ICEB 2009 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-4-2009

Mining Implicit Patterns of Customer Purchasing Behavior Based On The Consideration Of RFM Model

Liewen Cheng

Jashen Chen

Che-Wei Chang

MINING IMPLICIT PATTERNS OF CUSTOMER PURCHASING BEHAVIOR BASED ON THE CONSIDERATION OF RFM MODEL

Liewean Cheng[1], Ja-Shen Chen[2], and Che-Wei Chang[3]
Department of Information Management, Department of Business Administration, and
Graduate Institute of Business Administration
Ta Hwa Institute of Technology, and Yuan Ze University, Taiwan
[1]liewen.cheng@gmail.com; [2]jchen@saturn.yzu.edu.tw; [3]s957143@mail.yzu.edu.tw

Abstract

Association rules have been developed for years and applied successfully for market basket analysis and cross selling among other business applications. One of the most used approaches in association rules is the Apriori algorithm. However the Apriori algorithm, has long known for its weaknesses that generate enormous amount of rules and already-known facts. In this study, we integrate the RFM attributes with the classical association rule mining, Apriori. Based on RFM model, two indicators, RF score and Sale ratio, are used as measure of interestingness. We propose two algorithms, DWRF and DWRFE, to mine for implicit pattern. In our experimental evaluation, the performance of Apriori, DWRF and DWRFE are compared. The result of our algorithms offers an effective measurement of interesting patterns. Moreover, the DWRF algorithm that uses the RF score as a measure of interestingness seems to be able to promptly reflect the fast-changing customer's purchase patterns.

Key words: Data mining; Association rules; RFM; CRM

Introduction

Because of the industry competitiveness and free communication of market information, customers are easier to change their preferences and buying behaviors. While the traditional marketing tools are unlikely to aim at customers' behavior or transformation of preferences (Shaw et al., 2001), database marketing becomes more and more popular. With the transaction records in database, it is important to analyze the correlation information of the customer, because these records represent the accumulation of customers' past decisions (Holtz, 1992). If the company can use this data to discover the patterns of customers' consumption behaviors, it can help marketing associates to make a more effective marketing program (Chen et al., 2005). Many enterprises have gathered great number of data into their data warehouses (Inmon, 1996). With the tradition marketing analysis tools, it is too difficult for employees to analyze such a huge database (Shaw et al., 2001). Data mining solves the difficulties of data analysis and also helps to discover hidden information (Fayyad et al., 1999). The mining technique depends on different distance of functions which can be distinguished as cluster, classification, estimation, perdition, and affinity group, description and profiling (Berry and Linoff, 2004)..

The most frequently used tool in the affinity group is the association rule analysis. Association rules have developed for many years; they are useful and easy to understand, and they are widely applied in various industries such as finance, telecommunications, retail, and online commerce (Kotsiantis and Kanellopoulos, 2006). It has increasingly attracted much academic research interest in recent years. Its main purpose is to discover the relationship of associated products. However, for capturing other interesting insights within these patterns, we can only depend on the expert domain knowledge to do the matching. Some of researches called these problems the "interestingness" problem (Liu et al., 1999). Our main purpose is to develop a new algorithm which will take RFM attributes as parameters in order to mine the implicit patterns. This research incorporates recency and frequency to discover the latest purchase patterns. We will assign a RF score based on its recency attribute and a sale-ratio score based on its monetary attribute to every rule in order to discover the most recent patterns and also the patterns that customers are willing to spend more.

Literature Review

Association rule mining, one of the famous researched friend of data mining According to Agrawal and Srikant (1994) , Association Rule is defined as:" to find out these association rule among a set of product items frequently purchased together".

The rule is written in the form as: (milk→ noodles}. It implies that if the customer purchases the milk, he would probably buy noodles.

The Association Rule Mining has two indicators to evaluate the meanings of these rules, which are support and confidence (Cavique, 2007; Chen, 2006; Kotsiantis and Kanellopoulos, 2006). support of an association rule can be defined as the percentage of records that contain X $\cup$ Y to the total number of records in the database. The user needs to set the threshold for support called minimum support; it is a key element to prune the search space and to limit the number of rules (Chen et al., 2006).

Another indicator, confidence is defined as the percentage of the number of transactions that contain X $\cup$ Y to the total number of records that contain X. Confidence mainly measures the strength of the association rules, Apriori is the most traditional approach of association rule mining.

