Online Customer Value Identification Based On Site Usage Time through Data Mining Analysis

[1]R. Deepa, [2] R. Hamsaveni
[1]M.Phil
[1] rdeepa_25@yahoo.co.in.com,
[2] amsavenisasikumar@rediffmail.com

GJCST Computing Classification
H.2.8 & J.1

***Abstract*- CRM is required to evaluate the customer performance, discover the trends or patterns in customer behavior, and understand the factual value of their customers to their company. Data mining is playing an important role in analyzing and utilizing the abundant scale of information gathered from customers. This paper is on identifying online customer value based on the time duration of his stay in the online shop, through Data Mining Analysis. In this paper, the sequence of online site time usage can be modeled by using a Hidden Markov Model (HMM) and show how it can be used for the detection of a potential buyer and hence use appropriate promotion campaigns. The clustering technique is applied to identify the stay time category of the customer when he ends up buying a product. An HMM is initially trained with the normal web site stay of a site visitor. If a customer stay time reaches the time as calculated by the trained HMM with sufficiently high probability, he is considered to be a potential buyer. A detailed experimental result to show the effectiveness of our approach is presented.**

Keywords-HMM,K-Means,Model arameter Estimation

I INTRODUCTION

The popularity of online shopping is growing with every passing day. According to the ACNielsen study conducted in 2005, one-tenth of the world's population is shopping online. Customers visit online shops frequently with the intention of buying or making the decision process for buying. Companies usually aim at their produces, products and then give all the customers same sales promotion. This kind of sales promotions neglects the differences among customers. In most cases, these promotions cost a lot, but only get few real profits from customers. That means many promotions are waste. Data mining can solve many typical commercial problems, such as Database Marketing, Customer Segmentation and Classification, Profile Analysis, Cross-selling, Churn Analysis, Credit Scoring, Fraud Detection, and so on. Since those data mining technologies appeared, companies have changed their sales target from products to customers. How to classify customers? How to find out the common character of customers from database? How to dig up the potential customers? How to find out the most valuable customers? These kinds of questions become the most popular data mining applications in marketing.

Nevertheless, the recent customer relation analyses techniques do not consider the time duration factor of the customer's stay on the site and hence has drawbacks. The most important one is that based on those analyses company usually consider the customer visit and stay time on the site as an isolated object and having value only when he/she buys products deals with this company. The duration of the customers stay is neglected in analyzing the value from potential purchase probability. This paper reveals about the application of data mining in online customer value identification, and to use new visual angle to improve these drawbacks.

II RELATED WORK ON ONLINE CUSTOMER VALUE ANALYSIS

The web Usage analysis includes straightforward statistics, such as page access frequency, as well as more sophisticated forms of analysis, such as finding the common traversal paths through a Web site. Usage information can be used to restructure a Web site in order to better serve the needs of users of a site. Long convoluted traversal paths or low usage of a page with important site information could suggest that the site links and information are not laid out in an intuitive manner. The design of a physical data layout or caching scheme for a distributed or parallel Web server can be enhanced by knowledge of how users typically navigate through the site. Usage information can also be used to directly aide site navigation by providing a list of "popular" destinations from a particular Web page.

Web Usage Mining is the application of data mining techniques to large Web data repositories in order to produce results that can be used in the design tasks mentioned above. Some of the data mining algorithms that are commonly used in Web Usage Mining are association rule generation, sequential pattern generation, and clustering. Association Rule mining techniques [1] discover unordered correlations between items found in a database of transactions. In the context of Web Usage Mining a transaction is a group of Web page accesses, with an item being a single page access. Examples of association rules found from an IBM analysis oft he server log oft he Official 1996 Olympics Web site [7] are: – 45% oft he visitors who accessed a page about Indoor Volleyball also accessed a page on Handball.

- 59.7% oft he visitors who accessed pages about Badminton and diving accessed a page about Table Tennis.

The percentages reported in the examples above are referred to as *confidence*. Confidence is the number of transactions containing all of the items in a rule, divided by the number of transactions containing the rule antecedents (The

antecedents are Indoor Volleyball for the first example and Badminton and Diving for the second example).

