Development of Marketing Analysis System Using Association Rule Mining

San Thet Nwe, Myo Myint
Computer University of Myitkyina
santhetnwemkn@gmail.com,mmyoemyint@gmail.com

Abstract

Marketing is a societal with process which includes advertising, distribution and selling. It is also concerned with anticipating the customer's future needs and wants, which are often discovered through market research. The purpose of this paper is to identify associated itemsets, which then grouped in mix merchandise with the use of association rule mining. This association between products then will be applied in deciding about which products are the best selling itemsets in the supermarket. The process of identifying the related itemsets bought together in one transaction is done by using data mining technique. The market basket analysis is a powerful tool for the implementation of one of the strategies of marketing. Apriori algorithm is chosen as a method in the data mining process. The computational results show strong association rules and applied for marketing which items are most susceptible to promotional efforts and which items to put on sales at reduced prices.

1. Introduction

The marketing teams (marketers) are tasked to create consumer awareness of the product of services through marketing techniques. The success of marketing is influenced by its fast response and its ability in understanding consumer's behaviors. Marketing must focus to its consumer since retail business plays its role at the end of distribution channel.

Consumer buying behaviors can be comprehended by observing how someone interacts and reacts to the marketing mix. According to Wikipedia [9], company determines the decisions related to the 4P (Product, Place, Promotion, and Price) by focusing to its consumer; while each individual considers the option to buy which products under the psychological influences of culture, attitude, experience, previous usage of the products, and personal perception. Effects of both inputs (marketing and psychological) somewhat influences the customers to decide whether they will buy or not, where to buy, which bran to buy, and another choices.

Aside of observing customers reactions to marketing mix, another way for understanding customers behavior is by using historical data, which is transactional data. From customer transactional database, it can be observe customer's shopping patterns which show associated categories or even associated itemsets.

Association rule mining search for interesting relationships among items in a given data set. Mining association rules, also market basket analysis is one of the application areas of Data Mining. Mining Association Rules has been first introduction in [1]. The objectives of this study were to discover the associate itemsets, and to determine which items are most susceptible to promotional efforts.

The remainder of the paper is organized as follows. Chapter 2 examines the concepts of association rule mining. Chapter 3 discusses the association rule discovery for marketing analysis. Chapter 4 explains system design and implementation of the system. Finally, Chapter 5 concludes the paper.

2. Related works

Association rule mining raised by Rakesh Agrawal is an important research problem in data mining field. In essence, association rule mining is to find the rule sets satisfying minimum support threshold and minimum confidence threshold in the dataset or database [5]. The Apriori algorithm performs as many passes over the data as the size of the itemsets. Apriori algorithm to improve the effectiveness of the various studies, is mainly towards reducing the amount of computation and by less the number of scanning the database to improve, then introduced a method of improving the algorithm Apriori: ApriofiMend algorithm [1].

The Apriori algorithm [5] takes all of the transactions in the database into account in order to define the market basket. The market basket analysis is a powerful tool for the implementation of cross-selling strategies. Especially in retailing it is essential to discover large basket, since it deals with thousands of items. In Luis Cavique [3], the condensed data is used and is obtained by transforming the market basket problem into a maximum-weighted clique problem.

3. Background theory

3.1. Mining association rule discovery

Association rule mining finds interesting association or correlation relationships among a large set of data items. The discovery of interesting association relationships among huge amounts of business transaction records can help in many business decision making process.

This study used marketing analysis to find association rules between sets of items in transactional database and the Apriori algorithm is used in the data mining process. The goal of the association discovery was to find items that imply the presence of other items. For example, 90% of customers that purchase frozen pizza also buy sodo. The Apriori Algorithm [7] takes all of the transactions in the database into account in order to define the market basket. The market basket can be represented with association rules, with a left and a right site Left⇒Right. For given an itemset {A, B, C} the rule {B, C} ⇒{A} should be read as follows: if a customer bought {B, C} he would probably buy {A}. This approach was initially used in pattern recognition and it became popular with the discovery of the following rule: "on Thursdays, grocery store customers often purchase diapers and beer together" [4].The problem of finding association rules was first introduced by Agrawal, et. al. [6].Association rule mining is a two-step process:

1. Frequent Itemset Generation
2. Rule Generation

The computational requirements for frequent itemset generation are generally more expensive than those of rule generation.

