World Academy of Science, Engineering and Technology
International Journal of Economics and Management Engineering
Vol:5, No:8, 2011

Customer Need Type Classification Model using Data Mining Techniques for Recommender Systems

Kyoung-jae Kim

Open Science Index, Economics and Management Engineering Vol:5, No:8, 2011 publications.waset.org/5197.pdf

***Abstract*—**Recommender systems are usually regarded as an important marketing tool in the e-commerce. They use important information about users to facilitate accurate recommendation. The information includes user context such as location, time and interest for personalization of mobile users. We can easily collect information about location and time because mobile devices communicate with the base station of the service provider. However, information about user interest can't be easily collected because user interest can not be captured automatically without user's approval process. User interest usually represented as a need. In this study, we classify needs into two types according to prior research. This study investigates the usefulness of data mining techniques for classifying user need type for recommendation systems. We employ several data mining techniques including artificial neural networks, decision trees, case-based reasoning, and multivariate discriminant analysis. Experimental results show that CHAID algorithm outperforms other models for classifying user need type. This study performs McNemar test to examine the statistical significance of the differences of classification results. The results of McNemar test also show that CHAID performs better than the other models with statistical significance.

***Keywords*—**Customer need type, Data mining techniques, Recommender system, Personalization, Mobile user.

I. Introduction

RECOMMENDER systems are now regarded as an important marketing tool in the e-commerce because many users who sue e-commerce suffer serious information overload. They can filter and provide useful information to customers. Recently, many researchers proposed several kinds of context-aware recommender systems.

Barkuus & Dey [1] categorized context-aware services as three common categories. The first one, a simple context-aware service, accepts personal preference and context data only from corresponding customers. The second one, an inactive context-aware service, gets the customer's current context. However, it can start its service only after the user's approval process. The last one, an active context-aware service, is similar to the inactive service. However, the active context-aware service can start its service without the customer's approval.

In general, recommender systems use some important information about users to facilitate accurate recommendation. Schilke et al. [2] proposed three dimensions, location, time, and interest, for personalization of mobile users. The location and time dimensions use information about the user's position and time from mobile devices. The interest dimension considers user preferences to match relevant products or services. We can easily collect information about location and time because mobile devices communicate with the base station of the service provider. However, user interest can't be easily collected because user interest can not be captured automatically without user's approval process. User interest usually represented as a need.

A need is something that is necessary for humans' healthy life [http://www.wikipedia.org]. In general, needs are defined as requirements for something essential or desirable that is lacking. That is, needs are the most basic factors and the starting point of the generating process of behavioral outcomes. Therefore, understanding a user's needs is quite important to satisfy the user. Prior research in the marketing context has identified numerous kinds of needs that influence the process of stimulating user behavior. However, we may classify them into two types: utilitarian and hedonic [3].

Utilitarian needs are explained as requirements for products that remove or avoid problems with life, while hedonic ones are requirements for products that provide social or aesthetic utility to users. For example, a user who participates in an online social network to obtain useful information for his/her life has utilitarian needs, but the user has hedonic needs when he/she uses it for a social relationship or amusement. Users are generally conscious of the needs stimulated by advertisements. Thus, advertisers can use utilitarian or hedonic appeals to stimulate users' utilitarian or hedonic needs.

This study investigates the usefulness of data mining techniques for classifying user need type for recommendation systems. We employ several data mining techniques including artificial neural networks, decision trees, case-based reasoning, and multivariate discriminant analysis.

The remainder of the paper is organized as follows. Section 2 reviews the basic concepts of data mining techniques in this study. Section 3 describes research data and experimental design. Section 4 presents experimental results. In the final section, the conclusions of the study are presented.

II. Basic Concepts of Data Mining Techniques

A. Artificial Neural Networks

This study uses three-layer back-propagation neural network model. The model is most popular for the purpose of business application. The basic algorithm of the model is well-known for the researcher, so this study do not mention about the

Kyoung-jae Kim is with Dongguk University-Seoul, South Korea (phone: +82-2-2260-3324; fax: +82-2-2260-3684; e-mail: kjkim@ dongguk.edu).

