Market Basket Analysis: A Profit Based Approach to Apriori Algorithm

WJ Samaraweera[1#], CP Waduge[1], and RGUI Meththananda [2]

[1] *Faculty of Computing, General Sir John Kotelawala Defence University, Sri Lanka*
[2] *Faculty of Built Environment and Spatial Sciences, General Sir John Kotelawala Defence University, Sri Lanka*
[#]wishjeewa88@gmail.com

Abstract—*The field of data mining seeks to recognize the regularities, patterns and behaviours of large data collections. Association mining is used to discover elements that occur frequently within a dataset consisting of multiple independent selections of elements and to discover rules. This mining approach can find rules which predicts the occurrence of an item, based on the occurrences of other items in a particular transaction. Apriori algorithm is an influential algorithm designed to operate on data collections enclosing transactions such as in market basket analysis. To address various issues Apriori algorithm has been extended in different perspectives. In real world scenario, one of the major objectives in performing a market basket analysis is to maximize the profit. In Apriori algorithm, Support value and Confidence value are the dominant factors in generating association rules which seems to be insufficient to achieve the said objective as the algorithm does not consist a variable to maximize the profit gain. Moreover, consideration of frequent items, rather than rare items, significantly impact the profit maximization. Therefore, this research was focused to develop a new algorithm based on an extended Apriori approach which maximize the profit of a transaction using frequent items as well as rare items in a market basket analysis. The developed new algorithm and the extended Apriori algorithm were applied to a real world data set and the results were compared focusing the profit gain from each algorithm separately. Finally, the results conclude that the proposed algorithm derives association rules which significantly increase the profit gain, disregard of the number of items involving in the transaction.*

Key Words: Apriori Algorithm, Support Value, Confidence Value, Market Basket Analysis

I. INTRODUCTION

Retailing is one of the leading businesses in the world and supermarket is a commercial establishment based on customer self-service in retailing. In supermarkets variety of products are showcased in shelves, according to their own arrangement, making the customer comfortable with purchasing products. The objective of this facility is to provide the customer an opportunity to explore different brands and prices offered by different companies in accordance with their requirements. Conversely, it provides easy access to go through products with less necessity before attaining true requirements. This empowers the objective of making the basket of the customer outsize to enhance the profit of the vender, by tempting customers to buy items which were not intended to buy before entering the place. This illustrates the importance of high concentration in arranging items for floors and shelves. Association rule mining is a branch of data mining, which is sourced to address this issue, arranging shelves or floors, by finding rules that will predict the occurrence of an item based on combinations of products that frequently co-occur in transactions. It helps the retailers in supermarkets to identify the relationships among the purchased items by customers in order to upgrade better customer satisfaction and retention.

There are several algorithms to generate these association mining rules such as Apriori algorithm, FP-Growth algorithm, K-means, K-nearest Neighbour Classification, Naïve Bayes, K-Apriori, Eclat etc. This research is based on the influential algorithm, Apriori, to address market basket analysis, identifying products that go well together, to gain better rules for floor and shelf arrangements.

The Apriori Algorithm was introduced by Aggarwal and Srikant (1994) which delivers a way to find frequent itemsets in a market basket analysis. Predefined Minimum Support Value, a factor based on industrial exposure, Minimum Support, how frequent the item appears in the transaction set calculated with respect to the data set, and

Confidence Value, the probability of purchasing an item when an another item has already been purchased, are the basic filters, help to generate association rules in this algorithm which absorbs patterns associates with frequent items only.

Although the ultimate objective of the vender is to maximize profit, it is identified that the constraints confederate with frequency of an item are not sufficient to encounter this objective. Alternatively, exponentially proportional growth of association rules with the expansion of the dataset is another major observation in utilizing the Apriori algorithm. The research is conducted to discover a new approach to incapacitate the above mentioned observations. The proposed modification to the Apriori algorithm is an introduction of a new constraint associates with profit. The new constraint, running parallel to the minimum support and the support, enhances the total profit gain by generating rules, considering both rare and frequent items, to arrange the shelves of a supermarket. Further it prunes the unnecessary rules generates with the expansion of the data set intensifying the efficiency of the process.

II. LITERATURE REVIEW

A large number of association rule mining algorithms have been developed with different mining efficiencies. Apriori (Agrawal and Srikant, 1994), FP Growth (Han et al., 2000), Eclat (Han and Kamber, 2001), K-Apriori (Annie and Kumar, 2011), K-Means (Liu et al., 2014), K-Nearest Neighbor (Larose and Larose, 2005) and Naïve Bayes (Kamruzzaman and Rahman, 2010) are some of the association rule mining algorithms. These algorithms can be categorized into two types called candidate generation or pattern growth. Apriori Algorithm is one of the most popular and influential algorithms in association rule mining categorized under candidate generation. It is an algorithm for frequent itemset mining and association rule learning over transactional databases. The Apriori Algorithm was first introduced by Agarwal and Srikant (1994) which generates frequent itemsets based on a threshold called 'Minimum Support'.