3.1.1. Rule Measure : Support and Confidence

Rule support and confidence are two measures of rule interestingness. In general, each measure is associated with a threshold that can be controlled by the user or domain experts.

The support of an association pattern refers to the percentage of task-relevant data tuples (or transactions) for which the pattern is true. Support is an important measure because a rule that has very low support may occur simply by chance. Support is often used to eliminate uninteresting rules.

A certainty measures for association rules is confidence. Confidence on the other hand, measures the reliability of the inference made by a rule. Confidence also provides an estimate of the conditional probability of Y given X.

$$\text{Support } (X \rightarrow Y) = P(X \cup Y) \quad (1)$$

$$\text{Confidence } (X \rightarrow Y) = P(Y/X) \quad (2)$$

3.2. The apriori algorithm

Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules. Apriori Algorithm was also discussed by Agrawl et. al. [8], which considered as one of the most contributions to this subject. Its main algorithm, Apriori, has affected not only the association rule mining community, but other data mining fields as well. The Apriori algorithm for finding all large item sets makes multiple passes over the database. In the first pass, the algorithm counts item occurrences to determine large 1-item sets. The subsequent pass, say pass k, consist of two steps. First, the large item sets L_{k-1} found in the (k-1)-th pass are used to generate the candidate item setsC_k. Then, all those item sets which have some (k-1) subset that is not in L_{k-1} are deleted, yielding C_k. Figure 1 gives the Apriori algorithm as described by Agrawal et. al. [2]. In the second step, the Apriori algorithm generates sets of large frequent itemsets and then generates association rules Left⇒Right. For each rule, the support measure and the confidence measure are calculated.

The outputs of the Apriori algorithm are easy to understand and many new patterns can be identified. However, the sheer number of association rules may make the interpretation of the results difficult. A second weakness of the algorithm is the computational times when it searches for large itemsets, due to the exponential complexity of the algorithm.

```
1)  L1={large 1-item sets};
2)  For (k = 2; Lk-1 ≠ ; k++) do begin
3)    Ck = apriori-gen(Lk-1);//New candidates
4)  forall transactions t   D do begin
5)    C1 = subset(Ck, t);//Candidates contained in t
6)    forall candidates c   C1 do
7)          c.count++;
8)  end
9)  Lk={c   Ck| c.count≥minsup}
10) end
11) Answer= kLk;
```

Figure 1. Apriori algorithm

3.3. The marketing analysis system

Marketing tends to seen as a creative industry, which includes advertising, distribution and selling. It is also concerned with anticipating the customers' future needs and wants, which are often discovered

through market research. The marketing teams (marketers) are tasked to create consumer awareness of the product of services through marketing techniques. For a marketing plan to be successful, the mix of the four "Ps" must reflect the wants and desires of the consumers or shoppers in the target market. Marketers depend on insights from marketing research, both formal and informal, to determine what consumers want and what they are willing to pay for.

4. System design and implementation

4.1. System design

This system is designed to suggest new products to frequent customers based on previous purchase patterns and effective used for marketing analysis. In this design, the marketing analysis system presents three main processes. The first step was to choose which items to be included in this analysis. The second is items information process for transaction database D. In this process, collected transactions are stored in database to use in market analysis. The last process is implemented for testing. In this section, the user can test step-by-step the marketing analysis system according to the Apriori algorithm. As a result of this section, the system shows the strong association rules with good scalability properties. The following figure2 is the system design for our system.

