Market Basket Analysis using Association Rule Learning

Nidhi Maheshwari
Computer Science and Engineering
IMS Engineering College
Ghaziabad, 201009, India

Nikhilendra K. Pandey
Computer Science and Engineering
IMS Engineering College
Ghaziabad, 201009, India

Pankaj Agarwal, PhD
Computer Science & Engineering
IMS Engineering College
Ghaziabad, 201009, India

ABSTRACT
The proposed paper focusses on the basic concepts of association rule mining and the market basket analysis of different items. In the current study, the market analysis would be done by collecting the real, primary data directly from retailers and wholesalers. The efficiency of the FP-Growth algorithm can be measured in terms of mining of the frequent pattern. Precisely, we apply FP-Growth algorithm on the various data collected from different stores in order to trace the various association rules comprising of a basket. One discrete advantage is that it avoids the generation of candidate sets, which is computationally exhaustive. The results and conclusions drawn can be used in optimizing the market. This will help in predicting future trends and behaviours, allowing businesses to make knowledge-driven decisions.

Keywords
Market Basket Analysis, Association Rule Mining, FP-Tree algorithm, Frequent Itemsets, Support, Confidence

1. INTRODUCTION
Association rule mining is one of the most important technique of data mining. The process of knowledge discovery in databases (KDD) includes selection of data, its preprocessing, transformation, data mining and interpretation. The major goal of data mining algorithms is to extract the hidden predictive information and transform it into an understandable structure [8]. It aims at extracting interesting patterns, relations, associations among sets of items in databases. One of the major task of ARM is to find the relationship among various data items in database. An association is defined in the form of A->B where A is the antecedent and B is the consequent and the meaning of the rule is deduced as: A and B, both are itemsets and the rule says that if a customer who purchases the A item are likely to purchase the B item as well with a conditional probability percentage factor known as %C where C is the confidence value of a rule. This helps the business managers to study the behaviour and buying habits of the customers in order to increase their sales. Based on this study, items that are closely related or the items that have attraction to each other are put under closed proximity. For instance, a customer who purchases milk is also likely to purchase bread together [5].

The interestingness measures such as support and confidence plays a very important role in the association rule analysis. The support value of any transaction X with respect to T is defined as the proportion of the transactions in dataset which contains itemset X [4]. It is given as: Supp(X)=X∪Y. The confidence value of a rule is defined as the proportion of the transactions that contains X which also contains Y [4]. It is given by: Conf(X=>Y) = Supp(X∪Y)/Supp(X). The itemsets that comply with minimum support and minimum confidence values are called strong association rules [5].

Currently, one of the fastest and most popular algorithm for itemset mining is FP-Growth algorithm [3]. The algorithm is based on a prefix tree representation of a stated database of a transaction. It can save considerable amounts of memory used for storing the transactions. The algorithm can be described as a recursive elimination scheme which means that in the preprocessing step, we delete all the itemsets from the transactions that are not frequent individually, i.e., delete all the items that do not appear in a user-defined minimum number of transactions. Then select all those transactions which have the least frequently occurring item among those which are frequent and also delete them. Recurse the whole process in order to obtain the minimized (also known as projected) dataset in which items found in recursion share the deleted item as a prefix. On return, all the processed items are removed from transactions and the whole process is started again. The initial FP-Tree is built from the preprocessed database and then, a FP-tree is projected which means the infrequent items are removed from the transaction database [1]. The size of the FP-tree is reduced by pruning. Finally, results are deduced on the basis of charts by doing the comparison between FP-tree algorithm [1] and Apriori algorithm [5].