The problem of discovering *sequential patterns* [17, 26] is that of finding inter transaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set. By analyzing this information, a Web Usage Mining system can determine temporal relationships among data items such as the following Olympics Web site examples:

- 9.81% oft he site visitors accessed the Atlanta home page followed by the Sneak peek main page.
- 0.42% oft he site visitors accessed the Sports main page followed by the Schedules main page.

The percentages in the second set of examples are referred to as support. Support is the percent of the transactions that contain a given pattern. Both confidence and support are commonly used as thresholds in order to limit the number of rules discovered and reported. For instance, with a 1% support threshold, the second sequential pattern example would not be reported.

A. *Clustering Analysis*

Clustering analysis allows one to group together users or data items that have similar characteristics. Clustering of user information or data from Web server logs can facilitate the development and execution of future marketing strategies, both online and off-line, such as automated return mail to visitors falling within a certain cluster, or dynamically changing a particular site for a visitor on a return visit, based on past classification of that visitor. As the examples above show, mining for knowledge from Web log data has the potential of revealing information of great value. While this certainly is an application of existing data mining algorithms, e.g. discovery of association rules or sequential patterns, the overall task is not one of simply adapting existing algorithms to new data. Ideally, the input for the Web Usage Mining process is a file, referred to as a user session file in this paper that gives an exact accounting of who accessed the Web site, what pages were requested and in what order, and how long each page was viewed. A user session is considered all of the page accesses that occur during a single visit to a Web site. The information contained in a raw Web server log does not reliably represent a user session file for a number of reasons that will be discussed in this paper. Specifically, there are a number of difficulties involved in cleaning the raw server logs to eliminate outliers and irrelevant items, reliably identifying unique users and user sessions within a server log, and identifying semantically meaningful transactions within a user session.

There are several data preparation techniques and algorithms that can be used in order to convert raw Web server logs into user session files in order to perform Web Usage Mining. The specific contributions include (i) development of models to encode both the Web site developer's and users' view of how a Web site should be used, (ii) discussion of heuristics that can be used to identify Web site users, user sessions, and page accesses that are missing from a Web server log, (iii) definition of several transaction identification approaches, and (iv) and evaluation oft he different transaction identification approaches using synthetic server log data with known association rules.

B. *Data Mining Process*

There are three main steps of data mining process.

i. Data Preparation

In the whole data mining process, data preparation is somehow a significant process. Some book says that if data mining is considered as a process then? Data preparation is at the heart of this process. However, nowadays databases are highly susceptible to noise, missing and inconsistent data. So preprocessing data improve the efficiency and ease of the data mining process, this becomes an important problem. Several consulting firms, such as IBM, have approved that data preparation costs 50% - 80% resource of the whole data mining process. From this view, there is a need to pay attention for data preparation. There are three data preprocessing techniques should be considered in data mining:

- Data cleaning-**Inconsistent Data**- Not all the data taken is "clean". For example, a list of Nationality may have the values of "China", "P.R.China", and "Mainland China". These values refer to the same country, but are not known by the computer. Therefore, this is a consistency problem. b) Missing values Data from a company's database often contains missing values. Sometimes the approaches require rows of data to be complete in order to mine them, but the database may contain several attributes with missing values. If too many values are missing in a data set, it becomes hard to gather useful information from this data. c) Noisy data Noise is a random error or variance in a measured variable.
- Data integration usually the data analysis task will involve data integration. It combines data from multiplying sources into a coherent data store. Those sources include multiple database or flat files. Several issues should be considered during data integration, such as schema integration, correlation analysis for detecting redundancy, and detection and resolution of data value conflicts. Careful integration of the data can help improve the accuracy and speed of the mining process.
- Data reduction If you select data from a data warehouse, you probably find the data set is huge. Data reduction techniques can be applied to obtain a reduced representation of the data set. Mining on reduced data set should be more efficient yet produce the same analytical results. It includes several strategies, such as data cube aggregation, dimension reduction, data compression, numerosity

reduction, and discretization and concept hierarchy generation

ii. Knowledge Discovery In Database

As a core data mining techniques, knowledge and information discovery has several main components:

