A Personalized Recommendation Methodology based on Collaborative Filtering*

Jae Kyeong Kim
School of Business Administration, Kyung Hee University
(jaek@khu.ac.kr)

Ji Hae Suh
School of Business Administration, Kyung Hee University
(jaek@khu.ac.kr)

Do Hyun Ahn
School of Business Administration, Kyung Hee University
(jaek@khu.ac.kr)

Yoon Ho Cho
Department of Internet Information, Dongyang Technical College
(yhcho@dongyang.ac.kr)

The rapid growth of e-commerce has made both companies and customers face a new situation. Whereas companies have become to be harder to survive due to more and more competitions, the opportunity for customers to choose among more and more products has increased. So, the recommender systems that recommend suitable products to the customer have an important position in E-commerce. This research introduces collaborative filtering based recommender system which helps customers find the products they would like to purchase by producing a list of top-N recommended products. The suggested methodology is based on decision tree, product taxonomy, and association rule mining. Decision tree is used to select target customers, who have high possibility of purchasing recommended products.

We applied the recommender system to a Korean department store. The methodology is evaluated with the analysis of a real department store case and is compared with other methodologies.

Key words: Recommendation system, Personalization, Collaborative filtering, Association Rule Mining, Decision Tree, Data Mining

1. Introduction

E-commerce has been growing rapidly, keeping the pace with the web. However, its rapid growth has made both companies and customers face a new situation. Whereas companies have become to be harder to survive due to more and more competitions, the opportunity for customers to choose among more and more products has increased(Kim, et al., 2000; Schafer, et al., 2001). As a result, the need for new marketing strategies such as one-to-one marketing, web personalization, and customer relationship management(CRM) have been stressed from researchers as well as from practical affairs(Sarwar, et al., 2000; Mobasher, et

* 이 연구는 2002학년도 경희대학교 연구비 지원에 의하여 이루어졌음.

al., 2000; Berson, et al., 2000; Changchien & Lu, 2001; Yuan & Chang, 2001).

One solution to achieve these goals in e-commerce is the use of recommender systems. Recommender system is a personalized information filtering technology used to help customers find the products they would like to purchase by producing a list of top-N recommended products for a given customer. Several hybrid approaches that combined content-based filtering and collaborative filtering have been proposed currently(Cho et al., 2002).

Fab is a well-known hybrid content-based collaborative systems for recommending web pages(Balabanovic & Shoham, 1997). User profiles based on the pages a user liked are maintained by using content-based techniques. The profiles are directly compared to determine similarity between users in order to make collaborative filtering predictions. GroupLens implements a hybrid collaborative filtering systems for Usenet news that support content-based "filterbots"(Sarwar, et al., 1998) that evaluate and enter ratings for articles as soon as they are published. Ptango(Claypool, et al., 1999) combines content-based and collaborative filtering for an online newspaper. The user profile is made up of both explicit keyword entered by the user and implicit keywords gathered by automatically from articles. Lawrence, et al.(2001) developed a personalized recommender system designed to suggest new grocery products to supermarket shoppers who use Personal Digital Assistants(PDAs). Core of the system is a content-based filtering that generates recommendations by matching products to customers based on the expected appeal of the product and the previous spending of the customer. The system also applies the collaborative filtering to two sources of information both to refine the content model and to make recommendations.

But, there are common problems on existing recommender systems. One problem is sparsity. In practice, many commercial recommender systems are used to evaluate large product sets(e.g., Amazon.com recommends books and Cdnow.com recommends music albums). Another one is scalability. All the recommender systems require computation that grows with both the number of customers and the number of products. With millions of customers and products, a typical web-based recommender system running existing algorithms will suffer serious scalability problem (Sarwar, et al., 2000).

In this study, we propose a new hybrid methodology for personalized recommendations to overcome above problems of existing systems. For such a purpose, the suggested methodology is based on collaborative filtering, decision tree, product taxonomy, and association rule mining. Decision tree is to select target customers, who have high possibility of purchasing recommended products. Product taxonomy is used to solve scalability and sparsity. We applied the recommender system to Korean department store. The methodology is evaluated with the analysis of a real Korean department store case.