A. Useful Concepts in Apriori Algorithm.

- Itemset - A collection of one or more items (that represents together a single entity)

  Eg: - {Milk, Bread, Diaper}
- Minimum Support – A user defined value which helps to eliminate non-frequent items from a database.
- Frequent Itemset – An itemset that occurs in at least a user specific percentage of the database (the sets of item which has minimum support).
- Support – The support of a rule, $X \rightarrow Y$, is the percentage of transactions in T that contains $X \cup Y$, and can be seen as an estimate of the probability, $P(X \cup Y)$. Support determines how frequent the rule is applicable in the transaction set T. The support of rule $X \rightarrow Y$ is computed as follows:

$$Support = P(X \cup Y) = \frac{count(X \cup Y)}{n}$$

- Confidence - The confidence of a rule, X → Y, is the percentage of transactions in T that contains X also contains Y. It is the conditional probability, $P(X|Y)$. The confidence of the rule X → Y is computed as follows:

$$Confidence = \frac{count(X \cup Y)}{count(X)}$$

B. Apriori Algorithm.

The Apriori algorithm works in two steps:

1. Generate all frequent itemsets – A frequent itemset is an itemset that has transaction support above minimum support.
2. Generate all confident association rules from frequent itemsets – A confident association rule is a rule with confidence above minimum confidence.

C. Pseudocode for Apriori Algorithm.

C_k- Candidate itemset of size k

L_k - Frequent itemset of size k

L_1= {frequent items};

For (k = 1; L_k != ∅; 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 (minimum support)

end

return $U_k L_k$;

Generation of Association Rules is one of the major tasks in Data Mining. Association Rule Mining is all about finding rules whose support and confidence exceed the threshold and minimum support values. These association rules can be used in numerous real world tasks such as Market basket analysis, Customer segmentation, Fraud detection, Detection of patterns in text and Medical diagnosis.

Apriori Algorithm can be mainly utilized to generate the association rules in Market Basket Analysis. Market Basket Analysis (Raorane et al., 2012; Annie and Kumar, 2011) is one of the most frequently used data mining technique used to generate association rules. The purpose of Market Basket Analysis is to discover purchasing patterns of products from a supermarket's transactional database. Typically in supermarkets very large and constantly growing databases are maintained. From these large collection of data, it is really difficult to extract the data related to the pattern of buying products of customers. Association rules in Market Basket Analysis are frequently used by retail stores to assist in marketing, advertising, floor placement and inventory control. Direct Marketers could use this technique to determine the layout of their catalogue and order forms also.

Eg: A grocery store noticed that 100% of the time that Peanut Butter is purchased, so is Bread. Furthermore, 33.3% of the time Peanut Butter is purchased, Jelly has also been purchased.

It is profit oriented that Peanut Butter and Bread or Peanut Butter and Jelly are arranged in side by side in shelves of the grocery store. Such information will help the grocery store to decide which items can be put together in order to tempt the customer to buy more things in a logical manner.

But Apriori Algorithm suffers from some main limitations such as unnecessary memory utilization by generating a vast number of candidate sets with higher frequent itemsets, low minimum support or large itemsets. (Rao and Gupta, 2012) Furthermore Apriori Algorithm has a high scanning time since it needs to check for many more itemsets and they have to be scanned repeatedly in consequent steps.

Several aspects of Apriori Algorithm have been studied in academic literature in order to generate association rules while declining limitations of Apriori Algorithm. One of such aspect is mining association rules with multiple minimum supports (Liu et al., 1999). The Extended Model (MSApriori) allows the user to specify multiple minimum supports to reflect the items and their frequencies in the database. It generates all large itemsets by making multiple passes over the data. This model emphasizes that having a single minimum support value is insufficient. If it is set too high, necessary rules may not be generated and on the other hand if it is set too low, combinatorial explosion will occur. It is proved here that using multiple minimum supports instead of single minimum support value will provide two conclusions; rare items will not be ignored and number of generated rules will be less compared to initial Apriori algorithm. Another approach of Apriori Algorithm is introducing new parameters to maximize profit (Trikha and Singh, 2014). This algorithm enhance the efficiency of generating association mining rules by making a model which will be beneficial in eliminating the shortcomings of Apriori Algorithm. Two new parameters called, Q-factor using profit ratio and Profit Weighting factor (PW factor) were introduced in order to identify interesting patterns from transactional databases and to maximize profit.