- Determine the type of data mining tasks in order to confirm that the functions and tasks to be achieved by recent system belong to which kind of classification or clustering.
- Choose suitable technologies for data mining. So the appropriate data mining technologies based on the tasks have to be confirmed. Such as, classification model often use learning neural network or decision tree to realize; while clustering usually use clustering analysis algorithms to realize; association rules often use association and sequence discovery to realize.
- Select a specific algorithm based on the technologies. Furthermore, a new efficient algorithm can be designed by the specific mining tasks. For choosing the data mining algorithms determine the hidden pattern in selecting the data.
- Mining data use the selected algorithms or algorithms portfolio to do repeated and iterative searching. Extract the hidden and innovative patterns from data set.

iii. Model Explain and Estimate

Explain and estimate the patterns got from data mining, get the useful knowledge. For instance, remove some irrespective and redundant patterns, after filtration the information should be presented to customers; Use visualization technology to express the meaningful model, in order to translate it into understandable language for users. A good application of data mining can change primal data to more compact and easily understand form and this form can be defined definitely. It also includes solving the potential conflict between mining results and previous knowledge, and using statistical methods to evaluate the current model, in order to decide whether it is necessary to repeat the previous work to get the best and suitable model. The information achieved by data mining can be used later to explain current or historical phenomenon, predict the future, and help decision-makers make policy from the existed facts.

III HMM BACKGROUND

An HMM is a double embedded stochastic process with two hierarchy levels. It can be used to model much more complicated stochastic processes as compared to a traditional Markov model. An HMM has a finite set of states governed by a set of transition probabilities. In a particular state, an outcome or observation can be generated according to an associated probability distribution. It is only the outcome and not the state that is visible to an external observer [18]. HMM-based applications are common in various areas such as speech recognition, bioinformatics, and genomics. Lane [23] has used HMM to model human behavior. Once human behavior is correctly modeled, it can be used to predict the future behaviour and hence help in managing future actions to the predicted behaviours.

An HMM can be characterized by the following [18]:

i. N is the number of states in the model and the set of states $S = \{S_1, S_2, \ldots . S_N\}$ where $S_i, i = 1,2, \ldots N$ is an individual state. The state at time instant t is denoted by q_t.

ii. M is the number of distinct observation symbols per state. The observation symbols correspond to the physical output of the system being modeled. Let the set of symbols be $V = \{V_1, V_2, \ldots . V_M\}$, where $V_i, i = 1,2, \ldots M$ is an individual symbol.

iii. The state transition probability matrix $A = [a_{ij}]$, where $a_{ij} = P(q_{t+1} = S_j | q_t = S_i), 1 \leq i \leq N; 1 \leq j \leq N; t = 1, 2, \ldots$ For the general case where any state j can be reached from any other state i in a single step, then $a_{ij} > 0$ for all i, j. Also, $\sum_{j=1}^{N} a_{ij} = 1, 1 \leq i \leq N$.

iv. The observation symbol probability matrix $B = [b_j(k)]$, where $b_j(k) = P(V_k | S_j), 1 \leq j \leq N, 1 \leq k \leq M$ and $\sum_{k=1}^{M} \mathrm{b_j}(\mathrm{k}) = 1, 1 \leq \mathrm{j} \leq \mathrm{N}$.

v. The initial state probability vector $\pi = [\pi_i]$, where $\pi_i = P(q1 = S_i), 1 \leq i \leq N$, such that $\sum_{i=1}^{N} \pi_\mathrm{i} = 1$

vi. The observation sequence $O = O_1, O_{i2} \ldots O_R$ where each observation O_t is one of the symbols from V, and R is the number of observations in the sequence.

It is evident that a complete specification of an HMM requires the estimation of two model parameters, N and M, and three probability distributions A, B, and π. The notations are used to indicate the complete set of parameters of the model, where A, B implicitly include N and M.