2. Backgrounds

2.1 Recommender Systems Using Collaborative Filtering

Collaborative filtering(CF) is known to be the most successful recommender system technology, and used in many of the successful recommender systems. CF systems recommend products to the target customer based on the opinions of other like-minded customers. These systems employ statistical techniques to find a set of customers known as neighbor who have a similarity with the target user. When neighbor is formed, these systems use several algorithms to produce recommendations. In general, the entire recommendation procedure of CF systems is divided into sub-processes namely, representation process, neighbor formation process, and recommendation generation process(Sarwar, et al., 2000).

(1) Representation

In a typical CF-based recommender system, the input data is a collection of historical purchasing transaction of *n* customer on *m* products. It is usually represented as an $m \times n$ customer-product matrix, R, such that $r_{i,j}$ is one if the i^{th} customer purchased the j^{th} product, otherwise is zero.

(2) Neighbor Formation

The most important step in CF-based recommender systems is that of computing the similarity between customers. It is used to form a proximity-based neighbor between the target customer and a number of like-minded customers. The neighbor formation process is the model-building or learning process for a recommender system algorithm. The main goal of neighbor formation is to find, for each customer *u*, an ordered list of *l* customers $M=[M_1, M_2, ..., M_l]$ such that $sim(u, M_1)$ is maximum, $sim(u, M_2)$ is the next maximum and so on. The proximity between two customers is usually measured using either the Cosine measure or Pearson correlation measure. Two measures are reported to give almost the same performance(Lawrence, et al., 2001; Sawar, et al., 1998). The following is the Pearson Correlation formula for customer a, b,

$$corr_{ab} = \frac{\sum_i (r_{ai} - \bar{r}_a)(r_{bi} - \bar{r}_b)}{\sqrt{\sum_i (r_{ai} - \bar{r}_a)^2 \sum_i (r_{bi} - \bar{r}_b)^2}}$$

(3) Recommendation Generation

The final step of a CF-based recommender system is to derive the *top*-N recommendation products from the neighbor. Product recommendation is usually based on *Most-frequent item set* and *association rule mining*. Most-frequent item recommendation looks into the neighbor *N* and scans through his/her purchase data and performs a frequency count of the products. Association rule mining is the discovery of all association rules higher than a user-specified minimum support and minimum confidence

2.2 Recommender Systems Using Association Rule Mining

Knowledge Discovery in Database (KDD) is interested in devising methods for making product recommendation to customers based on different techniques. One of the most commonly used data mining techniques for E-commerce is finding association rules between a set of co-purchased products.

(1) Association Rule Mining

Association rule mining (Agrawal, et al., 1993) is concerned with discovering association rules between two sets of products. More formally, let us denote a collection of *m* product P={P_1, P_2,..., P_m}. A transaction T⊆P is defined to be a set of products purchased together. An association rule between two sets of products X, and Y, such that X, Y⊆P, X∩Y =∅, states that the presence of products in a set X and transaction T indicates a strong association with products from the set Y which is also present in T. Such an association rule is often denoted by X⇒Y.

The quality of association rules is commonly evaluated by looking at *support* and *confidence.* The support *s* of a rule measures the occurrence frequency of the pattern in the rule. The confidence c is the measurement of the strength of implication. For a rule X⇒Y, the support s, is measured by the fraction of transactions that contains both X and Y,

$$S = \frac{\textit{number of transactions containing } X \cup Y}{\textit{number of transactions}} \quad .$$

In other words, support value indicates that s% of total transactions contain $X \cup Y$. For a rule X⇒Y, the confidence c, states that c% of transactions that contain X also contain Y,

$$C = \frac{\textit{number of transactions containing } X \cup Y}{\textit{number of transactions } X} \quad .$$

More formally, confidence is nothing but the conditional probability of seeing Y, given that we have seen X. On association rule mining, it is common to find rules that have support and confidence higher than a user-defined minimum threshold.