According to Chen (2002), rules generated by association rule mining could be can meaningless. It is not enough to discover the user's interesting patterns only by o=support and confidence. (Padmanabhan and Tuzhilin, 1999).

RFM model primarily analyzes and evaluates the consumer behavior (Miglautsch, 2002). RFM represents three reference indicators respectively, such that recent purchase time (Recently), frequency of buying (Frequency) and how much you pay (Monetary) (Hughes, 1994). By using these three indexes, it's able to evaluate the relationships between companies and customers and determine the value of every customer. Because this model can successfully analyze customer behaviors and segment customers, some data mining researches have already combined this model into their approaches, .such as in the areas of cluster analysis and classification (Kuo, 2007). In cluster analysis, these researches use the RFM point to segment the customers into a number of clusters with similar characteristics (Sung and Sang, 1998; Russell and Lodwick, 1999). Another research is in the area of classification and RFM variables are used to classify customers into different categories according to customer's value. (Kitayama et al., 2002; Kaymak, 2001). In those researches that incorporate the RFM Model, the RFM attributes are used mainly to identify the customers but not to improve the algorithms of apriori (Liu and Shih, 2005).

3.Problem Statement and Definition

Market basket analysis, a typical example of association rule mining, is widely used in the retail industry. This process analyzes customer buying patterns by finding associations between the different items that customers have purchased in their carts. However, the disadvantages of this approach, apart from generating enormous amount of rules, are that the strong rules are not always interesting. According to Lin (Lin, 2001), the findings always represent the ordinary rules, i.e. the fact that are already known. These rules are easily deduced from our life experience, thus they provide little value for marketing campaign (Balaji and Padman, 1999).

Market basket analysis varies according to different criteria, such as the types of values, dimension of data, level of abstractions and the methods of finding frequent itemsets. However, none of the above approaches takes account of the impact of rapid transitions on customer purchase behavior. In this study, we will integrate RFM attributes with resulting association rules in hopes that these attributes will help decision makers to extract interesting pattern programmatically. These attributes could be used as useful measures of interestingness of association rule.

Firstly, we will take the recency attribute, i.e. purchase time, into consideration. Recency attribute will be split into several intervals, for instance, we can categorize recency attribute into 3 intervals: customers with purchases within the last one month; between last one and last two months; and between last two and last three months. Such categories may be arrived at by applying business rules or by domain expertise.

Apriori ranks these association rules by the confidence and support measures, however, these measures are insufficient to provide the information of recency. Therefore, some strong rules could be outdated. Since the buying patterns might change dramatically with time, the most recent rules can accurately reflect customer purchase patterns more than others. Without the consideration of purchase time, the resulting rules generated by association rule mining might mislead the managers into the wrong marketing strategies. Secondly, the monetary attribute, i.e. sale amount of each rule, is regarded as a valuable factor to invest. In this study, we will explore the ratio of the sales amount contributed by the items from a specified frequent itemset to the total sales amount of the transactions that contains this frequent itemset. The frequent itemsets with a higher expense ratio indicate that these itemsets take up a considerable portion of the sale amount for every visit and. that customers are willing to pay more for these items. In the following section, we will define the terms used in this study.

Definition 3.1: Transaction database

Let the database, D= {T1, T2, T3…Tn} be a set of customer transaction records, T represents a transaction and is labeled by a cart number. For instance, T1 contains a set of products that customer has purchased, and is denoted as T1={P1, P2, P3, P4…..Pn}, P is regarded as items in T1.

Definition 3.2: Recency variable

Recency is defined as the duration from customer's last transaction to the time that analysis starts. The recency attribute can be split into intervals which are determined by domain expert.

Definition 3.3: RF_weight

Formula 3.4.1:

I=number of time intervals, {F} =total frequency, that is the number of carts that contain the specified frequent itemset

The weight for those transactions containing a specified frequent itemset, whose purchase time fall into interval i., the formula (see Appendix) is, so as $\frac{(F+I-1)!}{F!(I-1)!}$ at $\frac{(F+I-2)!}{F!(I-2)!}$

interval II, and so on.