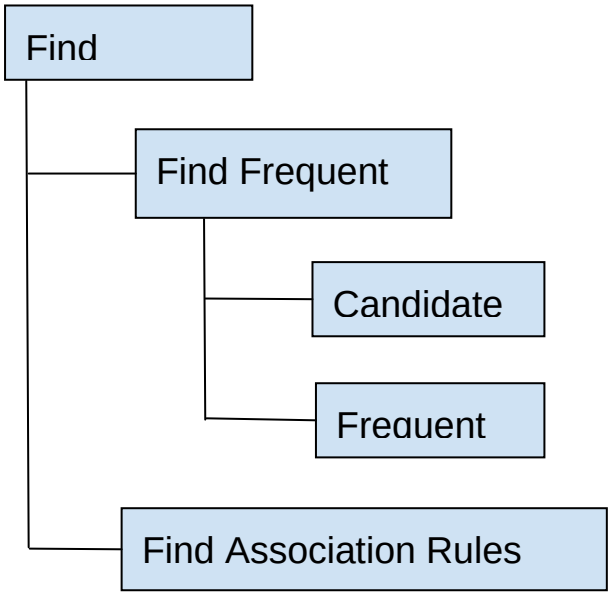

Figure 1: Generation of association rules

2. LITERATURE SURVEY
Trnka [5][9], in this paper describes the implementation of market basket analysis to Six Sigma methodology. The methods of data mining provide a great deal opportunities in the market sector. One of them is market basket analysis. By

implementing this to Six Sigma, the results can be improved and the performance level of the process can be changed.

Yanthy et al. [5][10], in this paper talks about the goal of data mining which is to reveal the hidden knowledge from provided data and the various algorithms that have been proposed so far. The interestingness of the rules can be determined by using various measures such as confidence, support, lift, information gain etc. since not all the rules generated are of interest to any given user. In this paper, he studied the relationship between interesting measures and algorithms.

Cunningham et al. [5][11], he provided a model for library circulation data and applied the Apriori tool for the task of detecting subject classification categories that co-occur in transaction records of the library borrowed books from university. The results of the paper provide insight into the degree of "scatter" that the classification scheme foster in a particular collection of documents.

Rastogi et al. [5][12], presented in his paper the optimised association approach on association rules that contain uninstantiated attributes. To determine the relationship between two items such that the support and confidence of the optimized rule is maximised. He presented effective techniques for pruning the search space while computing optimized association rules for both categorical and numerical data.

Neesha et al. [8] studied the various advancements in the field of data mining. In her paper, she described these advancements starting from year 2008, a novel frequent pattern generation algorithm had been proposed in order to tackle data imbalance problem. In 2009, an experiment was performed to compare three association rule mining algorithms: Apriori, Predictive Apriori and Tertius, on the basis of predictions made on the status of heart using heart disease data. The results of experiment showed that Apriori was best suited for this type of data.

In 2010, a new algorithm SC-BF Multilevel was introduced as a better version of Apriori algorithm which was faster and efficient since it required only single scan of database for mining frequent itemsets. In 2012, three data mining techniques were applied upon a specific set of data consisting of students' enrollment for the likeliness of the courses to be learnt. The techniques involved clustering (k-means algorithm), classification (ADTree algorithm) and association rule (Apriori algorithm). The comparison depicted that the combined approach is better than using association rule mining alone for such kind of task.

The work carried out in year 2013 involved: (1) Comparative analysis was done of three algorithms- Apriori, Predictive Apriori and Tertius. The author discussed after results, the limitations of Apriori algorithm and ways to improvise it. This showed that Apriori was faster than other two algorithms. (2) The association rule mining technique was applied on data set consisting of crimes against women. A comparison was made between Apriori and Predictive Apriori on the same data set in which Apriori was found better and faster. (3) A comparison was made among Apriori, FP-Growth and Tertius algorithm on a super-market data using Weka tool. The results depicted that FP-Growth performed better than the other two algorithms.

3. PROPOSED WORK

Market basket analysis is a technique that helps us in determining which products tends to be purchased together in accordance with the association rules. The primary objective is to improve the effectualness of sales and marketing strategies with the help of previously obtained customer data. Association rules aims to identify those items which frequently occur in a database [7]. This paper presents each item is represented by boolean value, i.e., 0 and 1, where 0 represents that item is not present whereas 1 represents that item is present. We have proposed a novel data structure, FP-Tree, frequent pattern mining, overcomes the main bottlenecks of Apriori. The frequent itemsets are generated only with two scans of the database. It is an extended prefix tree structure which is used for the storage of information about patterns. The nodes of the tree are arranged in such a way that the nodes occurring more frequently will have better chances of sharing nodes than nodes occurring less frequently. FP-Tree performs better than Apriori because there is no candidate set generation as well as the length of the frequent itemsets increases as support value decreases. FP-Growth algorithm is more efficient than latter one [2].