(2) Association rule Mining as a Product Class Unit

Brand product unit based association rule derives more specific and special information than product class level based association rule. When the number of rules higher than a predefined threshold is a few, the minimum support is set to be lower to find more rules. For researching strong association rules, it must be conducted at class level unit. Many algorithms are suggested to find multi-level association rules, which may find more beneficial rules than others (Agrawal, et al., 1993; Han & Fu, 1999; Han & Kamber, 2001)..

(3) Generation of top N Recommendation Products

Association rules may be used to develop *top-N* product recommendations in the following way. First, we find the rules higher than a pre-defined support value. Next, we sort these products based on the confidence of the rules that were used to predict them. Note that if a particular product is predicted by multiple rules, we use the rule that has the highest confidence value. Finally, we select the *N* highest ranked products as the recommendation set.

2.3 Combining Collaborative Filtering and Association Rule Mining

Lawrence, et al. (2001) proposed a personalized recommender system which combines collaborative filtering and association rule mining. This study segments the customer group as similar purchase history groups to recommend customer preferable products to individual customer. By using purchase history data, all customers are divided by Spending Vector. Spending Vector of customer $C^{(m)}$ is defined as follows.

$$C^{(m)} = [C_{m1}, C_{m2}, \ldots, C_{ms}, \ldots, C_{mS}]^T, m = 1, \ldots, M.$$

C_{ms} denotes frequency of product class s, M is a total customer count, and S is product class count. This Spending Vector is normalized as following formula,

$$\hat{C}_{ms} = \frac{C_{ms}}{\sum_{s'=1}^{S} C_{ms'}}.$$

When clustering is finished, list of most frequent purchase products is generated. This product list is delivered to matching engine. This matching engine performs to rank the product. Also, matching engine reflects similarity of spending vector and product vector, and gives a score to each product. Association rule mining is used as matching algorithm of matching engine to find the association rule of product class.

About each product n , product vector $P^{(n)}$ is defined as follows,

$$P^{(n)} = [P_1^{(n)}, P_2^{(n)}, \ldots, P_s^{(n)}, \ldots, P_S^{(n)}]^T \quad n = 1, \ldots, N,$$

$S(n)$ is defined as a product class of product n, and $C(s)$ is defined as one level upper product class of product class s. Then $P_s^{(n)}$ is defined as follows.

$$P_s^{(n)} = \begin{cases} 1.0 & : \text{if } s = S(n). \\ 1.0 & : \text{if } S(n) \Rightarrow \text{s}. \\ 0.5 & : \text{if } C(s) = C(S(n)), \\ 0.25 & : \text{if } C(S(n)) \Rightarrow C(s), \\ 0.0 & : \text{otherwise.} \end{cases}$$

$P_s^{(n)}$ means how much product n and product class s are associated. Similarity score σ_{mn} about customer m and product n, is calculated by using cosine coefficient of $C^{(m)}$ and $P^{(n)}$,

$$\sigma_{mn} = \rho_n \frac{C^{(m)} \cdot P^{(n)}}{\left\|C^{(m)}\right\| \left\|P^{(n)}\right\|}, \quad where\ \rho_n = \left[\frac{PM_n}{\overline{PM}}\right]^{\alpha}.$$

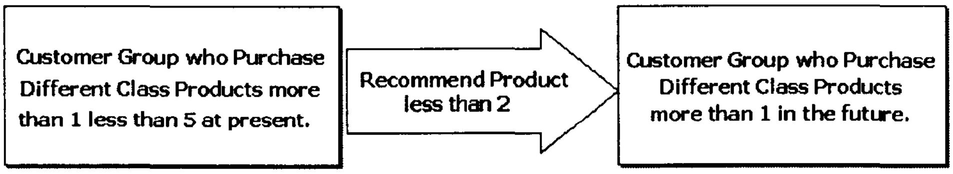
ρ_n is a modulation factor, so it is used to reflect the marketing condition like inventory maintenance cost and profit. Product vector generates it as a class unit. PM_n is a profit margin about product n and, PM_n is a mean profit about all products. Also is an empirical factor to control the degree of effect.