Open Science Index, Economics and Management Engineering Vol:5, No:8, 2011 publications.waset.org/5197.pdf

algorithm.

B. MDA

Multiple discriminant analysis (MDA) is a method for compressing a multivariate signals to produce a lower dimensional signal amenable to classification [4]. MDA finds the criteria for determining population membership of data using the information about each data. It forecasts membership of each component by use of discriminant function which is derived from the characteristics two predictable classes.

C. CART

Classification and regression tree (CART) allocates components to some classes according to the resulting tree. The term CART analysis is first introduced by Breiman et al. [5]. It is an umbrella term used to refer to both of classification and regression tree procedures.

D. CHAID

CHAID is a kind of decision tree technique, based on adjusted significance testing. It was developed by Kass [6]. CHAID stands for CHi-squared Automatic Interaction Detector because it uses the Chi-square test for determining the splits in the resulting decision tree. It generally detects interaction between variables in the data set. Using this technique it is possible to construct relationships between a dependent variable and other independent variables.

E. QUEST

QUEST is a binary-split decision tree algorithm for classification developed by Loh & Shih [7]. QUEST stands for Quick, Unbiased and Efficient Statistical Tree. The basic objective and algorithm of QUEST is similar to that of the CART algorithm. However, QUEST uses an unbiased variable selection technique by default and uses imputation instead of surrogate splits to deal with missing values.

F. CBR

CBR stands for Case-Based Reasoning. CBR is a reasoning technique that reuses past cases to find a solution to the new problem. In general, it selects k-nearest cases from population using similarity measure (usually Euclidean distance), then composites the solutions of selected cases and produces solution for the new case. In general, it is called as the k-nearest neighbor algorithm.

III. Research Data And Experiments

In this study, we need information on location, time, and user's needs type to predict other user's needs type. We built a Web-based data collection system to gather appropriate empirical data from users. This data collection system contained the places for shopping, eating, drinking, enjoyment, and learning in five major commercial locations of Seoul, South Korea. The system contained the information on 275 places in the Chongro, Daehakro, Shinchon/Ewha Univ., Kangnam Station, and Myungdong areas.

The data collection system was designed to collect data including the visiting day, visiting time, and user's needs type at the point of visit for these spots from mobile phone users. To simplify the input process, we discretized the candidate values of the input variables, as presented in Table 1. As shown in Table 1, we assign the numeric code in an interval scale to each candidate value of the most input variables (visiting time and needs type). The needs type is categorized as three groups including hedonic, utilitarian needs and both. Finally, it is possible to apply simple numeric operations for the inputted values.

TABLE I
DESCRIPTION OF THE VARIABLES

Dimension	Variable	Candidate values
Location	Commercial Zone (CZ)	Chongro Daehakro Shinchon / Ewha Univ. Kangnam Station Myungdong
Time	Visiting day (VD)	Weekday (Mon.-Fri.) Weekend (Sat./Sun.)
	Visiting time (VT)	Morning / AM08:00 – AM11:00 Lunch / AM11:00 – PM02:00 Afternoon / PM02:00 – PM05:00 Dinner / PM05:00 – PM08:00 Night / PM08:00 – PM11:00
Interest	Needs type (NT)	Hedonic (H) Utilitarian (U) Both (B)

We collected the experimental from April to May 2006. In the two months, we collected 9980 ratings from 265 respondents in three universities in Korea. We deleted some cases that were inappropriate, and finally selected 200 respondents and their data for 3360 visits for the experimental dataset. We split the data as two or three (for artificial neural networks model) sub-data sets such as modeling and validation (hold-out) data sets for all models except artificial neural networks model or training, test, and validation data set for artificial neural networks model. The ratio of data is 7:3 for two sub-data sets and 6:1:3 for three sub-data sets.