The complete weight list for interval 1 to I is:

$$\frac{(F+I-1)!}{F!(I-1)!} : \frac{(F+I-2)!}{F!(I-2)!} : \frac{(F+I-3)!}{F!(I-3)!} :\ldots: \frac{(F+I-I)!}{F!(I-I)!}$$

Example 3.5.1:

I=3,F=10

$$\mathbf{W1:W2:W3=C_F^{F+I-1} : C_F^{F+I-2} : C_F^{F+I-3}}$$

$$\mathbf{= C_{10}^{10+3-1}: C_{10}^{10+3-2} : C_{10}^{10+3-3}}$$

$$= 66:11:1$$

Definition 3.6: sale ratio

Sale-ratio is defined as

$$\frac{\Sigma(S_i)}{\Sigma(E_k)} \quad \begin{array}{l} 0<i<=n, \\ 0<k<=m, \end{array}$$

For those transactions containing the specified frequent itemset, we sum up the total sale amount contributed by items in this frequent item set, as in

$\Sigma(S_i)$ 0<i<=n,

here ***n*** represents the size of the frequent itemset *l*; $\boldsymbol{S_1}$ represents the total sale amount of Item*1* of all the transactions that contain this frequent itemset

We sum up the totol of sale amount For those transaction containing the specified frequent itemset, as in

$\Sigma(E_k)$ 0<k<=m,

for those transaction that contain the specified frequent itemset, ***m*** represents the number of transactions ;where E_1 is the total sale amount of cart1

4.Algorithm

4.1The DWRF –Algorithm

(Dynamic Weighted Recency and Frequency -Algorithm)

Derived from the classical algorithm, Apriori, we like to propose an novel algorithm which integrates the time and frequency factors to help screen out outdated rules. The new measure, RF Factor(i.e. Recency-Frequency score), can be used as another objective measure of interestingness for the decision makers.

We propose a new algorithm, DWRF, which use a RF score to filter out out-of-date patterns. The algorithm emphasizes the relationship between purchase time and patterns, therefore it will timely reflect the customer purchase patterns. In this approach, manager can dynamically set up the time intervals for comparison, for instance, the recency can be split into the last week, the last 2nd week, the last 3 week and so on. The RF scores are calculated dynamically depends on 2 parameters, the time intervals and the frequencies in each time slot for those transactions with a particular frequent itemset..

The DWRF- algorithm includes 5 phases as follows.

Phases 1: Defining time intervals by managers/users

Phases 2: Retrieving transactions which contain the frequent itemsets generated by Apriori.

Phases 3.a: Assigning every transaction with this specified frequent itemset to the right time slot by its purchase time and recording the count of every time slot.

Phases 3.b: Calculating the sale amount that is contributed by the items in every transaction with this specified frequent itemset, and the total amount of this transaction.

Phases 4: Calculating RF score for the specified frequent itemset,, repeat phase 2 to 4 for next frequent itemset until all the frequent itemsets are scanned.

Phases 5: Retrieving the frequent itemsets or rules whose RF scores exceed the RF-threshold given by manager or domain expert.

Example 4.1.1: Divide into three time intervals in the research period

Large itemset: A and B (generated by Apriori)

Total frequency =10

Time intervals are 2006/01, 2006/02, 2006/03

TID	Item	Sale date
1	ABCD	2006/03/12
2	ABDE	2006/02/11
3	ABDEFG	2006/02/10
4	ABD	2006/01/15
5	ABFGHK	2006/03/17
6	ABJKL	2006/03/18
7	ABMK	2006/03/09
8	ABGH	2006/02/24
9	ABDEK	2006/01/16
10	ABGHHKL	2006/01/14

↓

Time interval	Frequency	Sale date
2006/03	4	2006/03/12 2006/03/17 2006/03/18 2006/03/09
2006/02	3	2006/02/11 2006/02/10 2006/02/24
2006/01	3	2006/01/15 2006/01/16 2006/01/14

Figure 4.2: time-frequency transaction table.