A. Use Of HMM For Identifying Online Customer Value

Ideally, the input for the Web Usage Mining process is a file, referred to as a user session file in this paper that gives an exact accounting of who accessed the Web site, what pages were requested and in what order, and how long each page was viewed. A user session is considered to be all of the page accesses that occur during a single visit to a Web site. The information contained in a raw Web server log does not reliably represent a user session file for a number of reasons. Specifically, there are a number of difficulties involved in cleaning the raw server logs to eliminate outliers and irrelevant items, reliably identifying unique users and user sessions within a server log, and identifying semantically meaningful transactions within a user session.
The analyzing program runs at the application layer or the business layer of the web application. Each web customer

time is submitted to the application on server. The application receives the current session duration details. It tries to find the user's normal site stay time that results in the purchase of a product based on the stay time profile calculated by clustering. If the customer stay time has approached the profile time, he can be identified as a more valuable customer and hence provided with appropriate campaigns. In this section, how HMM can be used for online customer value identification is explained.

B. HMM Model For Identifying Online Customer Value

To map the customer site stay or usage time in terms of an HMM, first decide the observation symbols in the model. Then quantize the stay time values x into M stay ranges V1; V2; . . . VM, forming the observation symbols at the business logic layer. The actual stay duration for each symbol is configurable based on the stay habit of individual customers. These duration ranges can be determined dynamically by applying a clustering algorithm on the values of each customer's site stay time, as shown in Section III.C.

In this work, it is considered only three time ranges, namely, low l, medium m and high h. Our set of observation symbols is, therefore, V = {l, m, h} making M = 3. For example, let l = (0,15 mins), m = (30 mins, 45 mins) and h =(1 hour,site logout time). If customer stays on site for 30 mins, then the corresponding observation symbol is m. A customer stays on the site for different amounts of time over a period. One possibility is to consider the sequence of duration of stay and look for similarities in them. However, the sequence of types of stay time is more stable compared to the sequence of exact stay times. The reason is that, a customer stays online depending on his need for procuring different types of items over a period. This, in turn, generates a sequence of site stay times.

Each individual stay time usually depends on the corresponding type of site stay. Hence, the transition in the type of stay time is considered to state the transition in our model. The type of each stay is linked to the decision making time for products from the corresponding merchant. This information about the customer's decision-making time is not known to the online shop application. Thus, the intention of the customer is hidden from the application. The set of all possible types of decisions forms the set of hidden states of the HMM.

One possibility is to consider the sequence of duration of stay and look for similarities in them. However, the sequence of types of stay time is more stable compared to the sequence of exact stay times. The reason is that, a customer stays online depending on his need for procuring different types of items over a period of time. This, in turn, generates a sequence of site stay times. Each individual stay time usually depends on the corresponding type of site stay. Hence, the transition in the type of stay time is considered as state transition in the model. The type of each stay is linked to the decision making time for products from the corresponding merchant. This information about the customer's decision-making time is not known to the online shop application. Thus, the intention of the customer is hidden from the application. The set of all possible types of decisions forms the set of hidden states of the HMM.

After deciding the state and symbol representations, the next step is to determine the probability matrices A, B, and _ so that representation of the HMM is complete. These three model parameters are determined in a training phase using the Baum-Welch algorithm[18] . The initial choice of parameters affects the performance of this algorithm and, hence, they should be chosen carefully.

One possibility is to consider the sequence of duration of stay and look for similarities in them. However, the sequence of types of stay time is more stable compared to the sequence of exact stay times. The reason is that, a customer stays online depending on his need for procuring different types of items over a period this, in turn, generates a sequence of site stay times. Each individual stay time usually depends on the corresponding type of site stay. The type of each stay is linked to the decision making time for products from the corresponding merchant. This information about the customer's decision-making time is not known to the online shop application. Thus, the intention of the customer is hidden from the application. The set of all possible types of decisions forms the set of hidden states of the HMM.

