Prospects Identification Scheme for Supermarkets by Classification of Customer Behavior Using Time Based Analysis of Transactional Data

Adil Mehmood Khan, Faraz Idriss Khan, Sungyoung Lee, Young-Koo Lee
Department of Computer Engineering, Kyung Hee University
Sochen-ri, Giheung-eup, Yongin-si, Gyeonggi-do, 449-701, South Korea
{adil, sylee}@oslab.khu.ac.kr, {faraz, yklee}@khu.ac.kr

Abstract

Prospecting means "finding new customers". It demands effort and resources. Traditional prospecting techniques involved surveying. During surveys raw data is generated which must be cleaned before usage. This cleaning is an overhead. Other techniques require customer demographics. Unfortunately supermarket databases lack customer demographics and acquiring these demographics by any mean is very expensive. So these traditional techniques cannot be used to perform prospecting for supermarkets. Hence our aim is to devise a way that will assist supermarket analysts in identifying future prospects by analyzing the semantics hidden in the customer's past transactional records without using any customer demographics and without conducting any surveys. We will classify the customers on the basis of date and time and then identify the prospective customer classes.

1. Introduction

Supermarkets are a great place to reach a mass audience. Non-food items now make up a third of supermarket sales. It is therefore not surprising that consumers have embraced the ability to purchase other items whilst in store and are spending more time in supermarkets as a result. But every year thousands of new products are launched and despite their support by large scale advertising and promotional campaigns 75% fail [1]. These failure rates can be reduced by taking more informed decisions before committing huge costs required to launch these products.

Supermarket analysts should focus on prospecting. They should start with asking the critical questions like "If we have the opportunity, with whom do we want to do business? ". The answer is to identify their prospective customers by analyzing their historical transactional data. Prospecting means to find new customers. Traditional prospecting techniques involve customer surveys that require considerable effort and result in huge amount of data. Since this data is raw, its analysis offers an overhead of cleaning the data. Another mean to identify the prospects is to analyze the existing customer's behavior, present in the form of customer's transaction history. Prospect analysis on these records require customer demographics, so that customers can be classified on their basis, but environments like supermarkets lack these demographics and acquiring them from external means is very expensive.

Thus our aim is to devise a way to first classify the customers in the absence of their demographics using some other dimension and then identify the prospective customer class using the buying history of that class. In order to do so this paper presents a new technique for identifying the prospects for the new products by analyzing the buying behavior of the customers for the older products. Since the most ideal customers for any new product are none other but the most profitable customers for similar previous products. Customers buying behavior will be classified on the basis of the time they perform these transactions and then the buying behavior of each class for the matched products will be analyzed. Depending upon this behavior, prospective class will be identified.

The paper is organized as follows. Second section presents the literature review. Third section deals with the detailed description of proposed method. In the fourth section some experimental results are presented and in the last section conclusion of the whole proposed system is presented.

2. Related Work

Planning a better promotional campaign for the sale of a new item requires its prospective customers to be identified, so that the advertising campaign can be altered according to their needs and it also helps in saving cost and resources. Frank M. Thiesing in [2] presented a technique of using artificial neural networks for predicting a short term forecast of the sale of articles in supermarkets but its focus was not to identify the prospective customers for a new product. Many techniques use Association Rule Mining to identify the hidden rules in the data called Association Rules [3, 4, 5]. These rules identify the products that are frequently bought by customers and set of products whose sale is affected by the sale of another set of products. But the

problem with this approach is that Association Rule Mining generates a lot of rules and it becomes difficult for us to tell the machines, which of them are significant and which are insignificant and also Association Rule Mining assumes the whole transaction set as a single cluster. Decision Trees have also been good candidates for such analysis [6, 7], but they require training data to train them so they demand a deep understanding of the data on user behalf.

Though many techniques of prospecting involved clustering but in all of them clustering is being carried out on high volume of customer's demographics [8]. Thus in order to perform clustering on a data that lacks customer demographics, we have to look for some other dimension. Most common transaction attributes in any Business environment are Customer, Products and Time of the transaction. Since we are assuming that there is no mean through which we can identify the customer performing the transaction, we cannot include the Customer attribute in this analysis. We will now explore how the other two attributes can be used in finding the prospects for the new products.

3. Prospect Identification

Though we do not have customer demographics but we have customers buying behavior in the form of transactional records. We can classify this behavior. Classifying the customer behavior will ultimately classify the customers. The key idea behind our proposed method is based on two elements of customer psychology. The author in [9, 10] highlighted these observations as

Observation 1: The most ideal customers for any new product are the one that had shown interest in similar products in past.