Formula.4.1:RF_scores of large itemset:

$$\text{RF\_Score} = \frac{\sum_{1}^{I} \text{Weighted of each time interval} * \text{Frequency of each time interval}}{\text{Weighted of the receny time interval} * \text{Total frequency}}$$

$$= \frac{\sum_{i=1}^{I} w_{\ell i} * F_{\ell i}}{w_{\ell i} * f_{\ell}} \quad ,1 \le i \le I \,,\ i \in \text{integer}$$

The weight for interval 2006/03 is C_{10}^{10+3-1}=12!/(10!2!)=66
The weight for interval 2006/02 is C_{10}^{10+2-1}=11!/(10!1!)=11
The weight for interval 2006/01 is C_{10}^{10+1-1}=10!/(10!0!)=1

The RF-score is:
(66*4+11*3+1*3)/(66*10)=(264+33+3)/660=300/660=0.45

Example 4.1.2: Table 4.3 list the frequent itemsets, the count at each time slot, and the RF_score. RF_score represents the strength of recency. When the RF_threthods is set to 0.45, the latest frequent itemsets are {c e} {c b} {b e} {e g}{c f}

Table 4.3: An illustrative of the generation of RF_Scroe for all of large itemset

Frequent itemsets	Total frequency	**Time intervals by** Month-frequency			RF_Scroe
A b c	805	272	179	354	0.338
a d	354	122	200	32	0.348
c e	322	145	139	38	0.453
b d	259	50	167	41	0.198
c b	249	182	52	15	0.733
b e	243	145	90	8	0.600
e f	232	96	90	46	0.417
e g	231	111	92	28	0.484
c f	218	9	149	60	0.048
b g	218	85	98	35	0.394

4.2 The Incremental Mining Large Itemsets from DWRFE Algorithm
We propose another algorithm, DWRF, which. uses another measure, sale amount ratio, to extract those patterns that take takes up a good portion of sale per visit. We present that this pattern has the stronger strength with purchase monetary behavior.
4.2.1 The Sale Ratio Indicator
Sale ratio is defined as
For those transaction that contain the specified itemset, the summation of sale amount which is contributed bye items in frequent itemset divided by the summation of sale amount of transactions. We can infer that the customers are willing to pay more portion for those patterns with higher sale ratio.

Example 4.2.1:
Assume the there are 10 transactions that contains a frequent itemset {A,B,C}:

Table 4.6: An illustrative for {si} and {ei} for one of large itemset

Transaction	items	The sale amount {A,B,C}	total sale amount in this caet
1	{A,B,C,D,E}	40	80
2	{A,B,C,F,G}	60	300
3	{A,B,C,G,H}	100	160
4	{A,B,C,D,H}	70	180
5	{A,B,C,H, I}	150	200
6	{A,B,C,M,I}	55	140
7	{A,B,C,M,N}	65	120
8	{A,B,C,O,P,Q}	90	170
9	{A,B,C,S,T}	160	190
10	{A,B,C,M, T}	350	500

From the above data, the sale ratio would be calculated, as the formula:
{40+60+100+70+150+55+65+90+160+350}/ /
{80+300+160+180+200+140+120+170+190+500}
= 0.537

4.2.2 The Incremental Mining Frequent Itemsets from DWRFE Algorithm

This approach can focus on the users interesting pattern to be discovered faster and we show a whole process of DWRFE pattern generation as follows.

< DWRFE -algorithm incremental mining process>

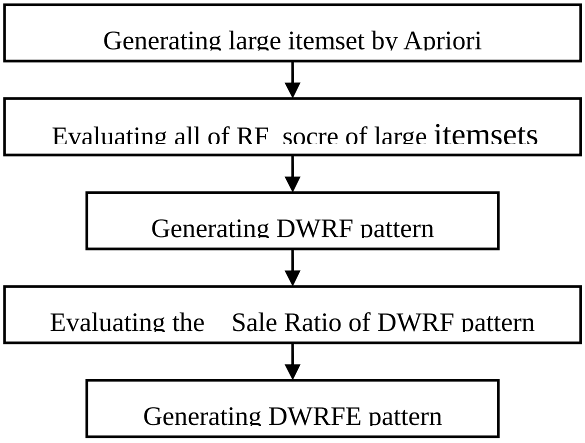

Generating large itemset by Apriori
Evaluating all of RF socre of large itemsets
Generating DWRF pattern
Evaluating the Sale Ratio of DWRF pattern
Generating DWRFE pattern

Figure 4.3: DWRFE -algorithm incremental mining process.

5. Experiments

The experiments were used to examine our algorithm performance which are compared with the original approach,Apriori. The tools include Microsoft SQL Server 2005 BI Suite and Java lanuage.

5.1 Data Sources and Data Structures

In this section, we assign four datasets in our experiment for testing our algorithm. All of the four datasets sources are from real datasets of one of major retailers in Taiwan during January 2005 to December 2006. The data are divided into a 75 percent and a 25 percent randomly, named RT-75 and RT-25; another two, CT-1 and CT-2, are modified data. Table1 summarizes the parameters used.