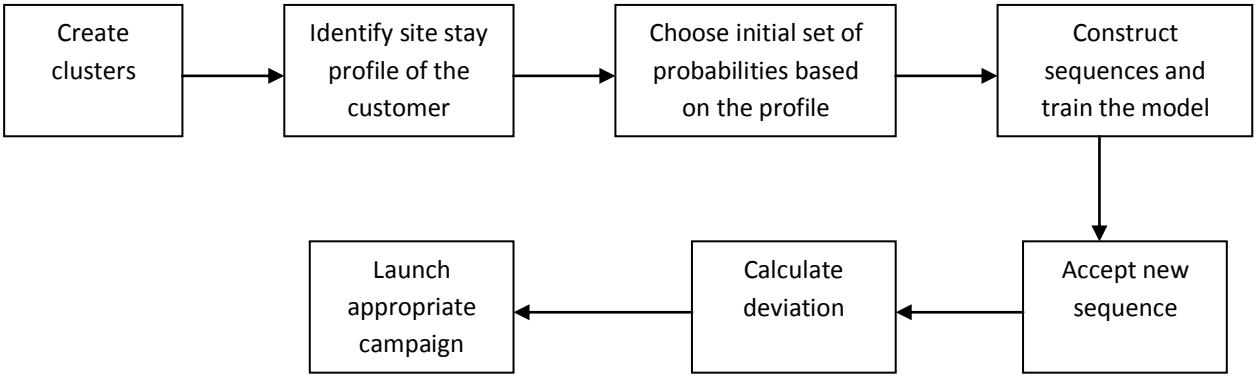

Fig 1. Flow of Process

Thus, the intention of the customer is hidden from the application. The set of all possible types of decisions forms the set of hidden states of the HMM.
The special case of fully connected HMM is considered, in which every state of the model can be reached in a single step from every other state. Site stay profiles of the individual customers are used to obtain an initial estimate for probability matrix B of (2).

C. Dynamic Generation Of Observation Symbols

Data Mining in general is an incremental process which includes the following phases as described by the CRISP-DM Process Model (Cross-Industry Standard Process for Data Mining):

- **Business Understanding**- defining objectives and requirements
- **Data Understanding**- data collecting, checking quality and consistency
- **Data Preparation**- preparing the data for the data mining algorithms / tools (includes selection, normalization etc.)
- **Modelling**-applying various data mining techniques to the data; stepping back to 3. may be necessary due to requirements of some algorithms
- **Evaluation**- evaluation of the model to be certain that it meets the requirements; if not, stepping back to 1. may be necessary
- **Deployment**-Clustering algorithms calculate a partitioning of a dataset into subsets (clusters) in a way that instances within a subset are more similar to each other than to instances within another subset. This is usually done using a distance measure D(r1,r2), which is specific for the data to be compared. Typical distance measures are the Euclidian and the Manhattan distance.

For each customer, we train and maintain an HMM. To find the observation symbols corresponding to individual customer's transactions dynamically. Finally a clustering algorithm is executed on his past transactions. Normally, the duration of site stay details are stored in the web session log contain many attributes. For our work, r only the time duration that the customer spent in his site visit. Although various clustering techniques could be used, we use K-means clustering algorithm [24] to determine the clusters.

A. K-Means

The K-Means algorithm is a simple, well known, and widely used clustering algorithm. The algorithm is based on the idea that objects (input vectors, records) are grouped into clusters according to a distance function, for example the Euclidian distance. The resulting clusters contain objects with a minimum within-cluster distance. The algorithm is performed as follows:

- A number of so called centroids are randomly spread among the input range. For each cluster one centroid will be calculated.
- Each object is assigned to the cluster represented by the closest centroid according to the distance function.
- The new position of the centroid is found by calculating the centre of all objects that are assigned to that cluster. This will cause the centroids to 'move around'.
- Point 2. und 3. are repeated until the centroids do not change any more.

