzing the different constraints such as total execution time, total memory usage, and the most frequent itemset will be feasible for when forecasting the promotion products in retail shops. During this research work, two variables have identified when calculating profit maximizing using linear programming concepts.

a) No. of frequent itemset maximization ($\acute{Z}_F$)

b) No. of non-frequent itemset maximization ($\acute{Z}_{NF}$)

Both variables that have identified above are maximizing the profit.

The set of Transactions given below.

Table 1: Transaction Set

The result of which is tabulated alongside the support value of each item, profit of frequent itemset ($\acute{Z}_F$) and profit of non-frequent itemset ($\acute{Z}_{NF}$).

Table 2: The result of which is tabulated alongside the support value of each item

So, the profit maximization will be as follows:

$$Z = \acute{Z}_F + \acute{Z}_{NF}$$

$$\text{Maximize } \acute{Z}_F = \text{Support Value}_{(A,B,C,D,E,F,G)} + \text{Profit } \acute{Z}_F$$

$$\text{Maximize } \acute{Z}_{NF} = \text{Support Value}_{(A,B,C,D,E,F,G)} + \text{Profit } \acute{Z}_{NF}$$

IV.DISCUSSION

When considering the executing time using Eclat algorithm, it takes less time rather than using the Apriori and FP-growth algorithms. Before, running the Eclat algorithm for the transaction dataset most of the researchers have had arrange the data in a bit matrix and then run the Eclat algorithm on it. Calculating of the frequency of the itemset is considering the counting of the occurrence in each dataset. Eclat algorithm performance takes the highest value by considering other frequent mining algorithms. The key idea of Eclat algorithm is to use the tid set (transaction id) intersection to compute the support of each candidate item set and to avoid the generation of subset that does not exist in prefix tree. To perform the Eclat algorithm database is scanned for few times. Virtual memory is needed to accomplish the operation. Eclat takes less execution time and memory than other two frequent mining algorithms. However, the main operation of Eclat is intersecting tidsets, and therefore the size of the tidset is one of the main factors that affect the running time and memory. Whereas the tidset is large, it takes more space to store the candidate set and it needs more time for intersecting tid sets. Even if the dataset is too large and it takes more time to mining Eclat counted more frequent itemset than the other two well-known frequent mining algorithms. But Eclat always takes much time when the transactional entries are larger. But fore medium and small dataset Eclat is faster than FP-growth and Apriori.

V.CONCLUSION

The above proposed research work generates the results of promotional products based on the non-frequent items in the transaction database which are not tend to buy, by the customers and maximizing the profit by selling the non-frequent items. The proposed methodology in this research consists operational research theories which is a only consider about the linear equations, *linear programming* commencing the maximizing profit process and generating promotional products using the Eclat algorithm which is going to extract the non-frequent items. This research enhances the profit gain in a transaction. Simultaneously this profit constraint facilitates the rare items without disturbing frequent items. When conveying with outsized real world data sets the results might be vary depending on the predefined values.

VI. REFERENCES

Item	Support Value	Profit $\acute{Z}_F$	Profit $\acute{Z}_{NF}$
A	0.6	10	5
B	0.8	20	10
C	0.4	30	15
D	0.5	40	20
E	0.9	50	25
F	0.3	60	30
G	0.3	70	35

Bogdan Hoanca, K. M. (2011). Using Market Basket Analysis to Estimate Potential Revenue Increases for a Small University Bookstore . *Conference for Information Systems Applied Research, 2011 CONISAR Proceedings,v4 n1822*, (pp. 1-11). Wilmington North Carolina, USA.

Borgelt, C. (2012). *Eclat:Find Frequent Item Sets with the Eclat Algorithm.* Retrieved 03 28, 2019, from Borlget.net Web site: www.borgelt.net

Borgelt, C. (2012). *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.* United Kindom.
Ferguson, T. S. (n.d.). *LINEAR PROGRAMMING A Concise Introduction.*

FREDERICK S. HILLIER, G. J. (2000). *Introduction to Operations Research.* New York,NY,10020.: McGraw-Hill Higher Education.
Heaton, J. (30 March-3 April 2016, Jan 30). Comparing Dataset Characteristics that Favor the Apriori, Eclat or FP-Growth Frequent Itemset Mining Algorithms. *SoutheastCon 2016.* Norfolk, VA, USA: IEEE.

Jiawei Han, J. P. (2000). Mining Frequent Patterns without Candidate Generation. *SIGMOD '00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 1-12). New York, NY, USA: ACM Press.

Lewis, C. (2008). *Linear Programming: Theory and Applications.*

Mamtora, S. G. (2014). A Survey on Association Rule Mining in Market Basket Analysis . *International Journal of Information and Computation Technology*, 409-414 .

Mohammed Javeed Zaki, S. P. (1997). *New Algorithms for Fast Discovery of Association Rules.* Rochester, New York.

Pooja Pandey, I. S. (September 2016). Improving Accuracy using different Data Mining Algorithms . *International Journal of Computer Applications (0975 – 8887) Volume 150 – No.10*, 10-13.

Rahman, S. M. (2010). Text Categorization using Association Rule and Naïve Bayes Classifier. *Asian Journal of Information Technology*, Vol. 3, No. 9, pp 657-665.

Rakesh Agrawal, R. S. (1994). Fast Algorithms for Mining Association Rules . *20th VLDB conference*, (pp. 487-499). CA 95120 .

S.O. Abdulsalam, K. A. (2014). Data Mining in Market Basket Transaction: An Association Rule Mining Approach . *International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 7– No.10, October 2014 – www.ijais.org* , 15-20.

Schmidt-Thieme, L. (January 2004). Algorithmic Features of Eclat. *FIMI '04, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations.* Brighton, UK, November: IEEE.

WJ Samaraweera, CP Waduge. (2016). Market Basket Analysis: A Profit Based Approach to Apriori Algorithm. *Proceedings in Computing, 9th International Research Conference-KDU, Sri Lanka* , (pp. 127-133). Colombo.