Datamining Application in Retailing

***Mrs. P. SRIDEVI - **Dr. V. J. SIVAKUMAR - ***Mrs. M.HEMALATHA**

Abstract

Retail *is identified as a sector that is poised to show highest growth in the coming years. The sector is set for a revolution by both present players & new entrants to explore the market. It is expected to grow by 25 – 30% by 2010. As, business has gone more customer – centric, this paper investigates Market Basket Analysis as an important component of analytical CRM in retail organizations. Selling as many different products as possible to customers maximizes the value to business.* ***Market Basket Analysis*** *(MBA) applies association of rules learning to purchase data with the goal of identifying cross selling opportunities. Given a data set, the algorithm trains and identifies product baskets and product association rules. Product baskets (referred to as* ***item sets****) are groups of products purchased together at checkout. Product association rules predict the purchase of one or more other products (the* ***consequent****) given the known presence of some products in a basket (the* ***antecedent****).So Market Basket Analysis (MBA) is used as a technique to make retailers understand what combinations of product items, customer may tend to purchase at the same time or later on as a follow-up purchase and thereby giving an idea to retailers which product items sell together. Therefore, this paper has focused on how MBA seeks to find relationship between purchases in a departmental store in Trichy region by formation of rules like* ***IF*** *{Detergent powder, No Detergent bar}* ***THEN*** *{Brush}, for the purchase items taken into analysis. This technique acts as an excellent cross-selling promotional measure in retail segment.* ***Key words****: Market Basket Analysis (MBA), Retail,*

* Lecturer, National Institute of Technology - Trichy
** Assistant Professor, National Institute of Technology - Trichy
*** Research Scholar, National Institute of Technology - Trichy

Rules, Product items, Cross selling

1. Introduction:

1.1Market Basket Analysis:

Market Basket Analysis is a modeling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items. For example, if you are in a retail outlet and buy detergent powder and don't buy a detergent bar, you are more likely to buy brush at the same time than somebody who didn't buy detergent bar.

The set of items a customer buys is referred to as an **item set**, and market basket analysis seeks to find relationships between purchases. Typically the relationship will be in the form of a rule:

IF {detergent powder, no detergent bar} THEN {Brush}.

The probability that a customer will buy detergent powder without a bar (i.e. that the antecedent is true) is referred to as the **support** for the rule. The conditional probability that a customer will purchase brush is referred to as the **confidence**.

The complexities mainly arise in exploiting taxonomies, avoiding combinatorial explosions (a supermarket may stock 10,000 or more line items), and dealing with the large amounts of transaction data that may be available.

A major difficulty is that a large number of the rules found may be trivial for anyone familiar with the business. Although the volume of data has been reduced, we are still asking the user to find a needle in a haystack. Requiring rules to have a high minimum support level and high confidence level risks missing any exploitable result we might have found. One partial solution to this problem is *differential market basket analysis*, as described below.

1.2 How is it used?

In retailing, *most purchases are bought on impulse*. Market basket analysis gives clues as to what a customer might have bought *if the idea had occurred to them*. As a first step, therefore, market basket analysis can be used in deciding the location and promotion of goods inside a store. If, as has been observed, purchasers of Barbie dolls have are more likely to buy candy, then high-margin candy can be placed near to the Barbie doll display. Customers who would have bought candy with their Barbie dolls *had they thought of it* will now be suitably tempted.

But this is only the first level of analysis. **Differential market basket analysis** can find interesting results and can also eliminate the problem of a potentially high volume of trivial results.

In differential analysis, we compare results between different stores, between customers in different demographic groups, between different days of the week, different seasons of the year, etc.

If we observe that a rule holds in one

store, but not in any other (or does not hold in one store, but holds in all others), then we know that there is something interesting about that store. Perhaps its clientele are different, or perhaps it has organized its displays in a novel and more lucrative way. Investigating such differences may yield useful insights which will improve company sales.

2. ASSOCIATION and SEQUENCING:

Selling as many different products as possible to your customers maximizes their value to your business. One way to accomplish this goal is to understand what products or services customers tend to purchase at the same time, or later on as follow-up purchases. Determining purchasing trends is a very common application of data mining, and association and sequencing techniques can perform this kind of analysis. Although originally devised for marketing purposes, these techniques also have important applications in medicine, finance, and other data rich environments where separate events might be related to each other and where knowing about such relationships can be valuable knowledge.

Association and sequencing tools analyze data to discover rules that identify patterns of behavior. An association tool will find rules such as:

When people buy Vegetables they also buy Noodles 50 percent of the time.It is highly unlikely that this rule is true. In fact, the oft-cited correlation between sales of vegetable and noodles is probably a myth. However, it is convenient to use it for illustrative purposes.