In this study, we try to estimate the user's needs type by using the information of the target user, and the background (location and time). Thus, the algorithm to estimate the user's needs type should be developed. We adopted several data mining techniques including artificial neural networks, MDA, CART, CHAID, QUEST, and CBR (k-NN) for this. Five variables were used as independent variables – (1) visiting day, (2) visiting time, (3) location (area), (4) the user's gender, and (5) the type of the place (shopping mall, restaurant, etc.).

The experimental software was SPSS 17.0 and its add-ins for artificial neural networks and decision trees.

IV. Experimental Results

In this study, we set the hit ratio as the criterion to evaluate the performance of the comparative models. Hit ratio is frequently used in the data mining literature, and represents the forecasting accuracy of the model. As mentioned above, the number of categories for dependent variable (need type) is three including hedonic, utilitarian need and both.

Open Science Index, Economics and Management Engineering Vol:5, No:8, 2011 publications.waset.org/5197.pdf

First, we experiment artificial neural networks model. As mentioned earlier, this study uses typical three-layer back-propagation model. We set the range of processing elements of hidden layer as 3, 5, and 10 because this study employs 5 input variables. Table 2 shows experimental results of the artificial neural networks models.

TABLE II
THE RESULTS OF THE ARTIFICIAL NEURAL NETWORKS MODELS

	Training Data	Test Data	Validation Data
3 PEs in hidden layer	46.2%	42.9%	42.5%
5 PEs in hidden layer	47.3%	42.0%	44.0%
10 PEs in hidden layer	46.2%	43.5%	43.7%

* PE stands for processing elements

As shown in Table 2, the hit ratio for the 5PEs model outperforms the other two models for the validation data set. Fig. 1 shows the graphical presentation of the best model.

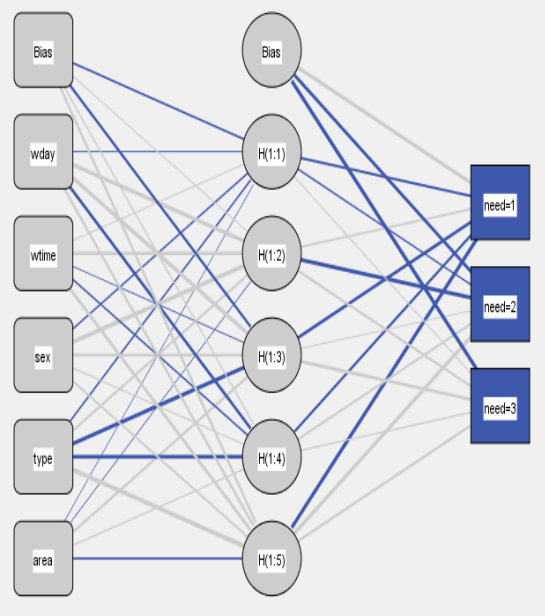

Fig. 1 Graphical presentation of ANN model

In addition, Table III shows experimental output of the best model.

TABLE III
THE RESULTS OF THE BEST ANN MODEL

Sample	Observed	Predicted			
		1	2	3	Percent Correct
Training	1 (U)	402	97	173	59.8%
	2 (H)	227	264	181	39.3%
	3 (B)	241	143	288	42.9%
	Overall Percent	43.2%	25.0%	31.8%	47.3%
Testing	1	59	28	25	52.7%
	2	30	51	31	45.5%
	3	33	32	47	42.0%
	Overall Percent	36.3%	33.0%	30.7%	46.7%
Validation	1	194	55	87	57.7%
	2	125	124	87	36.9%
	3	128	82	126	37.5%
	Overall Percent	44.3%	25.9%	29.8%	44.0%

Second, this study experiments MDA model. We use stepwise selection method for the input variable selection. The results of the MDA model are shown in Table 4.

TABLE IV
THE RESULTS OF THE MDA MODEL

Sample		need	Predicted Group Membership			Total
			1	2	3	
Modeling (Training and Test)	Original	1	386	173	225	784
		2	226	339	219	784
		3	190	216	378	784
Validation	Original	1	150	77	109	336
		2	100	140	96	336
		3	97	92	147	336

The hit ratio of the MDA Model is 46.9% for modeling data set and 43.4% for validation data set.