3. Methodology

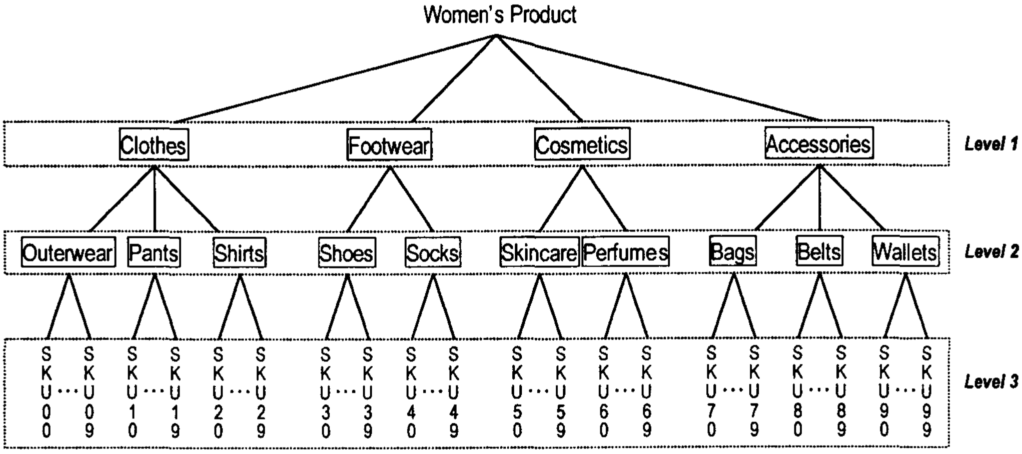
3.1 Product Recommendation

Product recommendation is a service that recommend proper product to the customer after analyzing purchase behavior. Proper product recommendation service can contribute to increase purchase by suggesting profer products to the customer. Product recommendation recommends less than N different products to the customer group purchasing more than P, less than P+I products from the different product classes. From this, the customer group can be induced to purchase the different class products more than 1. For instance, product recommendation is presented in Figure 1 in case of p=1, I=4, N=2.

A product taxonomy is practically represented as a tree form that classifies a set of low-level product into higher-level, a more general product. The leaves of the tree denote the product instances, SKUs (Stock Keeping Units) in retail jargon, and non-leaf nodes denote product classes obtained by combining several lower-level nodes into one parent node. The root node labeled by *Women's Product* denotes one of the most general product classes. Figure 2 shows an example of such taxonomy for a department store, where "Clothes", "Footwear", "Cosmetics", and "Accessories" are classified into "Women's Product", and so on.


Customer Group who Purchase Different Class Products more than 1 less than 5 at present.
Recommend Product less than 2
Customer Group who Purchase Different Class Products more than 1 in the future.

<Figure 1> Example of Product Recommendation

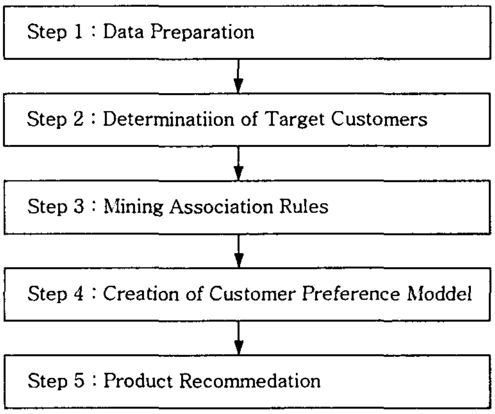


<Figure 2> Example of Product Taxonomy

The product taxonomy plays an important role for data mining analysis because choosing the higher levels of the product taxonomy may lead to improve the results of the analysis (Berry & Linoff, 2000). Several terms related to the tree, such as leaf node, non-leaf node, parent, child, ancestor, descendant, etc., are used under their original meaning. In Figure 2, "Outwear" is a non-reaf node and "SKU00" which is a leaf node, and at the same time a descendant of "Clothes". A number called *level* can be assigned to each node in the product taxonomy. The level of the root node is zero, and the level of other node is one plus the level of its parent. Please note that a higher-level product class has a smaller level number. The product taxonomy of Figure 2 has four levels, referred to as level 0 (for root), 1,2 and 3. Before recommending product, product class level is decided. Product class level, level 2 in Figure 2, is an analyzing unit in this study.