3.1 Preprocessing

There exists many other algorithms for mining of frequent itemsets viz., Apriori and Eclat, FP-Tree growth algorithm preprocesses the database only twice as follows: an initial scan of the database determines the frequencies of the items. All the uncommon items -- the items that do not appear in a minimum number of user-specified transactions -- are discarded from it as they can not be a part of frequent itemsets [1].

In addition to this, all the items in the transaction are sorted in descending order in reference to their frequencies. Whilst the algorithm does not depend upon the specific order of the frequencies of items, sorting in descending order may lead to much less execution time than ordered randomly. Sorting in ascending order leads to slower operations implementing even worse than in random order [1].

This preliminary processing is demonstrated in table 1. The descendingly sorted items are shown in the middle of the table. The user-specified minimal support discards all the items that are infrequent. If a user-specified support value is given as 3 then, items d and e are pruned. After doing sorting and pruning of the items, we get a new database on the right [1].

Table 1: Transaction database (left), item frequencies (middle) and reduced transaction database sorted descendingly (right) and the discarded items are shown in bold.

A B D	B=7	B A
B D	A=6	B
B C	C=6	B C
A B D	**D=3**	B A
A C	**E=1**	A C
B C		B C
A C		A C
A B C E		B A C
A B C		B A C

3.2 Fp-Tree Algorithm

The most popular algorithm for frequent itemset mining is FP-Growth algorithm. This algorithm is based on prefix tree representation of the database which saves considerable amount of memory for storing the same. The algorithm is described as a recursive elimination scheme [1]. The efficiency of FP-Tree algorithm accounts for three reasons: Firstly, it is a compressed representation of the transaction database. Secondly, this algorithm scans the database only twice. Thirdly, it uses divide-and-conquer approach which considerably reduces the size of the conditional FP-Tree [2]. After all the infrequent items have been deleted, the resultant is a FP-Tree. In this, each node corresponds to only one item and a set of transactions that share the same prefix are represented by one path [1]. The algorithm is divided into two steps [7]:

1. FP-Tree construction
2. FP-Growth

Algorithm 1: FP-Tree Construction

Input: A transaction database DB and a minimum support threshold.

Output: FP-tree, the frequent-pattern tree of DB.

Method: The FP-tree is constructed as follows:

1. Scan the transaction database DB once. Collect F, the set of frequent items, and the support of each frequent item. Sort F in support-descending order as FList, the list of frequent items.
2. Create the root of an FP-tree, T, and label it as "null". For each transaction Trans in DB do the following:

- Select the frequent items in Trans and sort them according to the order of FList. Let the sorted frequent-item list in Trans be [p | P], where p is the first element and P is the remaining list. Call insert tree([p | P], T).
- The function insert tree([p | P], T) is performed as follows: If T has a child N such that N.item-name = p.item-name then, increment N's count by 1; else create a new node N, with its count initialized to 1, its parent link linked to T, and its node-link linked to the nodes with the same item-name via the node-link structure. If P is nonempty, call insert tree(P, N) recursively.

Algorithm 2: FP-Growth

Input: A database DB, represented by FP-tree constructed according to Algorithm 1, and a minimum support threshold.

Output: The complete set of frequent patterns.

Method: call FP-growth(FP-tree, null).

Procedure FP-growth(Tree, a) {

1. if Tree contains a single prefix path then // Mining single prefix-path FP-tree {

2. let P be the single prefix-path part of Tree;

3. let Q be the multipath part with the top branching node replaced by a null root;

4. for each combination (denoted as ß) of the nodes in the path P do

5. generate pattern ß ∪ a with support = minimum support of nodes in ß;

6. Let freq pattern set(P) be the set of patterns so generated; }

7. else let Q be Tree;

8. for each item ai in Q do { // Mining multipath FP-tree

9. generate pattern ß = ai∪ a with support = ai .support;

10. constructß's conditional pattern-base and then ß's conditional FP-tree Tree ß;

11. if Tree ß ≠ Ø then

12. call FP-growth(Tree ß , ß);

13. Let freq pattern set(Q) be the set of patterns so generated; }

14. return(freq pattern set(P) ∪freq pattern set(Q) ∪ (freq pattern set(P) × freq pattern set(Q))) }

4. EXPERIMENTAL RESULTS

As a result, market basket analysis determines which sets of products tend to be purchased together. It is a tool to improve the efficiency of sales strategies by collecting data from past transactions. From the calculated values of the above mentioned dummy data, it is easily demonstrated which items should be coupled together [7].

