Sales Analysis of E-Commerce Websites using Data Mining Techniques

Anurag Bejju
Department of Computer Science
Birla Institute of Technology & Science,
Pilani, Dubai Campus, P.O.Box 345055,
Dubai International Academic City, Dubai,
UAE

ABSTRACT
In the emerging global economy, E-commerce is a strong catalyst for economic development. The rapid growth in usage of Internet and Web-based applications is decreasing operational costs of large enterprises, extending trading opportunities and lowering the financial barriers for active ecommerce participation. Many companies are restructuring their business strategies to attain maximum value in terms of profits as well as customer's satisfaction. Business tycoons around the globe are realizing that e-commerce is not just trading of products and information over Internet, rather it provides an opportunity to compete with other giants in the market. Data mining (DM) is used to attain knowledge from available information in order to help companies make weighted decisions. An organization needs to invest only on the group of products which are frequently purchased by its customers as well as price them appropriately in order to attain maximum customer satisfaction. The objective of this paper is to evaluate, propose and improve traditional pricing strategies by using web mining techniques to collect information from e- commerce websites and apply data mining methods to induce and extract useful information out of it. The proposed strategy can be generated by optimizing decision trees in an iterative process and exploit information about historical buying behavior of a customer.

Keywords
E-Commerce, Data Mining, ID3 Algorithm

1. INTRODUCTION
The Web is one of the most revolutionary technologies that changed the business environment and has a dramatic impact on the future of electronic commerce (EC). The future of EC will accelerate the shift of the power toward the consumer, which will lead to fundamental changes in the way companies relate to their customers and compete with one another. Previous studies in Information Science (IS) literature like The Consumer Behavior towards online shopping of electronics in Pakistan (Adil Bashir 2013), Online Consumer Behaviour (Dr. Bas Donkers 2013), Influencing the online consumer's behavior: the Web experience (Efthymios Constantinides 2010),) Post-purchase behavior (Dibb et al., 2004; Jobber, 2010; Boyd et al., 2012; Kotler, 2011; Brassington and Pettitt, 2013) have proposed various models explaining customer buying behavior. These research models typically derive hypotheses from a literature review. Based on this hypotheses, evaluation of a multi-channel customer choice data can bedone. Commerce networks involve buying and selling activities among individuals or organizations. [1]

Getting a deeper understanding of e-commerce networks, such as the Flipkart market space, in terms of structure, interactions, trust and reputation has tremendous value in developing business strategies and building eff ective user applications. Nowadays, web data provides comparative advantages for mass merchants to analyze and reveal important parts of online

consuming behavior [2]. This paper discusses examples of

multi-channel strategies and designs a pricing model which focus on 4 P's of Marketing mix. Based on the analysis of the retailer's transaction data and a literature review, we derive hypotheses to explain consumer purchasing behavior.

2. BACKGROUND
The E-Commerce industry represents one of the largest industries worldwide. For example, in the United States, it is the second largest industry in terms of both the number of establishments and profits, with $3.8 trillion in sales annually. [3] In addition, this industry is facing similar trends to those affecting other sectors, for instance, the globalization of markets, aggressive competition, increasing cost pressures and the rise of customized demand with high product variants.

Manual capture of sales information increases transaction costs and can cause inventory inaccuracies.

This kind of processing involves numerous human interventions at different levels such as order taking, data entry, processing of the order, invoicing and forwarding. The accuracy of the model is questionable and may not be consider few important factors while developing it. To overcome this problem, data mining can be used to analyze big data and develop efficient marketing strategies It is ideal because many of the ingredients required for successful data mining are easily satisfied: data records are plentiful, electronic collection

provides reliable data, insight can easily be turned into action, and return on investment can be measured [4].

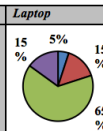
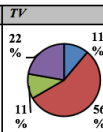
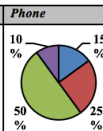
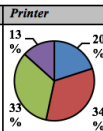
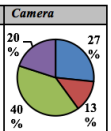
3. DATA MINING AND CONSUMER BEHAVIOR IN E-COMMERCE
In the past few years, the development of the World Wide Web exceeded all expectations. Retrieving data has become a very difficult task taking into consideration the impressive variety of the Web. Web consists of several types of data such as text data, images, audio or video, structured records such as lists or tables and hyperlinks. Web content mining can be used to mine text, graphs and pictures from a Web page and apply data mining algorithms to generate patterns used for knowledge discovery [5]. For a successful e-commerce site, reducing user-perceived latency is the second most important quality after good site- navigation quality.