Table 5.1: Datasets description

	Dataset1	Dataset2	Dataset3	Dataset4
Name	RT-25	RT-75	CT-1	CT-2
Size	19995	65819	103435	144332
Data source	Real data	Real data	Synthetic	Synthetic

5.2 Performance Evaluation

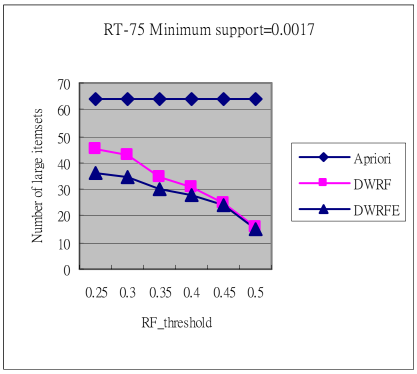
The comparison of the effectiveness among the three algorithms ,Apriori, DWRF and DWRFE was made by the following aspects; firstly the number of frequent itemsets; secondly, the importance of the top ten frequent itemsets.

Comparison with the Number of Frequent Itemset

We applied these algorithms to the four datasets and observed that the number of frequent itemsets decreased when the RF-threshold increased.(see Figure5.1), where the minimum support is 0.0017,minimum sale ratio is 0.3, and RF score ranges from 0.25 to 0.5. When the RF-threshold went up, there is no change in number for Apriori, however, the number of frequent itemsets in both DWRF and DWRFE dropped dramatically. We also observed that the numbers became closer between DWRF and DWRFE when the RF-threshold increased. This might imply that the DWRF has screened out the majority of unsatisfied rules. The consequence was similar in all datasets no matter whether the minimal support was changed (Fig. 5.1(b))

Thus, we can infer that DWRF and DWRFE can diminish numbers of patterns successful better than the classical approach.

(a)



(b)

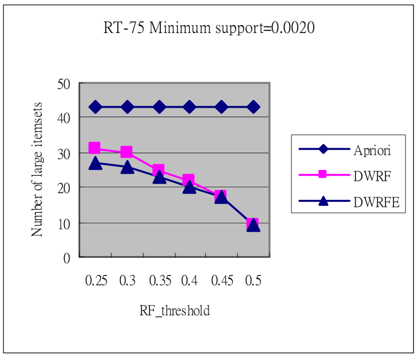


Figure 5.1: Comparing numbers of large itemsets of Apriori, DWRF and DWRFE

5.2.2 Comparisons of Top Rules with Algorithm Outputs

We applied the three algorithms to dataset CT-2. with a minimum support of 0.0020. and selected the top rules from these algorithms (table 5.2). Ranking by confidence, RF score, and sale ratio are listed also. We observed that the rules were ranked differently according to different criteria, especially the great difference between Apriori and DWRF. For instance, the fifth rule of RF algorithm is ranked as the 41th place in Apriori algorithm. This indicates that we might accidently delete some rules by increasing minimal confidence, thus we might miss some interesting patterns. There made little difference between the ranking of rules between DWRF and DWRFE. This is possibility due to the total sale amount of shopping carts is too huge, and sale amount contributed by the specified frequent itemset is relatively smaller.

Table 5.2: Top rules selected by Apriori, DWRF, and DWRFE

DWREF ranking	DWRF-Ranking	Apriori-ranking	RF_score	rules
8	4	13	0.529	222341->222343
4	1	17	0.601	8076-> 8074
7	2	19	0.578	8085-> 8074
1	3	23	0.554	8081->8074
3	8	39	0.468	107626->118909
2	5	41	0.497	118909->7685
5	7	43	0.479	115881->7685
6	6	49	0.483	118909->135008

e=0.005 and the time intervals were set by quarter.

6. Conclusions and Suggestion

In this research, in order to solve the interestingness problem, we present two novel approaches, DWRF-algorithm and DWRFE-algorithm to mine for the implicit patterns. Two parameters, RF score and Sale Ratio, are used as measures of recency and monetary characteristics of rules. The DWRF and DWRFE algorithm not only reduce the number of rules but also retrieve prompt and interesting patterns. Moreover, the time intervals can be set dynamically according to managers' wishes.; either in a consecutive or discrete manner. In addition, the RF weight formula for each time interval is devised mathematically. The DWRF provides a rational base for weight assignment. The approach is able to respond fast to user's needs in the different environments by analyzing the current customer purchase pattern. It seemed to be that our approach is more flexible than other classical algorithm.