A sequencing technique is very similar to an association technique, but it adds time to the analysis and produces rules such as:

People who have purchased a camcorder are three times more likely to purchase a burner in the time period two to four months after the camcorder was purchased.

Using an association or sequencing algorithm to find the kinds of rules we've just seen is frequently called market basket analysis. Because such analysis has become the primary use of the association technique, resulting it as frequently called a market basket technique.

Business managers or analysts can use a market basket analysis to plan:

Couponing and discounting. It is probably not a good idea to offer simultaneous discounts on Noodles and Egg if they tend to be bought together. Instead, discount one to pull in sales of the other.

Product placement. Place products that have a strong purchasing relationship close together to take advantage of the natural correlation between the products. Alternatively, place such products far apart to increase traffic past other items.

Timing and cross-marketing. For example,

assume that a sequencing analysis has produced the camcorder/burner rule described above. Clearly this suggests that mailing a burner promotion to camcorder purchasers is best done so that it will arrive in their mailbox approximately two to three months after the camcorder purchase.

Although most commonly used for market basket analysis, association and sequencing tools have useful applications in many other industries besides retail. Association and sequencing tools find patterns in transaction data, and many organizations capture transactional data. Understanding the patterns of behavior or activity can provide valuable insight.

In health care, there are possible applications in care management, procedure interactions and pharmaceutical interactions. Consider, for example, the following statements that might result from an analysis. Their possible application should be immediately obvious.

Patients who are taking drugs A, B, and C are two and a half times more likely to also be taking drug D.Patients receiving procedure X from Doctor Y are three times less likely to get infection Z.

There are many applications of association and sequencing in the financial industry. As with the medical applications, an example is worth a thousand words. (And to someone it may even be worth a million bucks!)

The prices of stocks in industry Q are 1.8 times more likely to close up one day after stocks in industry R closed down.

3. RULE FORMULATION AND ANALYSIS:

Each rule has a *left-hand side* (when people buy Vegetables) and a *right-hand side* (they also buy Noodles). Sometimes the left-hand side is called the *antecedent* and the right-hand side the *consequent*. In general, both the left-hand side and the right-hand side can contain multiple items, but for simplicity weýll stick with single items for now.

A rule has two measures, called *confidence* and *support*. Some products such, as MineSet from Silicon Graphics Inc., use the terms *predictability* and *prevalence* instead of confidence and support. Let us see what these terms mean and how they are computed,

Support (or prevalence) measures how often items occur together, as a percentage of the total transactions. Support is not dependent on the direction (or implication) of the rule; it is only dependent on the set of items in the rule.

Confidence (or predictability) measures how much a particular item is dependent on another.

In the absence of any knowledge about what else was bought, can also be ascertained from **expected confidence** by taking the ratio of particular item transaction by total number of

transactions.

Lift measures the difference between the confidence of a rule and the expected confidence. We can measure this difference either by subtracting the two values or, more commonly, by putting them a ratio. Lift is one measure of the strength of an effect.

3.1 PREPARING THE DATA:

The data that is used by an association algorithm needs to be in one of two formats that we'll call *horizontal* and *vertical*. In the horizontal format (see Table.1) there is one row for each entity, and there are columns for each attribute. For market basket analysis with the horizontal format, there is one row for each market basket, with columns for each (type of) product.

A significant problem for the horizontal format is that the number of columns can become quite large. For market basket analysis, where the number of products might exceed 100,000, similar products need to be grouped together to reduce the number of columns to a reasonable quantity. Another problem with the horizontal format is that the schema is data dependent. When a new product is added to the market basket analysis, or when products are categorized in a different way, then the schema needs to be changed to add or reorganize columns.

ID	Vegetables	Noodles	Egg	Grocery
1200	Yes			
1201	Yes	Yes		
1202				Yes

Table 1. Data in horizontal format.

The vertical format (seeTable 2), which is more commonly used by the data mining products, eliminates these problems by using multiple rows to store an entity, using one row for each attribute. The rows for a particular entity (that is, a market basket or a patient) are tied together with a common ID. This kind of representation is more normalized in the relational sense, and it works much better when an entity can have great variability in terms of the number of attributes. For example, some people check out of the supermarket with only two or three items when others fill two carts with hundreds of items.

ID	PRODUCT
121	Vegetables
122	Noodles
123	Egg

Table 2. Data in vertical format.

Because data may already exist in one format or the other, some of the products, such as IBM's Intelligent Miner, support a pivoting operation that converts a horizontal format to a vertical format.