The most successful approach towards reducing user-perceived latency has been the extraction of path traversal patterns from past users buying history to predict future user buying behavior and to fetch the required resources. [6] Vallamkondu & Gruenwald (2003) describe an approach to predict user behavior in e-commerce sites. The core of their approach involves extracting knowledge from integrated data of purchase and path traversal patterns of past users to develop a pricing model which focuses on profits as well as customer satisfaction. [7] Web sites are often used to establish a company's image, to promote and sell goods and to provide customer support. The success of a web site directly affects the success of the company in an electronic market.

3.1 Product Strategy

A product is anything that can be offered to a market for attention, acquisition, use, or consumption that might satisfy a want or need (Kotler, 2001). In an e-commerce marketing strategy, it is important to remember that information is now its own viable product. In the physical world, a shopper who wants to buy something has to manually sift through the millions of choices. A complete search of all offerings would be extremely expensive, time-consuming and practically impossible. Instead consumers rely on product suppliers and retailers to aid them in the search. This allows the suppliers and providers to use the consumers' cost-of search as a competitive advantage. However, on the Internet, consumers can search much more comprehensively and at virtually no cost[11].

Table 1. Rating v/s Type Comparison

Type	*Laptop*		*TV*		*Phone*		*Printer*		*Camera*	
Pie Chart										
Rating	*Count*	*%*	*Count*	*%*	*Count*	*%*	*Count*	*%*	*Count*	*%*
1 star	0	0%	0	0%	0	0%	0	0%	0	0%
2 star	15	5%	15	11%	45	15%	45	20%	60	27%
3 star	45	15%	75	56%	75	25%	75	33%	30	13%
4 star	195	65%	15	11%	150	50%	75	34%	90	40%
5 star	45	15%	30	22%	30	10%	30	13%	45	20%
Total Cost	**300**		**135**		**300**		**225**		**225**	

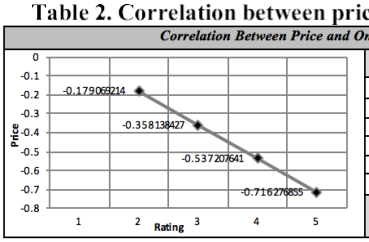
By using the direct access to consumers enabled by the Internet, companies can collect information, identify target consumers, and better introduce products or services to meet consumers' needs. If a customer finds all the desired product type it will directly affect the customer satisfaction index (refer to table 1.1). After analyzing the training set with 1185 instances, It was found that the 60% customers were satisfied with phone and TV had least customer satisfaction terms of online product rating.

3.2 Price Strategy

In the earliest days of Internet commerce, many economists and media observers predicted that competition among Internet retailers would quickly resemble perfect competition. After all, the Internet already reduces search costs relative to visiting physical stores and comparison sites could be expected to lower search costs still further. The question of how pricing impacts consumer purchasing behavior is interesting. In this paper, we

discuss one such application, measuring the potential magnitude of bias in the consumer price index arising from underweighting Internet commerce. Price is the only element of the marketing mix to generate revenues. Internet pricing decisions will be just as crucial as they traditionally have been.

Table 2. Correlation between price and online rating



Rating	*Review*
1 star	0
2 star	-0.17907
3 star	-0.35814
4 star	-0.53721
5 star	-0.71628
Correlation (Slope of graph) : -0.17907	
$Rating \propto \frac{1}{Price}$	

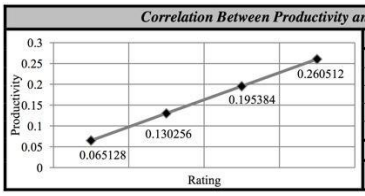
The Internet will lead to increased price competition and the standardization of prices. Also, the ability to compare prices across all suppliers using the Internet and online shopping services will lead to increased price competition. Finally, the price of providing Internet-based services often contains little or no marginal costs. Organizations will have to employ new pricing models when selling over the Internet. A negative correlation of -0.17907 was found on applying it to training set of 1185 instances. It was also found that

Rating 1/Price

3.3 Productivity Concept

Business profits can be increased by increasing revenue through stronger sales and/ or by decreasing the costs associated with constant sales. One of the major factors in customer satisfaction is the availability and timeliness of the delivery of products. If a customer has to wait to receive a product, it can be detrimental to their feeling of satisfaction. [13] With that in mind, avoiding back orders should be a major goal of any business.