We tested Apriori, DWRF, DWRFE on four datasets and compared the result.. All the results show that our approach is more effective to reduce the number of the frequent itemsets than Apriori. We observe that there is a big difference between the ranking of DWRF and that of Apriori., and the result proves to be that our algorithm can retrieve timely purchase pattern.

References

[1] Agrawal, R. & Srikant, R. (1994), "Fast Algorithms for Mining Association Rules," Proceedings of the 20th International Co nference on Very Large Database, pp. 487 -499.

[2] Berry, M., & Linoff, G. (2004), "Data mi ning techniques: for marketing, sales, and customer relationship management," John Wiley & Sons, Inc., NY.

[3] Cavique. L. (2007), "A scalable algorithm for the market basket analysis," Journal of Retailing and Customer Services 14(6) pp.400-407.

[4] Chen, G., H. Liu, Yu. L., Wei. Q., & Zhang. X. (2006), “A new approach to classification based on association rule mining,” Decision Support Systems, 42(2), pp. 674-689.

[5] Chen, M.C., Chiu, A.L., & Chang, H.H.(2005), “Mining changes in customer behavior in retail marketing”, Expert Systems With Applications, 28(4), pp. 773-781.

[6] Chen, Y. L., Zhao, S. R., & Chen, Y. C. (2003), “Several Improved Data Mining Algorithms for Finding Association Rules,” Journal of E-Business, 5(2), pp. 1-10.

[7] Fayyad, U.M., Piatetsky, G. S., & Symth, P. (1996), “From data mining to knowledge discovery in databases,” AI Magazine, pp.37-54.

[8] Frawley, A., & Thearling, K. (1999), “Increasing Customer Value by Integrating Data Mining and Campaign Management Software,” Database Management, pp.49-53.

[9] Hughes, Arthur M. (1994), “Strategic Database Marketing,” Chicago: Probus Publishing.

[10] Holtz, H. (1992), “Databased Marketing－Every Manager’s Guide to the Super Marketing Tool of the 21st Century,” Wiley, New York.

[11] Inmon, W. (1996), “Building the Data Warehouse,” N.Y.: Wiley Press

[12] Kitayama, M., Matsubara, R., Izui, Y. (2002), “Application of data mining to customer profile analysis in the power electric industry,” Power Engineering Society Winter Meeting, IEEE,1, pp. 632-634.

[13] Kotsiantis, S & Kanellopoulos, D (2006), “Association Rules Mining: A Recent Overview,” GESTS International Transactions on Computer Science and Engineering, 32 (1), pp. 71-82.

[14] Kuo, M.H. (2007),“Discovering RFM sequential patterns from customers’ purchasing data” Department of Information Management, National Central University, Taiwan.

[15] Liu, B., Hsu, W., Mun, L.F., & Lee, H.Y. (1999), “Finding Interesting Patterns Using User Expectations,” IEEE Transactions on Knowledge and Data Engineering,11, No. 6, pp. 817-832

[16] Liu, D.R., & Shih, Y.Y. (2005), "Integrating AHP and data mining for product recommendation based on customer lifetime value," Information & Management, 42(3), pp. 387-400.

[17] Miglautsch, J. (2002), “Application of RFM principles: what to do with 1-1-1customer?”Journal of Database Marketing, Jul, 9.4

[18] Padmanabhan, B., & Tuzhilin.A.(1999), “Unexpectedness As A Measure Of Interestingness In Knowledge Discovery,” Decision Support Systems, 27, pp. 303-318

[19] Russell, S., Lodwick, W. (1999), “Fuzzy clustering in data mining for telco database marketing campaigns, Fuzzy Information Processing Society,”NAFIPS. 18th International Conference of the North American, pp. 720-726.

[20] Shaw, M.J., Subramaniam, C., Tan, G.W., & Welge, M. E. (2001), “Knowledge Management and Data Mining for Marketing,” Decision Support Systems, 31, pp.127-137.

[21] Sung, H. H., Sang, C. P., (1998) “Application of data mining tools to hotel data mart on the Intranet for database marketing,” Expert Systems with Applications, 15 (1), pp.1-31.