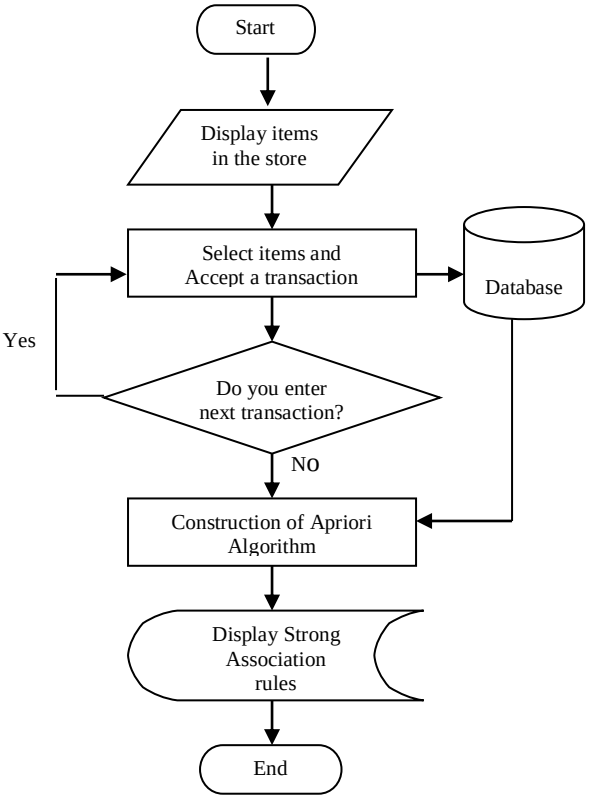

Figure 2. The system design for our system

4.1.1. The definition of the problem and result. The input for the marketing analysis is a dataset of purchases. A market basket is composed of items bought together in a single trip to a store. The most significant attributes are the transaction identification and item identification. While ignoring the quantity bought and the price. Each transaction represents a purchase, which occurred a specific voucher number and items ID.

The dataset with multiple transactions can be shown in a relational table (transaction, item). Corresponding to each attribute there is a set called domain. The table (transaction, item) is a set of all transactions $T= \{T_1, T_2, T_3, \ldots, T_n\}$ where each transaction contains a subset of items $T_k= \{I_a , I_b , I_c \ldots\}$.

To exemplify this problem, an instance with 10 items and 9 transactions is given in table 1 and 2. The domain (item) is equal to {BML ,CM ,CF ,HC ,ES ,LC ,LP ,MS ,NC ,Q-10}and the domain(transaction) is equal to {v001 ,v002 , v003 ,v004 ,v005 ,voo6 ,voo7 ,voo8 ,coo9}. Initial processing of the 9 transactions using Apriori algorithm with minimum support of 35% (count 3) and minimum confidence of 70%.

Table 1. Items categories

NO	ITEM ID	ITEM NAME
1	BML	Body Moisturizing Milk Lotion
2	CM	Cleansing Milk
3	CF	Cream Puff
4	ES	Eye Shadow
5	HC	Hair Color
6	LC	Lip Color
7	LP	Lucent Powder
8	MS	Milk Shampoo
9	NC	Nail Color
10	Q-10	Q-10 Day Cream
11	RO	Rollon
12	SB	Scrub
13	SC	Shower Cream
14	SG	Styling Gel
15	TR	Toner

Table 2. Sample data for the marketing analysis

Voucher No	Customer Name	Item	Purchase Date
V001	Ma Thet	{BML,CF,ES,LC,MS,NC}	3/5/09
V002	Su Su	{CM,CF,ES,LP,NC,Q-10}	3/5/09
V003	Hla Hla	{BML,CM,CF}	4/5/09
V004	Nilar	{CM,CF,ES,LP,NC}	4/5/09
V005	Thin Thin	{HC,LC,LP}	5/5/09
V006	Khin Thuzar	{CF,LC,LP,MS,Q-10}	5/5/09
V007	Seng Ra	{BML,CM,ES,HC,LC,NC}	6/5/09
V008	Hnin Mon	{BML,CF,ES,LC,LP}	6/5/09
V009	Htu Ra	{CM,HC,LP,MS,Q-10}	7/5/09

The following study gives the results of 4 frequent itemsets and 15 strong association rules as shown in table 3. The rule suggested that a strong relationship exists between the scales of products. The user can use this type of rules to help them for new opportunities. For example, a confidence of 70% for the association rule

$$(\text{BML, ES}) \Rightarrow \text{LC}$$

means that 70% of all customers who purchased body moisturizing milk lotion and eye shadow also bought lip color. So the user can estimate of the conditional probability of their market.