Table 3. Correlation Between Productivity and Online Rating



Rating	*Productivity*
1 star	0
2 star	0.065128
3 star	0.130256
4 star	0.195384
5 star	0.260512
Correlation (Slope of graph) : 0.065128	
Rating ∝ Productivity	

Through a strong inventory or warehouse management system you will be able to use product demand forecasts and lead time tracking to ensure that your warehouse is always stocked with the necessary products at the proper times. (Refer fig 1.3) A positive correlation of 0.065128 was found on applying it to training set of 1185 instances.

Rating Productivity

Step 1: Data Pre-processing

Mined Data (Data Base)					Pre Processed Data				Is Rating =4 (Yes/ No)
Name	Type	Price	Qty	Rating	Type	Price	Qty	Rating	
HP Notebook	Laptop	17700	1100	4 star	Laptop	R2	B	4 star	Yes
Dell Inspiron	Laptop	26700	750	4 star	Laptop	R2	B	4 star	Yes
Macbook Pro	Laptop	40100	500	5 star	Laptop	R3	A	5 star	No
Sams. Galaxy	Phone	14500	1300	4 star	Phone	R1	C	4 star	Yes
Moto E	Phone	6700	525	5 star	Phone	R1	A	5 star	No
Canon 1200D	Camera	27000	1260	4 star	Camera	R2	B	4 star	Yes
Nikon D5100	Camera	24000	1440	2 star	Camera	R1	C	2 star	No
HP Deskjet	Printer	13999	820	3 star	Printer	R1	B	3 star	No
Canon Inkjet	Printer	27999	1020	3 star	Printer	R3	B	3 star	No
Epson Inkjet	Printer	30999	2450	2 star	Printer	R3	C	2 star	No
Panason. LED	TV	16899	75	3 star	TV	R2	A	3 star	No
Videocon TV	TV	19100	650	3 star	TV	R2	B	3 star	No

Discretization of Price		Discretization of Quantity (Qty)	
Condition	Assignment	Condition	Assignment
(Price>=1500)&&(Price<14600)	Range 1 (R1)	(sales>=25)&&(sales<650)	A
(Price>=14600)&&(Price<27700)	Range 2 (R2)	(sales>=650)&&(sales<1275)	B
(Price>=27700)&&(Price<40800)	Range 3 (R3)	(sales>=1275)&&(sales<2525)	C

Step 3: CALCULATIONS

Set (S) = { Y, Y, N, Y, N, Y, N, N, N, N, N, N } $P_y = \frac{4}{12}$; $P_n = \frac{8}{12}$

Entropy(S)

$$Entropy(S) = -\sum_{j=1}^{k} (P_{ij}) \log_2 (P_{ij}) = -(P_y) \log_2 (P_y) - (P_n) \log_2 (P_n) = 0.91492$$

Attribute : TYPE

Laptop = [2Y, 1N]; Phone = [1Y, 1N]; Camera = [1Y, 1N]; Printer = [0Y, 3N]; Tv = [0Y, 2N];

$$Entropy_{Laptop}(S) = -(P_y) \log_2 (P_y) - (P_n) \log_2 (P_n) = -(\tfrac{2}{3}) \log_2 (\tfrac{2}{3}) - (\tfrac{1}{3}) \log_2 (\tfrac{1}{3}) = 0.92481$$

$$Entropy_{Phone}(S) = Entropy_{Camera}(S) = 1\ (Impure\ set)\ ;\ Entropy_{Printer}(S) = Entropy_{Tv}(S) = 0\ (pure\ set)$$

$$Info\ Gain_{Attribute}(S) = Entropy(S) - (\sum_{k=1}^{n} \frac{|S_k|}{S} Entropy(S_k))$$

$$Info\ Gain_{Type}(S) = E(S) - \left[\left(\frac{3}{12} E_{Ltp}(S)\right) + \left(\frac{2}{12} E_{Cam}(S)\right) + \left(\frac{2}{12} E_{Tv}(S)\right) + \left(\frac{3}{12} E_{Prtr}(S)\right) + \left(\frac{2}{12} E_{Phne}(S)\right)\right] = 0.3517175$$