Observation 2: Different customers like to visit supermarkets at different part of the week and also different part of the day.

Suppose a supermarket is planning to place a new type of a Malt Beer on its shelves. Everybody coming to supermarket does not buy Malt Beer, so the management is interested in finding the prospective customers for this beer in order to shapeup their promotional campaigns. Suppose supermarket is currently selling some existing types of malt Beer. If by any mean we can identify and classify the customers who are buying these existing types of beer, they will be the ultimate prospective customers for the new beer. Since we lack customer demographics so for classifying the customer behavior we will use the time when these customers prefer to perform shopping. We will classify the transaction data into four classes i.e. Weekday-Day, Weekday-Night, Weekend-Day, and Weekend-Night. For each class we will analyze the buying behavior for the previously selling similar products and decide the prospective class for the new product. Fig.1 shows the block diagram of the proposed method.

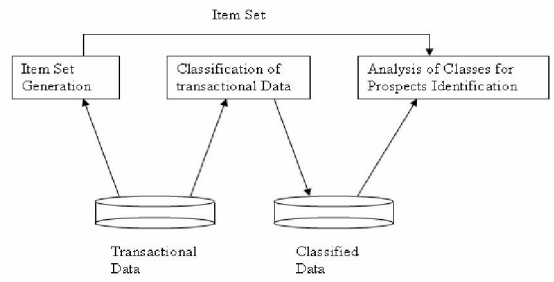

Figure 1. Block Diagram of the proposed system

3.1. Item Set Generation

This step involves two tasks; first it matches the attributes of the new product with the existing products to find a set of matched products X, choice of these attributes mainly depends upon the business environment, examples include type of the product, ingredients etc. Then transactional data is analysed to determine another set of products Y, where Y consists of products which are not similar to new product but being frequently bought along with the products of set X. This task is essential as in order to identify the prospects in this way we have to identify all the hidden patterns in data. We cannot decide the prospects only on the basis of matched similar products as there can be a case that the customers who mainly visit the supermarket on weekend nights show interest in buying beer and children dippers, which apparently are two different products. The stepwise description of this component is as follows.

Step 1: Match the attributes of the new product with the existing products to generate a set of matched similar products X, such that

$$\mathbf{X} = \{ X_1, X_2, X_3, ..., X_n \} \quad (1)$$

Where ***n*** is the total numbers of matched products.

Specify a threshold value Th_{count} for the number of transaction records that should be considered to identify the frequent item set **Y**. Since there will be huge amount of transactions involving products in **X**, so to reduce the complexity, Th_{count} puts an upper limit on the number of records. Transactions involving items of **X** will be selected, starting from the latest and moving to the older transactions until,

$$\text{Record\_Count} \leq Th_{count} \quad (2)$$

Thus we have a set of transactions **D** such that

$$\mathbf{D} = \{ t_1, t_2, t_3, ..., t_{Th_{count}} \} \quad (3)$$

Let us assume a set **Y** which is initially empty, set **Y** will consists of frequent items along with their frequencies, number of times they appear with the items of **X** in transaction set **D**.

1. For every $X_i \in X$
2. For every $t_i \in D$
3. Find K (set of products in t_i which are not in X)
4. If $\{X_{i-1}, \ldots, X_1\}$ are not in t_i
5. For every $K_i \in K$
6. If K_i is already in Y
7. Increment is frequency in Y
8. Otherwise
9. Insert the item in Y and increment its frequency to 1

Step 2: Since set **Y** can contain n items with frequencies ranging from 1 to Th_{count} . Percentage of each item in transaction set **D** is calculated as

$$\text{Percentage} = \frac{Item\ Frequency}{Th_{count}} \quad (4)$$

The items with lower percentages can be ignored, so we specify a threshold for percentages, Th_{per} .

1. For every $Y_i \in Y$
2. If percentage of Y_i is less than Th_{per}
3. Remove it from Y

Finally generate an item set **A** such that

$$\mathbf{A = X\ U\ Y} \quad (5)$$

3.2. Classifying the transactional data

We have decided to classify the transactions into four different classes with clear class boundaries as shown in the Fig.2. Since week can be divided into N parts and same is the case with the day, one can have N x N total classes. But with a large number of classes, the data points in a single class will be less and hence it does not remain a good candidate for an efficient analysis. So we have restricted ourselves to four classes i.e. Weekday-Day, Weekday-Night, Weekend-Day and Weekend-Night. In Fig.2, M represents the mean or centre point for each class. It is a two dimensional point and is represented as