A different approach called Improvised Apriori Algorithm using frequent pattern tree was suggested for real time applications. This algorithm focuses on reducing time spent to scan large

number of candidate itemsets and saving space utilized by unnecessary association rules (Bhandari et al., 2015). The improvised algorithm will scan only some transactions by a formula which partitions the set of transactions into sections and select one particular section among them. In new model it has been observed that the time consumed in group of transaction is less than the classical Apriori Algorithm and the difference increases more when the number of transactions increases. Though this approach reduces consumed time than the original Apriori Algorithm, it only reduces the time consuming by 67.87%.

There are several other contemporary approaches to Apriori Algorithm such as a secure mining of Association Rules which is based on the Fast Distributed Mining Algorithm (Tassa, Open, and Road, 2014). Furthermore an Adaptive Implementation of Apriori Algorithm was proposed in order to reduce the response time significantly by using the approach of mining the frequent itemsets (Balaji et al., 2013). An association classification based on compactness of rules is proposed but it suffers from a difficulty of over fitting (Qiang et al., 2009). How to maximize the efficiency of the parallel Apriori Algorithm is discussed and it is suggested that the efficiency can be improved effective load balancing (Shah and Mahajan, 2009).

A new approach primarily based on Apriori Algorithm is proposed in this paper, which considers profit as a variable when generating frequent itemsets. Though under theoretical framework the main variables that have been considered in Apriori Algorithm are Minimum Support and Confidence, when considering the real world scenarios, profit is the main variable that should be considered. This paper focuses on a new variable called profit of each product which calculates the profit margin with respect to the number of transactions other than the mean constraint of minimum supports (Samaraweera et al., 2014). Furthermore, proposed algorithm controls the exponential growth of association rules quantity as the size of the dataset increases. In addition to these reasons, Rare Item Problem is also addressed through this new approach. Since rare items generate more profit than frequent items, it is necessary to consider rare items as well.

III. METHODOLOGY

Proposed research work is based on an improvement of MS Apriori algorithm that enhances the effectiveness of the process by constructing a model which is beneficial in overcoming the shortcomings of Apriori algorithm. The frequent itemset which gives even 100% confidence with the classical Apriori algorithm may not provide maximum profit gain to the vender. The proposed algorithm calculates a profit factor which supports to maximize profit, associates with various frequent itemsets generated. The proposed improvement of the algorithm is implemented using Matlab.
Different data sets from different market environments has been used to check the internal consistency of the proposed algorithm.

The workflow of the proposed work is shown below in Figure 1.

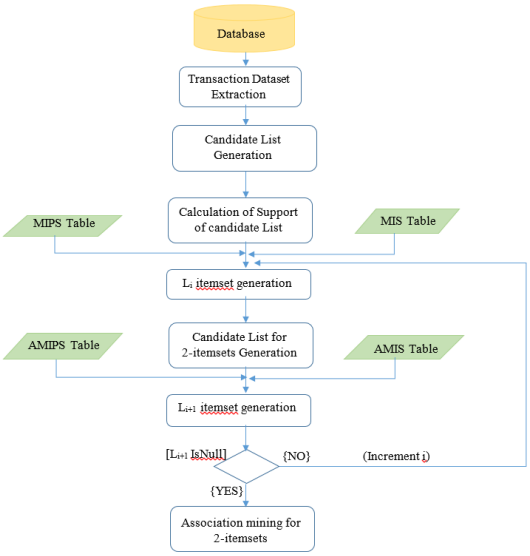

Figure 1: Work flow of the proposed algorithm

Following are the steps involved in the proposed methodology. It explains how the proposed work has been done.

INPUT

A set of n transaction data, each item, $i = 1, \dots, m$ with support (sup_i), user defined minimum support (ms_i), profit $(prof_i)$ and user defined minimum profit (mp_i), and a minimum confidence value λ where,

$sup_i = \frac{count(i)}{n}$ and $prof_i = \frac{count(i) * unit\ Profit_i}{n}$

STEP 01

In order to determine frequent items (L_1) which are highly consumable and profitable in vender's perspective, should satisfy the following condition.

$$sup_i \geq ms_i \text{ and } prof_i \geq mp_i$$

The items are filtered based on the minimum support and minimum profit.

STEP 02

Generate the candidate set of k-itemsets (C_{k+1}) by pairing the items in L_k, $k = 1,2,3, \ldots$. Then compute the average minimum support of i^{th} and j^{th} items (ams_{ij}) and average minimum profit of i^{th} and j^{th} items (amp_{ij}) of each candidate item. So as to sort the highly consumable and profitable k-itemsets (R_k), individual support and profit of items should be greater than or equal to average minimum support and profit respectively.