Attribute : Price

R_1 = [3Y, 0N]; R_2 = [1Y, 5N]; R_3 = [0Y, 3N];

$$Entropy_{R1}(S) = Entropy_{R3}(S) = 0\ (pure\ set)$$

$$Entropy_{R2}(S) = -(P_y) \log_2 (P_y) - (P_n) \log_2 (P_n) = -(\tfrac{1}{6}) \log_2 (\tfrac{1}{6}) - (\tfrac{5}{6}) \log_2 (\tfrac{5}{6}) = 0.63431$$

$$Info\ Gain_{Price}(S) = Entropy(S) - [\left(\frac{3}{12} Entropy_{R1}(S)\right) + \left(\frac{6}{12} Entropy_{R2}(S)\right) + \left(\frac{3}{12} Entropy_{R3}(S)\right) = 0.597765$$

Attribute : Quantity

A = [0Y, 3N]; B = [3Y, 3N]; C = [1Y, 2N];

$$Entropy_A(S) = 0\ (pure\ set) \qquad Entropy_B(S) = 1\ (impure\ set)$$

$$Entropy_C(S) = -(P_y) \log_2 (P_y) - (P_n) \log_2 (P_n) = -(\tfrac{1}{3}) \log_2 (\tfrac{1}{3}) - (\tfrac{2}{3}) \log_2 (\tfrac{2}{3}) = 0.91492$$

$$Info\ Gain_{Quantity}(S) = Entropy(S) - [\left(\frac{3}{12} Entropy_A(S)\right) + \left(\frac{6}{12} Entropy_B(S)\right) + \left(\frac{3}{12} Entropy_{RC}(S)\right) = 0.18519$$

Summary:

$\boldsymbol{Info\ Gain_{Price}(S) = 0.537765}$; $Info\ Gain_{Type}(S) = 0.3517175$; $Info\ Gain_{Quantity}(S) = 0.18519$