$$\mathbf{M}\ (m_{day}, m_{week})$$

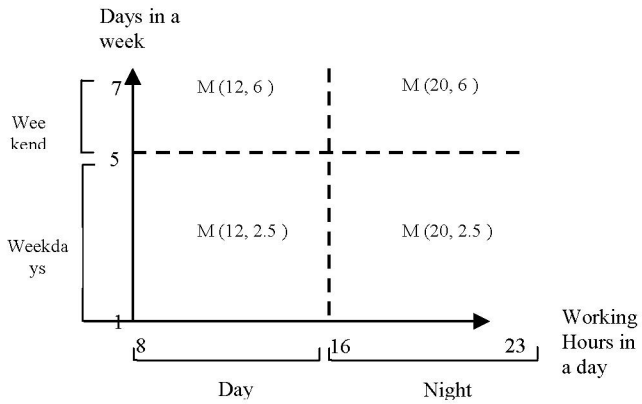

Figure 2. Two dimensional representation of four classes with mean points, where a mean point can be represented as M

$$(m_{day}, m_{week})$$

Classification of transactional data involves two main tasks; first it sums up all the transactions occurred at the same time at different counters of the supermarket into a single transaction and in the next step it assigns this transaction its appropriate cluster, depending upon the part of the week and part of the day the transaction occurred.

3.2.1. Summing up the Transactions. Since there can be multiple counters in a supermarket, so there are two possibilities.

1. Multiple transactions can be recorded at the same time.
2. Multiple transactions can be recorded in a small interval of time.

Since the customer himself is not important in our analysis, rather it is the behavior of the class to which that customer belongs. So in order to reduce the complexity for the analysis component it is wise to sum up the transactions occurring at the same time or in a small interval of time in the supermarket. All these customers have showed the same behavior with respect to the time of the transactions. This will ultimately reduce the number of records in each class and will make the work of analysis part easier.

3.2.1. Classifying the Transactions. After summing up the transactions, next task is to identify some behavior hidden in this transaction. This single transaction, which is actually the sum of different transactions performed by different customers at a single time, contains information about the products bought at that particular time by those different customers. Thus the common thing among all these customers is the time they visited supermarket. An important element of human psychology is that people are in the habit of visiting supermarket at some particular part of the week and also at some particular part of the day [9]. People who remain busy during the whole week in their jobs usually visit supermarkets on weekends and more precisely at night. Housewives on the other hand can visit

at any part of the week but they usually prefer the day time. Since different people exhibit different behaviors with respect to the time they visit supermarkets, we can use this behavior to cluster the transactions made by them and it will ultimately divide the customers into different clusters. Thus we have devised a way to classify the customers even in the absence of their demographics using only their behavior. The stepwise description of this component is as follows

Step 1: Since there can be billions of transactions in a supermarket database but for our analysis we focused on the transactions of last year. Let these transactions be represented by set D. We identify the time of the latest and the oldest transaction and divide the whole range into a set of ***n*** discrete small time intervals such that

$$\mathbf{T}=\{S_1, S_2, S_3, \cdots, S_n\} \quad \textbf{(6)}$$

If Δ represents the length of each interval then the i^{th} time interval can be written as

$$S_i = T_i \rightarrow T_{i-\Delta} \quad \textbf{(7)}$$

For accurate classification, Δ should be small.

Step 2:
1. For each $S_i \in T$
2. Select transactions $\{t_o \rightarrow t_n\} \in D$, satisfying criteria S_i
3. Sum up these transactions into a single transaction ***t***.
4. Determine the part of the week and part of the day for ***t***
5. Determine the class for ***t***.

Class of transaction ***t*** is determined by calculating the distance ***d*** between the time of transaction ***t*** and mean point ***M*** of each class using the equation (8)

$$d = |m_{week} - t_{week}| + |m_{day} - t_{day}| \quad \textbf{(8)}$$

The class whose centre ***M*** corresponds to the smallest ***d*** is the desired class for transaction ***t.*** At the end four different classes of transactions are produced (see figure 3).

4. Analyzing the behavior of each class for item set A and identifying the prospective class

Once all the transactions have been assigned their respective classes, the analysis component analyzes the behavior of each class for all the items in set ***A***, where ***A*** is the final item set produced in the item set generation step.