Table 3. Strong association rules extracted from the sample data

Strong Association rules	Confidence
(BML , ES) ⇒ LC	100%
(BML , LC) ⇒ ES	100%
(ES , LC) ⇒ BML	100%
BML ⇒ (ES , LC)	75%
(CM , ES) ⇒ NC	100%
(CM , NC) ⇒ ES	100%
(ES , NC) ⇒ CM	75%
NC ⇒ (CM , ES)	75%
(CF , ES) ⇒ LP	75%
(CF , LP) ⇒ ES	75%
(ES , LP) ⇒ CF	100%
(CF , ES) ⇒ NC	75%
(CF , NC) ⇒ ES	100%
(ES , NC) ⇒ CF	75%
NC ⇒ (CF , ES)	75%

Support is an important measure because a rule that has very low support may occur simply by chance. Thus it may not be profitable to promote items. Figure3 shows the effect of support threshold on the number of frequent itemsets.

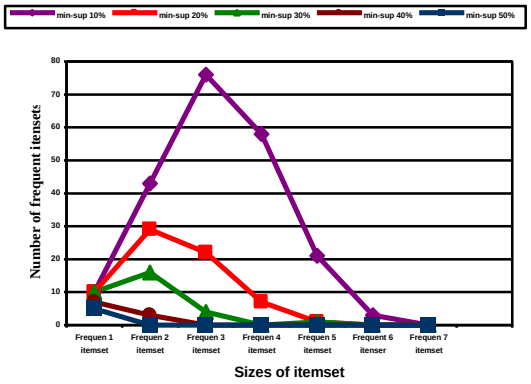

Figure 3. Effect of support threshold on the number of frequent itemsets.

4.2. Implementation of the system

When the user clicks the Items Info tab, it will display the available items list form as figure 4. If you want to see the details information of an item, first, you may select the item in table and then press enter key. The system shows the user the information of the selected item. The user can insert the new item and can update the exiting item by using new button and save button. In this way, the user can use the item information form easily by using the system.

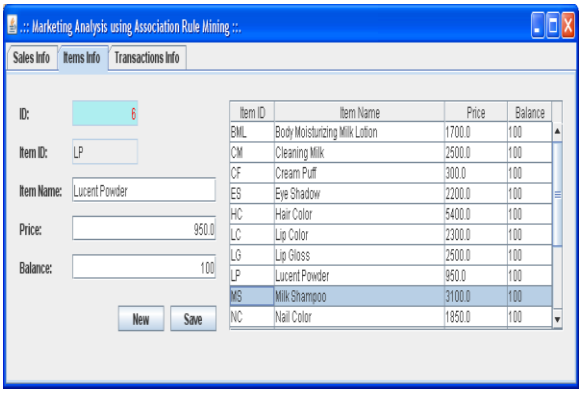

Figure 4. Items information form

When the user clicks the Sales Info tab, it will display sales information form as figure 5. This form can be use to enter the sale transactions list into the system. When the user wants to entry the new transaction to the system, it can be used by pressing the new button and save button. If the user wants to update the sales transaction, the user enters a voucher number and press the enter key. Then the user can do the changes the items and deleting the transaction by pressing the delete button.

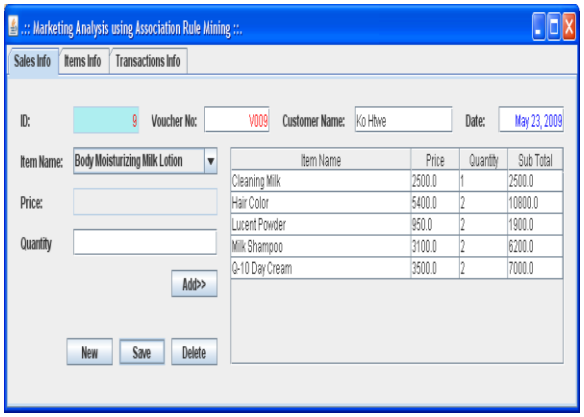

Figure 5. Sales information form

When the user clicks the Transactions Info tab, the form appears as figure6, figure7 and figure 8. In this form, the user can see the all sales transactions list and can start generate the Apriori Algorithm. The user must enter min_support count and min_confidence, and then press the start button. After pressing the start button, the system shows the step by step frequent itemsets generation and rules generation.