Third, we implement CART model. This study appropriately prunes the resulting tree for preventing over-fitting problem. Fig. 2 shows the resulting CART after experiments.

Open Science Index, Economics and Management Engineering Vol:5, No:8, 2011 publications.waset.org/5197.pdf

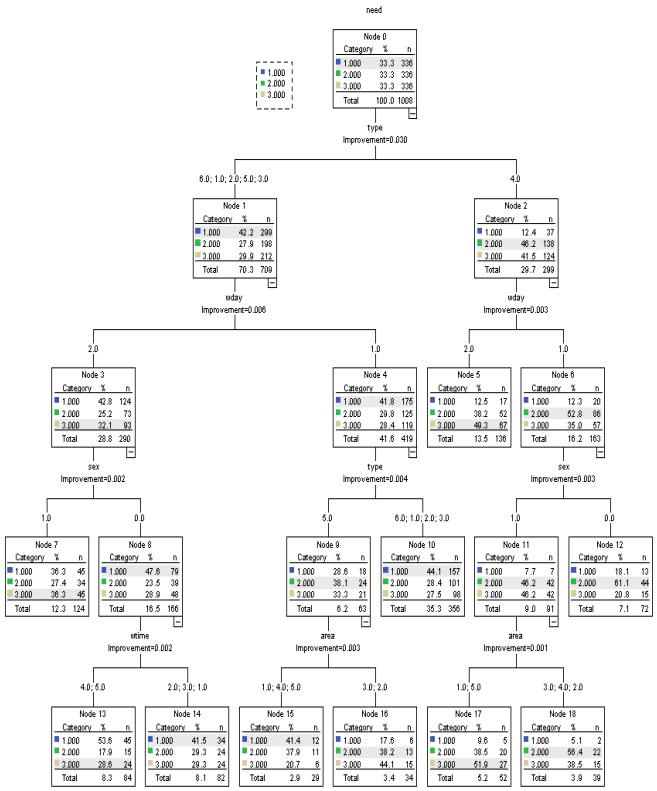

Fig. 2 The best CART model

In addition, Table V presents the experimental results of the best CART model.

TABLE V
THE RESULTS OF THE CART MODEL

Sample	Observed	Predicted			
		1	2	3	Percent Correct
Modeling	1 (U)	503	51	230	64.2%
	2 (H)	300	215	269	27.4%
	3 (B)	252	101	431	55.0%
	Overall Percent	44.9%	15.6%	39.5%	48.9%
Validation	1	203	21	112	60.4%
	2	136	79	121	23.5%
	3	128	45	163	48.5%
	Overall Percent	46.3%	14.4%	39.3%	44.1%

As shown in Table V, the hit ratio of the best CART model is 44.1% for the validation data set.

Fourth, this study experiments CHAID model using Chi-square statistics. The resulting model is depicted as Fig. 3.