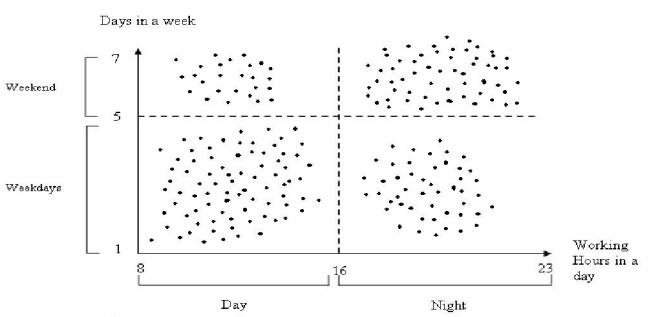

Figure 3. Classes after the classification of transactional data

For each item *i* in set A, its frequency is computed in each class, such that

$Freq_i$ = Number of transactions containing item *i* in a particular class **(9)**

Over all weight of the item *i* in a particular class is calculated as

$Weight_i = Freq_i$ / Total number of transactions in a particular class **(10)**

Weights for each item in set **A** in all four classes are stored in a matrix W having ***n*** rows and 4 columns, such that

$$W_{n\times 4} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \\ \vdots & \vdots & \vdots & \vdots \\ w_{n1} & w_{n2} & w_{n3} & w_{n4} \end{bmatrix} \quad \textbf{(11)}$$

Rows of matrix W correspond to the items in set **A** while columns of W correspond to four classes. W_{ij} represents the weight of i^{th} item in j^{th} class. Finally the weights of individual columns are summed up to provide total weight for all the items of A in that particular column, such that

$$TW = \begin{bmatrix} TW_1 & TW_2 & TW_3 & TW_4 \end{bmatrix} \quad \textbf{(12)}$$

Where

$$TW_i = \sum_{j=0}^{n} W_{ji} \quad \textbf{(13)}$$

Where **n** is the total number of products in set **A**

The class with the largest TW value will be the most profitable and prospective class for items of set **A** .

5. Experimental Results

Our proposed system performs three main tasks

1. Identify the matched products and generate frequent item set.

2. Classify the customer behavior by classifying the transaction data on the basis of date and time.
3. Identify the prospective customer class.

The accuracy of the third task depends upon the accuracy of the former tasks. If the item set generation and classification of the transactions have been done accurately, it will lead to the identification of correct prospective customer class. We simulated a virtual store, considering only the most common transactional attributes in any retail industry, as shown in Table 1. Total number of products used was 100 and 500000 transaction records were generated spanning a sale period of one year.

Before simulating the data, we chose a set of products similar to our new product and make the sale of some other products dependent on the sale of these matched products, thus we had already chosen our **Item Set**. During the simulation we kept the record of actual number of transactions happening at different parts of the week and at different times of the day, thus we have actual number of data points belonging to different classes on the basis of date and time of the transactions. Part from keeping the record of numbers, for analyzing the accuracy we divided the transactions into classes during simulation so that the classes generated by the system could be compared with them to identify the number of wrongly classified transactions. Finally we introduced some hidden patterns in the data which reveals a large percentage of sales for items of our chosen item set at a particular part of the week and a particular time of the day, thus we already knew the most profitable customer class for our item set. We then applied our proposed method on the same data set and compare the results of the system with the results of simulation to calculate the accuracy of the system.

Trans_date	Trans_time	Lane	Prod_upc	Quantity	amount
1/3/2003	12:54:00 PM	3	10	1	24

Table 1. Type of simulated transactional record

The indices we used were

1. Recall

Recall is the ratio between the total number of results identified by the system and the total number of results present in the actual data.

2. Precision

Precision is the ratio between the total number of correct results identified by the system and the total number of results identified by the system

3. F1 Measure

F1 measure uses both of recall and precision to give a final value that depicts the viability of the algorithm employed or our system.

$$F1 = \frac{2rp}{(r+p)} \quad \textbf{(14)}$$

Where **r** stands for recall and **p** stands for precision.

5.1. F1 Measure for Item Set generation

We ran simulation ten times, each with different number of products in the chosen item set and compared these with the item set generated by the system to calculate recall, precision and F1 values using the equations (15), (16) and (14) respectively. Results are presented in Table 2 and Fig 4.

$$r = \frac{\text{No of products in item set identified by the System}}{\text{No of Products in chosen item set}} \quad \textbf{(15)}$$

$$p = \frac{\text{No of actual products in the item set identified by the system}}{\text{Total no of products in the item set generated by the system}} \quad \textbf{(16)}$$

Experiment	No of products in chosen item set	No of products in item set identified by the System	r	p	F1
1	16	11	0.685	0.90	.77
2	12	10	0.83	0.80	.81
3	9	7	0.77	0.77	.74
4	12	10	0.833	0.7	.75
5	15	11	0.93	0.81	.77
6	20	16	0.8	0.75	.77
7	36	28	0.77	0.78	.80
8	25	25	1	0.84	0.91
9	10	7	0.7	1	0.8
10	18	11	0.61	0.81	0.694