K-means is an unsupervised learning algorithm for grouping a given set of data based on the similarity in their attribute (often called feature) values. Each group formed in the process is called a cluster. The number of clusters K is fixed a priori. The grouping is performed by minimizing the sum of squares of distances between each data point and the centroid of the cluster to which it belongs.
In our work, K is the same as the number of observation symbols M. Let c1; c2; . . . cM be the centroids of the generated clusters. These centroids or mean values are used to decide the observation symbols when a new entry comes in. Let x be the amount spent by the customer u in visit times T. As mentioned before, the number of symbols is 3 in our system. Considering M ¼ 3, and executing K-means algorithm on the example transactions in Table 2, the clusters are obtained, as shown in Table 3, with cl, cm, and ch as the respective centroids. It may be noted that the duration values 5, 10, and 10 have been clustered together as cl resulting in a centroid of 8.3. The percentage ðpÞ of total number of transactions in this cluster is thus 30 percent. Similarly, duration time values 15, 15, 20, 25, and 25 have been grouped in the cluster cm with centroid 20, whereas time values 40 and 80 have been grouped together in cluster ch. cm and ch, thus, contain 50 percent and 20 percent of the total number of transactions.
When the application layer receives a new entry T for this customer, it measures the distance of the stay time x with respect to the means cl, cm, and ch to decide the cluster to which T belongs and, hence, the corresponding observation symbol. As an example, if x =10mins, then in Table 2 using (9), the observation symbol is V1= l.

Table 1 Example stay times in Each visit

Visit No	1	2	3	4	5	6	7	8	9	10
Duration of stay	40	25	15	5	10	20	15	20	10	80

B. Site Stay Profile Of Customers

The site stay profile of a customer suggests his normal site stay behavior. Customers can be broadly categorized into three groups based on their browsing duration habits, namely, long-staying (hs) group, medium-staying (ms) group, and low-staying (ls) group. Custmers who belong to the hs group, normally take more time to make their decisions about purchasing an item. Similar definition applies to the other two categories also. Site stay profiles of customers are determined at the end of the clustering step.

Let pi be the percentage of total number of visits of the customer that belong to cluster with mean ci.
Thus, site stay profile denotes the cluster number to which most of the site visits of the customer belong. In the example in Table 2, the site stay profile of the customer is 2, that is m and, hence, the customer belongs to the ms group.

Table 2 Output of K-Means Clustering Algorithm

Cluster mean/ centroid name	cl	cm	ch
Observation symbol	V1=l	V2=m	V3=h
Mean values (Centroid)	8.3	20	60
Percentage of total transactions (p)	30	50	20

C. Model Parameter Estimation and Training

To estimate the HMM parameters for each customer Baum-Welch algorithm is used. The algorithm starts with an initial estimate of HMM parameters A, B, and _ and converges to the nearest local maximum of the likelihood function. Initial state probability distribution is considered to be uniform, that is, if there are N states, then the initial probability of each state is 1=N. Initial guess of transition and observation probability distributions can also be considered to be uniform. However, to make the initial guess of observation symbol probabilities more accurate, site stay profile of the customer, as determined in Section 3.4, is taken into account. So three sets of initial probability is made for observation symbol generation and also for three site stay groups—ls, ms, and hs. Based on the customer's site stay profile, the corresponding set of initial observation probabilities is chosen. The initial estimate of symbol generation probabilities using this method leads to accurate learning of the model. Since there is no a priori knowledge about the state transition probabilities, the initial guesses are considered to be uniform.

From now start training the HMM. The training algorithm has the following steps:

- initialization of HMM parameters,
- forward procedure, and
- backward procedure.

Details of these steps can be found in [18]. For training the HMM, the customer's site stay duration is converted into observation symbols and form sequences out of them. At the end of the training phase, an HMM corresponding to each customer is obtained. Since this step is done offline, it does not affect the application performance, which needs online response.

D. Launch The Appropriate Campaign

After the HMM parameters are learned, the symbols from a customer's training data are taken and form an initial sequence of symbols. Let O1;O2; . . .OR be one such sequence of length R. This recorded sequence is formed from the customer's transactions up to time t. We input this sequence to the HMM and compute the probability of acceptance by the HMM. If the deviation is not there, the customer is most probably going to make a purchase now and hence an appropriate campaign to close the sale can be launched.