Table 2: Frequent Patterns for Sampled Data

S. No.	Premises	Conclusion	Support	Confidence(%)
1.	A , D	A	1	10
2.	C , A , B , E	A	1	10
3.	A , B , E	B	1	10
4.	C , B , E	A	1	10
5.	B , E	B	1	10
6.	A , E	B	1	10
7.	C , E	A	1	10
8.	B , A	D	3	30
9.	B , C	D	3	30

5. CONCLUSIONS AND FUTURE SCOPE

The knowledge of what a customer or a group of customers is going to purchase can be very useful for the retailers. These results could also be helpful in determining which products appeal each other so that they can be put together in a market in order to increase the sales. For the same reason, we have proposed a novel data structure, frequent pattern tree (FP-tree), for storing compressed, crucial information about frequent patterns, and developed a pattern growth method, FP-growth, for efficient mining of frequent patterns in large databases. Using the inputs of the support and confidence values, we obtain the output in the form of association rules of the itemsets to be purchased thus deriving the patterns. This output would help in decision-making to the business organizations and determining the nature of the purchase of products. Based on the association rules, regularities between products in supermarkets are discovered [4]. This paper demonstrates a review on the association rule mining. Firstly, it talks what association rule mining is all about. It then, presents a generalized association rule mining algorithm. It also surveys the research work done by other authors in this field. While reviewing the literature, it was found that those algorithms which do not involve candidate set generation process are faster than those which involves the same [8].

In moving a step to enhance the existing system, we can fetch the data from the website database, in real time and analyze them. This will provide the real time market analysis and we can also implement Apriori algorithm so that we can analyze the data using either of the algorithm depending on the type of data.

6. REFERENCES

[1] Christian Borgelt, "An Implementation of the FP-growth Algorithm"

[2] Sotiris Kotsiantis, DimitrisKanellopoulos, "Association Rule Mining: A Recent Overview", GESTS International Transactions on Computer Science and Engineering, Vol. 32(1), 2006, pp. 71-82

[3] J. Han, H. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation, In: Proc. Conf. on the Management of Data (SIGMOD'00, Dallas, TX). ACM Press, New York, NY, USA 2000

[4] https://en.wikipedia.org/wiki/Association_rule_learning

[5] Phani Prasad, MurlidherMourya, "A Study on Market Basket Analysis Using a Data Mining Algorithm", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Vol. 3, Issue 6, June 2013

[6] Akansha Singh, K. K. Singh, Data Mining and Data Warehousing, India: Umesh Publications, 2011-2012

[7] Harpreet Kaur, Kawaljeet Singh, "Market Basket Analysis of Sports Store using Association Rules", International Journal of Recent Trends in Electrical & Electronics Engg., ISSN: 22316612, Dec. 2013

[8] Neesha Sharma, C. K. Verma, "Association Rule Mining: An Overview", IJCSC, Volume 5, Number 1, March 2014, pp.10-15, ISSN-0973-7391

[9] Trnka., "Market Basket Analysis with Data Mining Methods", International Conference on Networking and Information Technology (ICNIT), 2010

[10] W Yanthy, T. Sekiya, K. Yamaguchi, "Mining Interesting Rules by Association and Classification Algorithms", FCST 09

[11] Cunningham, S. J. and Frank, E., "Market Basket Analysis of Library Circulation Data", International Conference on Neural Information Processing, Vol. 2, 1999

[12] Rastogi, R. and Kyuseok Shim, "Market Optimised Association Rules with Categorical and Numerical Attributes", IEEE transactions on Knowledge and Data Engineering, Vol. 14, 2002