Fig 1.1. Application of ID3 algorithm on data extracted from e-commerce website

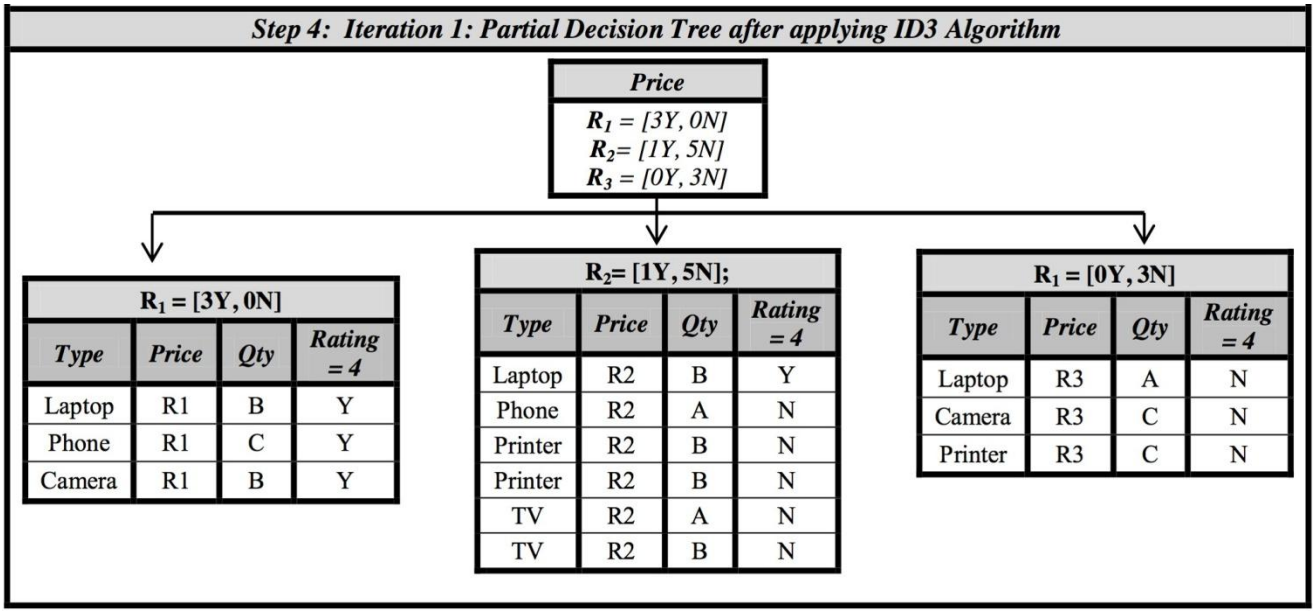


Fig 1.2. Partial Decision Tree after applying ID3 Algorithm

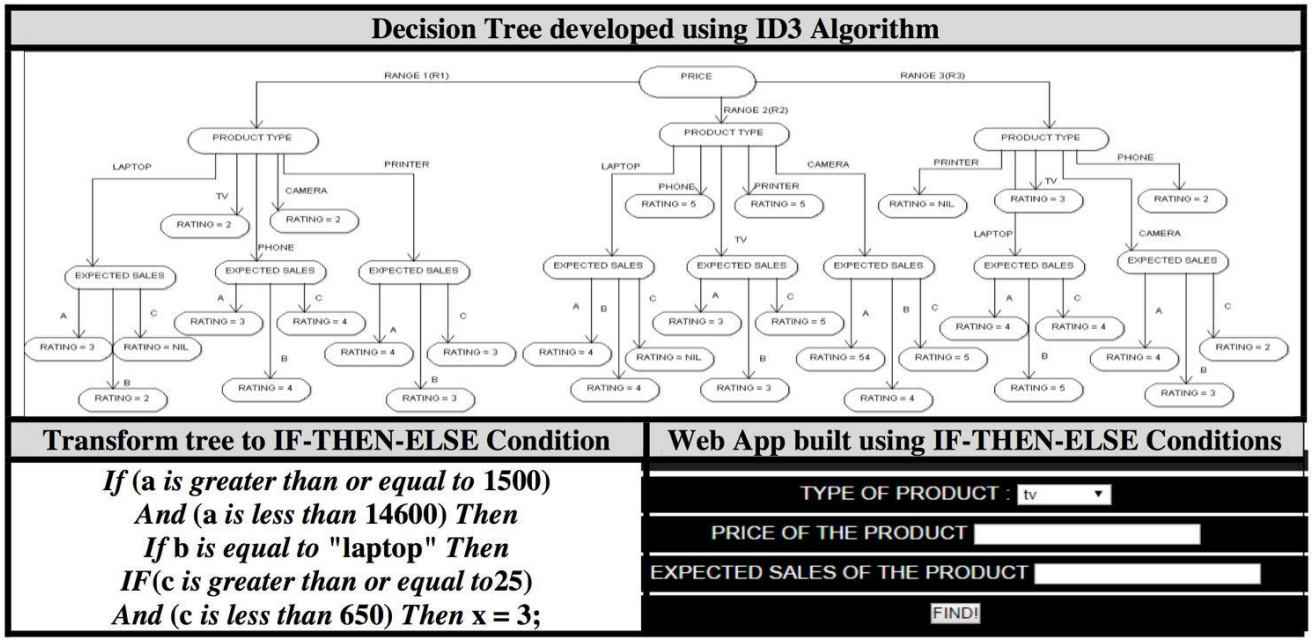


Fig 1.3. Decision Tree developed using ID3 Algorithm

4. APPLICATION OF ID3 ALGORITHM TO PREDICT ONLINE RATING OF A PRODUCT

Decision trees are used in visualization of probabilistic business models. Through generation of a tree customer's area of interest for the products can be determined. ID3 (Iterative Dichotomiser) is a simple decision tree algorithm developed by Ross Quinlan (1983). It is used to create a decision tree of given data set, by using top-down greedy approach to check each attribute at every tree node. In the decision tree method, information gain approach is generally used to determine

Performance Evaluation		
Total number of Instances in Test set	355 Instances	
Correctly Classified Instances	307	86.4789 %
Incorrectly Classified Instances	48	13.5211 %
Kappa statistic	0.802	
Mean absolute error	0.1032	
Root mean squared error	0.2276	
Relative absolute error	29.5419 %	
Root relative squared error	54.052 %	

Fig 1.4. Performance Evaluation

suitable property for each node of a generated decision tree. So, entropy of each attribute is calculated first and accordingly information Gain is calculated. Attribute which has maximum information gain set at a root node of the tree and accordingly it generates sub tree with another node. In this case, Initially,

The product name, product price, quantity, type and online rating is mined from flipkart.com website. Unique attribute values are deleted and continuous attributes like price, quantity are discretized to get effective results. Equal width binning technique divides the range of possible values into N sub ranges of the same size. [14].