IV Results

Predicting a potential buyer using real data set is a difficult task. Online shops do not, in general, agree to share their data with researchers. There is also no benchmark data set available for experimentation. Therefore, the large-scale simulation studies are performed to test the efficacy of the system. A simulator is used to generate a mix of visit times. The customers are classified into three categories as mentioned before—the low, medium, and hs groups. The effects of group and the percentage of visits that belong to the low, medium, and high-range clusters. First carry out a set of experiments to determine the correct combination of HMM design parameters, namely, the number of states, the sequence length, and the threshold value. Once these parameters were determined, the process was tested to appropriately predict a customers stay time and classify the customer to identify his buying decision and hence his value.

V Conclusion

In this paper, we have proposed an application of HMM in online customer value identification. The different steps in customer site stay and decision-making process to buy a product are represented as the underlying stochastic process of an HMM. The ranges of site stay times as the observation symbols are used, whereas the types of stay have been considered to be states of the HMM. We have suggested a method for finding the site stay profile of customers, as well as application of this knowledge in deciding the value of observation symbols and initial estimate of the model parameters. It has also been explained how the HMM can identify a potential buyer and predict a customer session duration. Experimental results show the performance and effectiveness of our system and demonstrate the usefulness of learning the session duration profile of the customers. Comparative studies reveal that the Accuracy of the system is close to 80 percent over a wide variation in the input data. The system is also scalable for handling large volumes of transactions.

VI Reference

1) "Global Consumer Attitude Towards On-Line Shopping," http://www2.acnielsen.com/reports/documents/2005_cc_online shopping.pdf, Mar. 2007.
2) D.J. Hand, G. Blunt, M.G. Kelly, and N.M. Adams, "Data Mining for Fun and Profit," Statistical Science, vol. 15, no. 2, pp. 111-131, 2000.

3) Charles X. Ling and Chenhui Li : Data Mining for Direct Marketing Problems and Solutions , KDD-98
4) Peter Van Der Putten, Data Mining In Direct Marketing Databases, World Scientific, October 15, 1998
5) Data Mining in the Insurance Industry,Solving Business Problems using SAS® Enterprise Miner™ Software
6) Shailendra K, 2005, "Understanding the Relationship Between loyalty and Churn", BI Reports, December 2005
7) Berry, M.J.A., Linoff, G.S.: Mastering Data Mining. Wiley, New York (2000)
8) Adomavicius, G. Tuzhilin, A., (2001), Using data mining methods to build customer profiles. IEEE computer,Volume: 34, Issue: 2, pp.74-82
9) Reichardt, Ch., One-to-One-Marketing im Internet. Wiesbaden: Gabler verlag, 2000.
10) Greenberg, P., CRM – Customer Relationship Management at the Speed of Light. 2nd ed., Berkeley: McGraw-Hill/Osborne, 2002.
11) Feelders, A.J.; Daniels, H.A.M.; Holsheimer, M., (2000), Methodological and practical aspects of datamining, Information & Management, vol.37 , pp.271-281
12) Liu,H.;Hussain, F.;Tan, C.L&Manornjan Dash, (2002), Discretization: An Enabling Technique, Data Mining and Knowledge Discovery, Volume 6, Number 4, pp.393–423
13) Coltman, T., "Why build a customer relationship management capability?", in Journal of Strategic Information Systems, vol. 16, 2007, pp. 301-320
14) He, Z., Xu, X., Huang, J. Z., Deng, S., "Mining class outliers: concepts, algorithms and applications in CRM", in Expert Systems with Applications, vol. 27, 2004, pp. 681-697.
15) Böttcher, M., Nauck, D., Borgelt, C., & Kruse, R., (2006) , A framework for discovering interesting business changes from data, BT Technology Journal , volume 24, issue 2,pp. 219 ¬228