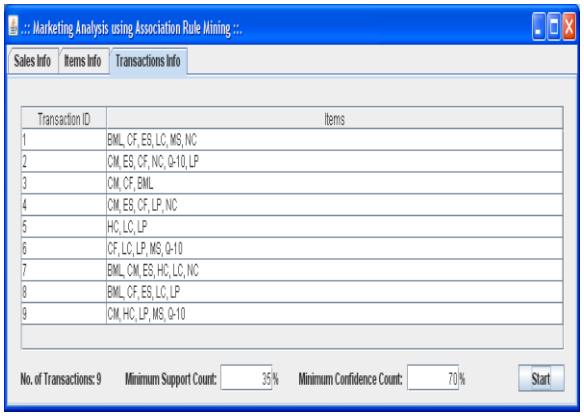

Figure 6. Transactions information form

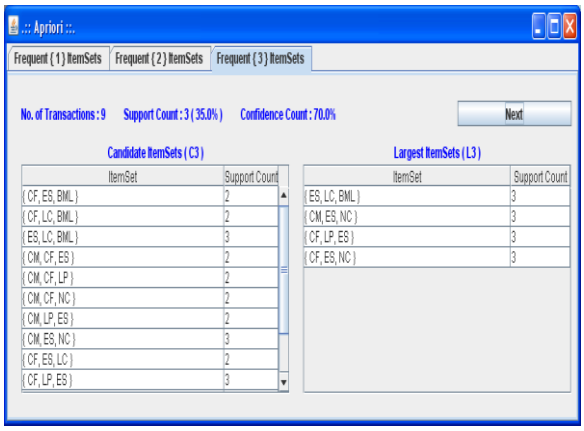

Figure 7. Largest frequent itemsets form

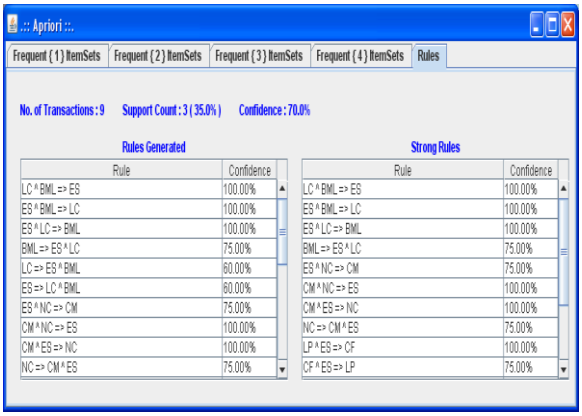

Figure 8. Strong rules form

5. Conclusion

This system implements the marketing analysis, can be applied to find items that are frequently bought together by customers. Using historical data of customers' shopping behaviors, this study has discovered several shopping patterns in the supermarket. This discovered patterns are typically represented in the form of implication rules or features subsets. This system, the Apriori algorithm generates sets of frequent itemsets and then generates strong association rules on market transaction. According to the first iteration of the algorithm, the system scans all of the transactions in order to count the number of occurrences of each item. So the marketing manager or user can know which products are the best selling in the supermarket and which are low value consumer products. Input parameters are defined as the minimum support (minsup) and the minimum confidence (minconf) by the user. The rule suggested that a strong relationship exists between the scales of products. The user can use this type of rules to help them for new opportunities. Based on these system can be extended as online marketing transaction analysis system.

References

[1] H.Feng, Z.Shu-mao, DU .Yinh-shuang," *The analysis and improvement of Apriori algorithm*", Journal of Communication and Computer, vol5,no.9(Serial no.46), Sep.2008.

[2] H. Jiawei and K. Micheline, "*Data Mining Concepts and Technique*", ISBN 1-55860-489-8.

[3] Luis Cavique, "*A Scalable Algorithm for the Market Basket Analysis*", ESCS, Instituto Politeconico de Lisboa, Purtugal.

[4] M.Berry and G. Linoff, "*Data Mining Techniques for Marketing, sales and Customer Support*", John Wiley and Sons, 1997.

[5] N.Hung Son, "*Transaction Data Analysis and Association Rules*", http://www.mimuw.edu.pl/~Son/datamining.

[6] P.N.Tan, V.Kumar and M.Steinbach, "*Introduction to Data Mining*".

[7] R.Agrawal, T.Imielinski, and A.Swami, "*Mining Association Rules Between Sets of Items in Large Database*", 1993.

[8] R.Agrawal and R. Srikant, "*Fast algorithms for Mining Association Rules in Large Database*" In Processing of the 20th International Conference on Very Large Databases,pp.407-419,Santiago, Chile, 1994.

[9] http://en.wikipeia.org//wiki/Marketing