5. EXTRACT CLASSIFICATION RULES

Data classification is an important data mining task that tries to identify common characteristics in a set of N objects contained in a database and to categorize them into different groups. We extract classification IF-THEN rules from those equivalence classes. For equivalence class {} , , If (a is greater

than or equal to 500 And (a is less than 14600) Then If b is equal to "laptop" Then IF(c is greater than or equal to25) And (c is less than 650) Then x = 3; can pruned by following a path in this tree. Here x is the rating which a product can get. Using these rules a web application using JavaScript, HTML and CSS is developed [15].

6. MODEL EVALUATION

The confusion matrix is a useful tool for analyzing how well your classifier can recognize tuples of different classes. TP and TN tell us when the classifier is getting things right, while FP and FN tell us when the classifier is getting things wrong mislabeling). Given m classes (where m≥2), a confusion matrix is a table of at least size m by m. An entry, CMi,j in the first m rows and m columns indicates the number of tuples of class i that were labelled by the classifier as class j. For a classifier to have good accuracy, ideally most of the tuples would be represented along the diagonal of the confusion matrix, from entry CM1,1 to entry CMm,m, with the rest of the entries being zero or close to zero In this case we m=4 so we have a 4*4 matrix. After applying ID3 algorithm this model has 86.4780% accuracy (i.e: out of every 100 test cases it has correctly predicted 87 test cases [16].

Confusion Matrix		***PREDICTED CLASS***			
		A	B	C	D
ACTUAL CLASS	A = 2 star	141	6	0	0
	B = 3 star	14	80	0	0
	C = 4 star	21	0	32	0
	D = 5 star	7	0	0	54

7. CONCLUSION

In this paper, a detailed study based on data mining techniques was conducted in order to extract knowledge in a data set with information about user's history associated to an e-commerce website. These datasets are directly mined from Flipkart.com using an online software which converts html documents to data tables. The main purpose to web mine data is to apply a set of descriptive data mining techniques to induce rules that allow data analyst working at ecommerce companies make strategic decisions to boost their sales as well as provide effective customer service. Techniques used to discover patterns are web mining and decision tree algorithms. In the future, this study can be used to analyze ecommerce websites and obtain interesting knowledge to further the companies' profits.

Many of the e-commerce strategy frameworks offer a unique contribution to strategic planning but with limited solution. This model based on web mining integrates the McCarthy's 4Ps to provide a complete analysis of e-business strategies. Thus managers can use an organized and precise process to make more successful and effective decisions. An aggressive competition has been observed in market space among the companies, thus accelerating the consumer dynamics. E-commerce will lead to increased price competition and this web application will provide an efficient way to price a particular product. It was found that price, product and production had an impact on online customer ratings. This model considers these three attributes which are correlated to customer satisfaction and help the marketer make an informed decision.

8. REFERENCES

[1] Han, Jiawei, Micheline Kamber, and Jian Pei. "Data mining: concepts and techniques" Morgan kaufmann, 2006.

[2] Quinlan J. R. (1986). "Induction of decision trees.Machine Learning," Vol.1-1, pp. 81-106.

[3] J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, Inc., 1993.

[4] Ding Xiang-wu and Wang Bin, "An Improved Pre-pruning Algorithm Based on ID3," Jisuanji Yuxiandaihua,Vol.9, pp. 47,2008.

[5] Ming Fan, Xiaofeng Meng translated, "Data mining techniques and concepts", Machinary Industry Press, Beijing, pp. 136-145, Feb., 2004.

[6] N R Srinivasa Raghavan, "Data mining in e-commerce: A survey," Sadhana Vol. 30, Parts 2 & 3, April/June 2005, pp.275–289.

[7] B. Schafer, J.A. Konstan, and J. Reidl, "E-Commerce Recommendation Applications," Data Mining and Knowledge Discovery, Kluwer Academic, 2001, pp. 115-153.

[8] P. Resnick et al., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," Proc. ACM 1994 Conf. Computer Supported Cooperative Work, ACM Press, 1994, pp. 175-186.

[9] J. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence, Morgan Kaufmann, 1998, pp. 43-52.

[10] Baesens, B. Verstreeten, G. Poel, D. (2004). Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. European Journal of Operation Research, 156, 508-523.

[11] Cheng, C.H. Chen, Y.S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. Expert System with Applications, 36, 3, 41764184. [12] Hwang, H., Jung, T. Suh, E., (2004). An LTV model and customer segmentation based on customer value