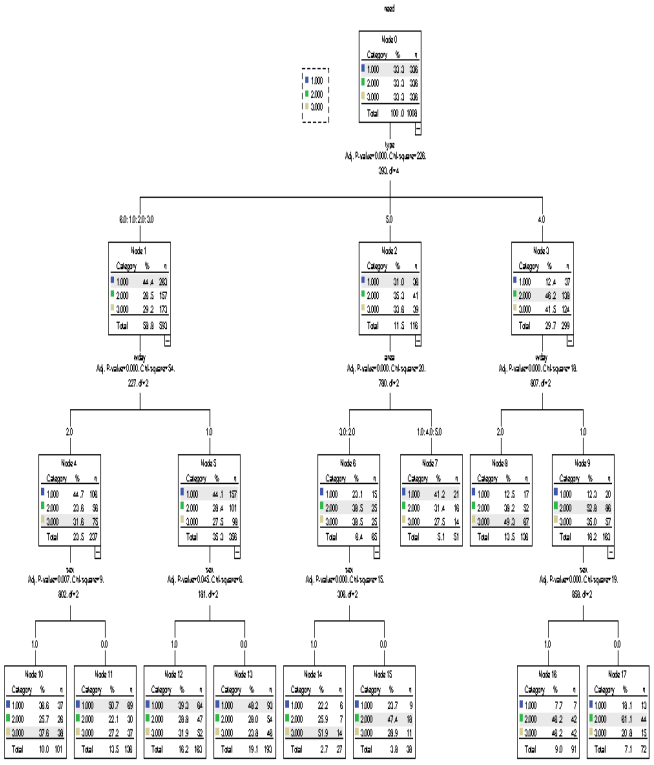

Fig. 3 The best CHAID model

The experimental results of the best CHAID model summarized as Table VI.

TABLE VI
THE RESULTS OF THE CHAID MODEL

Sample	Observed	Predicted			
		1	2	3	Percent Correct
Modeling	1 (U)	566	69	149	72.2%
	2 (H)	325	243	216	31.0%
	3 (B)	329	129	326	41.6%
	Overall Percent	51.9%	18.8%	29.4%	48.3%
Validation	1	247	29	60	73.5%
	2	147	104	85	31.0%
	3	149	68	119	35.4%
	Overall Percent	53.9%	19.9%	26.2%	46.6%

The results show that the hit ratio of the best CHAID model is 46.6% for the validation data set and 48.3% for the modeling data set.

Fifth, we implement QUEST algorithm for the research data. The resulting tree is presented as Fig. 4.

Open Science Index, Economics and Management Engineering Vol:5, No:8, 2011 publications.waset.org/5197.pdf

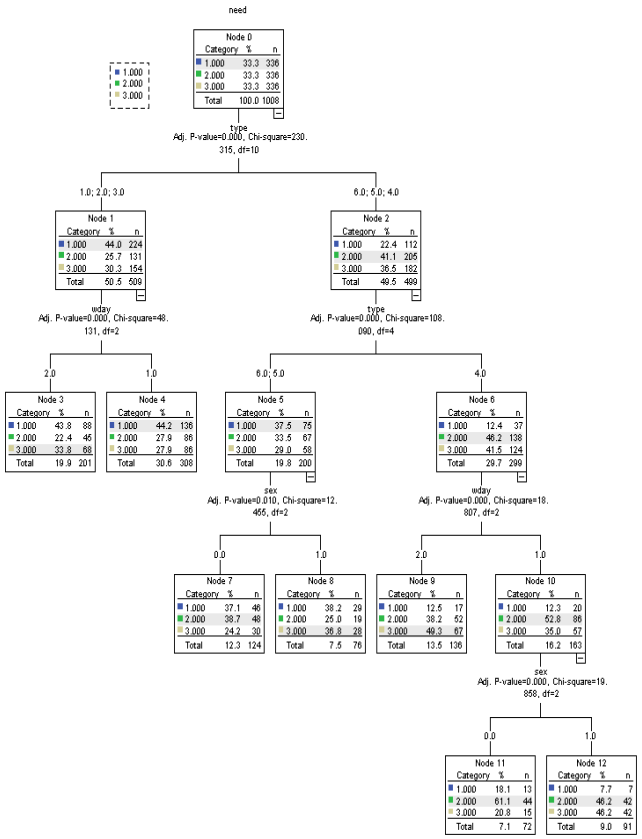

Fig. 4 The best QUEST model

In addition, Table 7 presents the experimental results of the best QUEST model.

TABLE VII
THE RESULTS OF THE QUEST MODEL

Sample	Observed	Predicted			
		1	2	3	Percent Correct
Modeling	1 (U)	351	130	303	44.8%
	2 (H)	185	294	305	37.5%
	3 (B)	159	176	449	57.3%
	Overall Percent	29.5%	25.5%	44.9%	46.5%
Validation	1	136	66	134	40.5%
	2	86	134	116	39.9%
	3	86	87	163	48.5%
	Overall Percent	30.6%	28.5%	41.0%	43.0%

As shown in Table 7, the hit ratio of the best QUEST model is 43.0% for the validation data set.