Table 2. Numerical results for item set generation in ten simulation runs

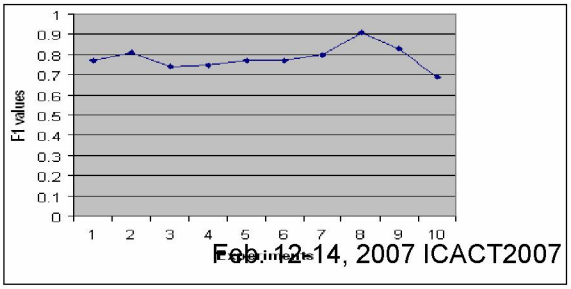

Figure 4. Graphical representation of F1 measure for item set generation in ten simulation runs

5.2. F1 Measure for Classification

We ran simulation ten times, each with a different number of transactions in four classes and compared these classes with the classes generated by the system to calculate recall, precision and F1 values for each class, using equations (17), (18) and (14) respectively. Finally mean F1 measure for a single simulation run is calculated using the F1 values of four classes. Results for the first simulation run are shown in Table 3, while the graphical representation of the final F1 values for all simulations is shown in Fig 5.

$$r = \frac{\text{No of transactions classified by the system for a class}}{\text{Actual no of transactions for that class}} \quad \textbf{(17)}$$

$$p = \frac{\text{No of actual transactions classified by the system for a class}}{\text{Total no of transactions classified by the system for that class}} \quad \textbf{(18)}$$

Class	Actual number of Transactions in the class	Number of transactions classified by the System for the class	r	p	F1
Weekday-Day	200000	174800	0.87	0.99	0.9
Weekday-Night	100000	93010	0.93	1	0.96
Weekend-Day	75000	100000	1.33	0.75	0.95
Weekend-Night	125000	132100	1.0568	0.80	0.91

Table 3. Numerical results for item set generation in ten simulation runs

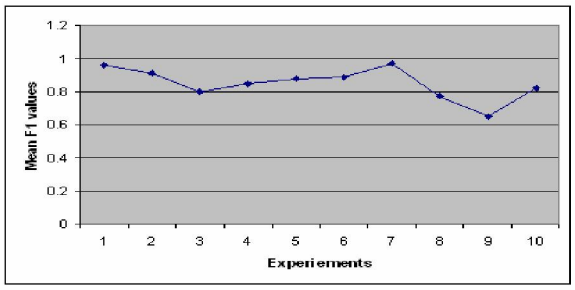

Figure 5. Graphical representation of mean F1 measure for classification in ten simulation runs

By observing the results presented in Fig.4 and Fig.5, we determined that the F1 measure for item set generation step is approximately 0.77 (77%) and for classification step is approximately 0.80 (80%). Since the accuracy of analysis component depends upon accuracy of above two tasks, it implies that system identifies the correct prospective customer class with an accuracy of approximately 77-80%. This is why, the system in all the 10 experiments identified the same prospective customer class, as the one simulated during the data generation.

6. Conclusion

It has been shown that performing prospecting by identifying actual semantics of transactional data is a logical way of doing prospecting. We have classified the customers on the basis of date and time of the transactions without using any customer demographics thus saving the cost, as supermarket databases lack customer demographics and acquiring such data by any other mean is very expensive. The analysis was based on the three most common transactional attributes in any retail industry i.e. Date, Time and Product.

7. References

[1] Geoff Wakeley, *Virtual Supermarket to Test Product success*, Aug 26, 1999

[2] Thiesing, F.M, Middelberg, U, Vornberger, O, "Short term prediction of sales in supermarkets". In Proceedings of IEEE International Conference on neural networks, 1995.

[3] Guoqing Chen, Qiang Wei, Etienne Kerre, "Fuzzy Data Mining: Discovery of Fuzzy Generalized Association Rules". In Bordagna & Pasi (eds.), Recent Research Issues on Management of Fuzziness in Databases. Physica-Verlag (Springer), 2000

[4] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", In ACM SIGMOD, 1993.

[5] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases". VLDB Conference, pages 487--499. 1994.

[6] S. K. Murthy, "Automatic construction of decision trees from data: a multidisciplinary survey". In Data Mining and Knowledge Discovery, number 2, pages 345--389, 1998

[7] Xingdong Wu, *Knowledge acquisition from databases*. Alex Publishing Corporation. USA,1995

[8] A. K. Jain, M.N. Murthy and P.J. Flynn: "Data Clustering: A Review". ACM Computing Reviews, Nov 1999

[9] Larry English, "Focusing on Customer a research report", Total DM, April 2004.
[10] Kim Humphery, *Shelf Life: Supermarkets and the Changing Culture of Consumption,* Cambridge University Press, July 27, 1998.