$sup_i \geq ams_{ij}$ and $sup_j \geq ams_{ij}$ with $prof_i \geq amp_{ij}$ and $prof_j \geq amp_{ij}$ where $ams_{ij} = \frac{(ms_i + ms_j)}{2}$ and $amp_{ij} = \frac{(mp_i + mp_j)}{2}$

STEP 03

The sorted k-itemsets(R_k) is pruned to obtain $L_{k+1,}$by comparing the support and profit of i^{th} and j^{th} items together, with average minimum support and profit respectively as follows;

$sup_{i \cup j} \geq ams_{ij}$ and $prof_{i \cup j} \geq amp_{ij}$ where i, j are items.

STEP 04

Repeat STEP 02 and STEP 03 until $L_{k+1} = \emptyset$.

STEP 05

Construct the association rules for each k-itemset in L_k. Compute the confidence values of all association rules and compare it with the user defined confidence value λ.

OUTPUT

Association rules of frequent itemsets which are giving maximum profit to the business.

IV. RESULTS AND DISCUSSION

Suppose a supermarket tracks sales data for seven items denoted by 'A', 'B', 'C', 'D', 'E', 'F', 'G'. The obtained results by the implementation of the proposed algorithm are discussed below.

INPUT

The predefined minimum support and minimum profit values are given in Table 1.

Item	**A**	**B**	**C**	**D**	**E**	**F**	**G**
$\boldsymbol{ms_i}$	0.4	0.7	0.3	0.7	0.6	0.2	0.4
$\boldsymbol{mp_i}$	1.0	2.2	2.0	1.9	2.5	1.4	2.0

Table 1: Minimum Support (ms_i)and Minimum profit (mp_i)

Consider following set of transaction, profit margin and support value in Table 2, Table 3 and Table 4 respectively.

ID	**Transaction**
1	ABDG
2	BDE
3	ABCEF
4	BDEG
5	ABCEF
6	BEG
7	ACDE
8	BE
9	ABEF
10	ACDE

Table 2: Transaction Data

Item	**A**	**B**	**C**	**D**	**E**	**F**	**G**
prof(i)	1.2	2.4	2.4	2.0	2.7	1.2	2.1

Table 3: Profit margin of each item with respect to the transactions in Table 2

Item	**A**	**B**	**C**	**D**	**E**	**F**	**G**
sup(i)	0.6	0.8	0.4	0.5	0.9	0.3	0.3

Table 4: support of candidate items

STEP 01

Utilizing the Table 1 to Table 4, the frequent 1-itemset L_1 is generated, according to the first step of the algorithm. The result of this step is in Table 5.

$\boldsymbol{L_1}$	{ A , B , C , E }

Table 5: The frequent 1-itemset

The items obtained in this result are relatively frequent and relatively profitable as it has been filtered from both profit and support (frequency) constraints.

STEP 02

Candidate 2-itemsets (C_2) are generated from L_1. Table 6 consist of arithmetic mean on support and profit of C_2 calculated according to the formulas given in the algorithm.

C_2	AB	AC	AE	BC	BE	CE
ams_{ij}	0.55	0.35	0.5	0.5	0.65	0.45
amp_{ij}	1.6	1.5	1.75	2.1	2.35	2.25

Table 6: ams and amp of C_2

As supports of the two items in each item set in C_2 must be larger than or equal to the ams_{ij} AND profits of the two items in each item set in C_2 must be larger than or equal to the amp_{ij}, $R_2 = \{BE\}$.

This step clearly depicts the prohibition of generating unnecessary rules as it has been filtered from arithmetic mean of profit and support.

STEP 03

Since $sup_{i\cup j} > ams_{ij}$ AND $prof_{i\cup j} > amp_{ij}$ of BE 2-itemset $L_2 = \{BE\}$.

R_2	BE
$sup_{i\cup j}$	0.7
$prof_{i\cup j}$	2.55
ams_{ij}	0.65
amp_{ij}	2.35

Table 7:R_2 2-itemsets

The outcome of this step illustrates the consistency of the algorithm as it provides an itemset comprises with a maximum profit and support.

STEP 04

Since L_3 is null, the process terminates.

STEP 05

Association rules are formed for L_2.

$B \rightarrow E$ and $E \rightarrow B$ are the generated association rules.

Calculate the confidence values of the above association rules.