Sixth, this study experiments CBR (k-NN) model using the k-nearest neighbor algorithm. For the CBR experiments, we vary the range of k-nearest neighbor as 1-15. The hit ratio of the model is best when the k is 10. Table 8 presents the experimental results of the best CBR model

TABLE VIII
THE RESULTS OF THE CBR MODEL (K=10)

Sample	Observed	Predicted			
		1	2	3	Percent Correct
Modeling	1 (U)	472	127	185	60.2%
	2 (H)	266	314	204	40.1%
	3 (B)	256	186	342	43.6%
	Overall Percent	42.3%	26.7%	31.1%	48.0%
Validation	1	195	53	88	58.0%
	2	119	118	99	35.1%
	3	124	85	127	37.8%
	Overall Percent	43.5%	25.4%	31.2%	43.7%

Table 8 shows the overall hit ratio of the best CBR model is 43.7% for the validation data set.

Finally, we summarize the experimental results of all comparative models as Table 9.

TABLE IX
OVERALL PERFORMANCES OF ALL COMPARATIVE MODELS

Model	ANN	MDA	CBR	CART	QUEST	CHAID
Hit Ratio	44.0%	43.4%	43.7%	44.1%	43.0%	46.6%

The results show that CHAID model outperforms the other comparative models. The second best model is CART and the worst model is QUEST.

In addition, the McNemar tests are used to examine whether the best model significantly outperforms the other models. This test is a nonparametric test for two related samples. This test may be used with nominal data and is particularly useful with before-after measurement of the same subjects [8]. Table 10 shows the results of the McNemar test.

TABLE V
MCNEMAR VALUES FOR THE PAIRWISE COMPARISON OF PERFORMANCE

Model	MDA	CBR	CART	QUEST	CHAID
ANN	0.343	0.035	0.000	0.461	2.790*
MDA		0.019	0.292	0.063	4.146**
CBR			0.072	0.127	2.731*
CART				0.637	4.267**
QUEST					7.005***

***significant at the 1% level, **significant at the 5% level, * significant at the 10% level

As shown in Table 10, CHAID outperforms the other models with statistical significance. In especial, CHAID outperforms QUEST with 1% statistical significance level, CART and MDA with 5% level, and ANN and CBR with 10% level.

V. Conclusions

In this study, we compared several data mining techniques for the prediction of user needs type. The experimental results showed CHAID performs better than other comparative models with statistical significance. Thus, we may conclude that CHAID model generates the most accurate prediction results for the inference of user context. The research findings may be used to build the active context-aware recommender systems for mobile users. Our study also has some limitations. The usefulness of CHAID should be validated in practice. The validation process in our study is quite restricted, because our model is not validated in the real-world mobile situation, although the experimental validation is performed using the data collected from real-world users. Thus, we hope to have a chance to implement and validate the model practically with a real-world mobile service provider in the future.

References

[1] Barkuus, L, Dey, A (2003) Is context-aware computing taking control away from the user? Three levels of interactivity examined. In: Proceedings of the Ubicomp, pp 150–156

[2] Schilke, SW, Bleimann, U, Furnell, SM, Phippen, AD (2004) Multi-dimensional personalisation for location and interest-based recommendation. Internet Research 14(5):379-385

[3] Maclnnis, DJ, Jaworski, BJ (1989) Information processing from advertisements: Toward an integrative framework. Journal of Marketing 53(4):1-23

[4] Duda R, Hart P, Stork D (2001) Pattern Classification. 2 Ed. John Wiley and Sons, NY

[5] Breiman, L, Friedman, JH, Olshen, RA, Stone, CJ (1984) Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA

[6] Kass, GV (1980) An exploratory technique for investigating large quantities of categorical data. Applied Statistics 29(2):119–127

[7] Loh, WY, Shih, YS (1997) Split selection methods for classification trees. Statistica Sinica 7:815-840

[8] Cooper, DR, Emory, CW (1995) Business Research Methods. Irwin, IL

Open Science Index, Economics and Management Engineering Vol:5, No:8, 2011 publications.waset.org/5197.pdf