Association algorithms can only operate on categorical data. If you use noncategorical attributes, such as income of the purchaser, the noncategorical data must be binned into ranges (for example, 0 to 20,000; 20,001 to 40,000; 40,001 to 70,000; and greater than 70,001), turning each range into an attribute.

As an aside, how many combinations would there be in a typical transaction? Knowing this number will give us an idea of how many counters need to be updated for each transaction. We can use the formula that tells us how many combinations there are of "*n* things taken *m* at a time. . $\frac{n!}{m!(n-m)!}$

TABLE	
Total Transaction	**3500**
Vegetable transaction	970
Noodles Transaction	80
Egg Transaction	55
Vegetable & Noodles	40
Noodles & Egg	12
Vegetables & Egg	15
Vegetables, Noodles & Egg	8

Table 3.1: Transaction Table

3.2 INTERPRETATION:

RULES		EXPECTED	CONFIDENCE	LIFT	SUPPORT
IF	THEN	CONFIDENCE			
Vegetable	Noodles	2.285714286	4.12	1.80	1.14
Noodles	Vegetable	27.71428571	50.00	1.80	1.14
Vegetable	Egg	1.571428571	1.55	0.98	0.43
Egg	Vegetable	27.71428571	27.27	0.98	0.43
Noodles	Egg	1.571428571	15.00	9.55	0.34
Egg	Noodles	2.285714286	21.82	9.55	0.34
Vegetable& Noodles	Egg	1.571428571	20.00	12.73	0.23
Vegetable& Egg	Noodles	2.285714286	53.33	23.33	0.23
Noodles& Egg	Vegetable	27.71428571	66.67	2.41	0.23

Egg

Table 3.2: Interpretation of rules for Market Basket Analysis

In general, analysts are looking for rules that have a very high or very low lift, that have support, that exceeds a threshold, and that do not involve items that appear on most transactions. The highest lift of 23.3 (from table 3.2) means that people who purchase vegetable and egg are 23.3 times more likely to purchase noodles than people who do not purchase vegetable and egg. The negative lifts of 0.98 (less than 1) means that people who buy vegetables / Egg are less likely to purchase Egg / vegetables.

From support one can determine how often items occur together as a proportion of total transactions. Here we can infer from the table that vegetables and noodles combinational purchase seems to be more than the rest of the combinations. So, by this way a retailer can understand where noodles can be placed in the shop to maximize the sales. Hence, Market basket analysis can be used as an effective tool to make analysis of all product items for deciding about the location and promotion of products.

4. CONCLUSION:

Association and sequencing techniques can help companies identify groups of products or services that customers have already demonstrated a tendency to acquire together, or in subsequent purchases. The infamous but probably apocryphal association between vegetable and noodles purchases is a good example of a retail application of association and sequencing. But detecting and assessing relevant patterns can benefit almost any business that accumulates large volumes of transactions.

In retailing product management process is a workable balance between the product mix target customers seek and the retailer's financial resources. If customers do not find the products they want, the store will lose the patronage. Maintaining an overabundance of merchandise that fails to turn over on the sales which is a serious threat to the firm's survival.

Similarly the number of units to be ordered can be decided on the basis of the proportion of buying behaviour of the consumers which need a proper Market Basket Analysis. Often retailer provide price discounts to reduce inventory, stimulates short term sales and to encourage new user. This MBA helps the retailer to identify the best combination to offer sales discount.

As retailers are interested in analyzing the data to learn about the purchasing behavior of their customers, such valuable information can be used to support a variety of business-related applications such as marketing

promotions, inventory management, and customer relationship management. Retailers can also use this type of rules to help them identify new opportunities for cross-selling their products to the customers.

References:

- Agrawal R., Imielinski T., and Swami A., "Mining association rules between sets of items in very large databases", Proceedings of the ACM SIGMOD Conference on Management of data, pages 207-216, Washington D. C., 1993
- Leonardo E. Auslender, " Online analytical tools for Market Basket Analysis", SAS institute, presented by NYC Informs, New York city, May 2004
- Levy, M., Weitz, B., Retailing Management. 3rd. Irwin / McGraw-Hill, 1998
- Nong Ye, "The Handbook of Data Mining", Lawrence Erlbaum Associates publishers, London, 2003
- Pang-Ning, Michael Steinbach, Vipin Kumar, "Introduction to data mining: Association analysis", Feb 2006
- Walters, R.G., "Assessing the impact of retail price promotions on product substitution, complementary purchase, and interstore sales displacement", Journal of Marketing. 1991