1. Confidence of $B \rightarrow E$ = 0.875
2. Confidence of $E \rightarrow B$ = 0.777

Predefined confidence value $\lambda = 0.8$

∴ Final result generated by the proposed algorithm is; $B \rightarrow E$

This association rule illustrates the reliability of the profit gain and the pruning capacity of the proposed algorithm.

V. CONCLUSION

The extended Apriori algorithm generates rules to arrange floors and shelves of a supermarket based on the frequent items in the transaction database which is insufficient to accomplish the requirement of the venders, maximization of profit gain. The proposed algorithm in this research consists of a profit constraint with effect from the commencement of the process so as to generate rules based on both frequent and rare items and profit of itemsets. This newly acquainted constraint enhances the profit gain in a transaction. Simultaneously this profit constraint facilitates the rare items without disturbing frequent items. It has been inspected with credible data sets and the outcomes conclude that the rules generated by the proposed algorithm heightens the profit gain while pruning unnecessarily generated rules. When negotiating with outsized real world data sets the results might vary depending on the predefined values. Therefore, the algorithm is subjected to further perfections to optimize the circumstances.

VI. REFERENCES

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules, in: Proceedings of the 20th VLDB conference, pp 487–499.

Annie M.C.L.C., Kumar D.A., 2011. Frequent Item set mining for Market Basket Data using K-Apriori algorithm, in: International Journal of Computational Intelligence and Informatics, Volume 1, No. 1, pp.14-18.

Annie M.C.L.C., Kumar D.A., 2012, Market Basket Analysis For A Supermarket Based On Frequent Itemset Mining, in: International Journal of Computer Science 9.5.

Balaji Mahesh, Rao, V.R.k., Subrahmanya, G., 2013. An Adaptive Implementation Case Study of Apriori Algorithm for a Retail Scenario, in: a Cloud Environment, ccgrid, pp.625629, 2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, 2013.

Bhandari, Akshita, Ashutosh Gupta, Debasis Das, 2015. Improvised Apriori Algorithm Using Frequent

Pattern Tree For Real Time Applications In Data Mining, in: Procedia Computer Science 46 (2015): 644-651.

Han, J., Kamber, M., 2001. Data Mining: Concepts and Techniques, in: Morgan Kaufmann Publishers, San Francisco, CA.

Han, J., Pei, H., Yin, Y., 2000. Mining Frequent Patterns without Candidate Generation, in: Proc. Conf. on the Management of Data SIGMOD'00, ACM Press, New York, NY, USA.

Kamruzzaman, S. M., Rahman, C. M., 2010. Text Categorization Using Association Rule And Naïve Bayes Classifier.

Larose, D.T., Larose, P.D.T., 2005. Discovering knowledge in data: An introduction to data mining. New York: Wiley-Interscience.

Liu, Bing, Wynne, H.s.u., Yiming Ma, 1999. Mining Association Rules With Multiple Minimum Supports. Knowledge Discovery & Data Mining (KDD-99). San Diego, in: ACM SIGKDD International Conference, 1999.

Liu, G., Huang, S., Lu, C. and Du, Y., 2014. An improved k-means algorithm based on association rules, in: International Journal of Computer Theory and Engineering, 6(2), pp. 146–149. doi: 10.7763/ijcte.2014.v6.853.

Qiang Niu, Shi-Xiong Xia, Lei, Zhang, 2009. Association Classification Based on Compactness of Rules,in: WKDD 2009, Second International Workshop on Knowledge Discovery and Data Mining, pp. 245-247.

Rao, S., Gupta, R.,2012. Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm, in: International Journal of Computer Science And Technology Mar. 2012, pp. 489-493.

Raorane A. A, Kulkarni R. V., Jitkar B.D., 2012. Association Rule - Extracting Knowledge Using Market Basket Analysis, in: Research Journal of Recent Sciences, Vo11 (2)19.

Trikha, R. and Singh, J. (2014) 'Improvement in Apriori Algorithm with New Parameters', International Journal of Science and Research, 3(9).

Tassa, T., Open, T. and Road (2014) 'Secure mining of association rules inHorizontally distributed databases', IEEE Transactions on Knowledge & Data Engineering, (4), pp. 970–983. doi: 10.1109/TKDE.2013.41.

Shah, K. and Mahajan, S. (2009) 'Maximizing the efficiency of parallel Apriori algorithm', 2009 International Conference on Advances in Recent Technologies in Communication and Computing, doi: 10.1109/artcom.2009.73.

Samaraweera, W.J., Vasanthapriyan, S. and Oza, K.S. (2011) Designing a multi-level support based association mining algorithm. Available at: http://www.ijsrp.org/research-paper-0414.php?rp=P282520 (Accessed: 2016).