3.2 Procedure of Recommender System

The recommender system uses decision tree technique to select target customer who has a higher possibility of purchasing recommended product. To recommend product efficiently, each customer's preference about products is used. Figure 3 shows an overall product recommendation process.


Step 1 : Data Preparation
Step 2 : Determinatiion of Target Customers
Step 3 : Mining Association Rules
Step 4 : Creation of Customer Preference Moddel
Step 5 : Product Recommedation

<Figure 3> The Procedure of Product Recommendation

3.2.1 Data Preparation

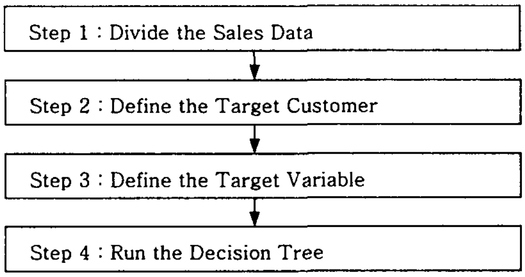
First, data collection is needed to perform product recommendation. Data is prepared from customer database, product database, and sales database. Customer data is composed of demographic data like age, sex, academic career, marriage, and job, and psychology data like life style and personality. Product data is composed of product ID, product name, price, brand, and manufacturer. Product data combined to sales data is used to acquire sales frequency and brand preference. Sales data is composed of customer ID, transaction ID, purchased date, and purchased product. Customer behavior and customer value can be detected by purchasing history from the sales data. These data will be needed to perform association rule mining and decision tree technique. From the sales data (customer ID, transaction ID, purchase data, and purchase product), product taxonomy is constructed.

3.2.2 Determination of Target Customers

Our recommender system selects not all

customers but customers with high a purchase possibility, target customer. A target customer is defined as the customer who purchases product more than P and less than P+I from the different classes till now.

To select target customer, this study uses decision tree technique. Decision tree technique is a powerful and popular tool for classification and prediction. On this study, Decision tree is used to classify the customer into target customers and the other customers. Figure 4 shows the procedure of decision tree to decide the target customer.


Step 1 : Divide the Sales Data
Step 2 : Define the Target Customer
Step 3 : Define the Target Variable
Step 4 : Run the Decision Tree

<Figure 4> The Procedure of Select Target Customer

3.2.3 Mining Association Rules

Product database is used to enable the user to find interesting patterns and trends in the data. To mine association rules, product taxonomy is needed first. On this study, product class unit (level 2 in Figure 2) is selected as an association rule unit.

Second, after selecting association rule mining unit, mining association rules is performed at a product class unit. Purchased product class set of customer 'm' is defined as Purset_m. Product class set associated with purchasing product class Purset_m is defined as AssoSet_m. Conf(s) is a confidence of association rule of product class s. If there are many rules that have lots of results about product class 's' and customer m, the rule with the most high confidence is selected. Figure 5(a) shows example of discovery of association rules and Figure 5(b) is an example of AssoSet_m about each customer m.

Product Association Rule

A=>C(0.8), A=>D(0.6), B=>E(0.8), B=>F(0.3),
C=>A(0.7), C=>F(0.4)

(a) Product Association Rules

CID	$PurSet_m$	$AssoSet_m$
101	A,E	C(0.8), D(0.6)
103	B	E(0.8), F(0.3)
112	B,C,E	A(0.7), F(0.4)
117	B,D	E(0.8), F(0.3)

(b) The Example of AssoSet_m

<Figure 5> Association Rules and Association Rule Set

Third, product affinity matrix is made based on product association rules of each customer. Table 1 shows an example of product affinity matrix.

3.2.4 Creation of Customer Preference Model

A customer preference model measures customer's preference toward products through the analysis of purchase data. Customer preference

<Table 1> Product Affinity Matrix

	A	B	C	D	E	F
A	1		0.8	0.6		
B		1			0.8	0.3
C	0.7		1			0.4
D				1		
E					1	
F						1

matrix is made from the purchase data, because customer preference is based on the number of transactions of product purchase of each customer. Table 2 shows an example of customer preference matrix.

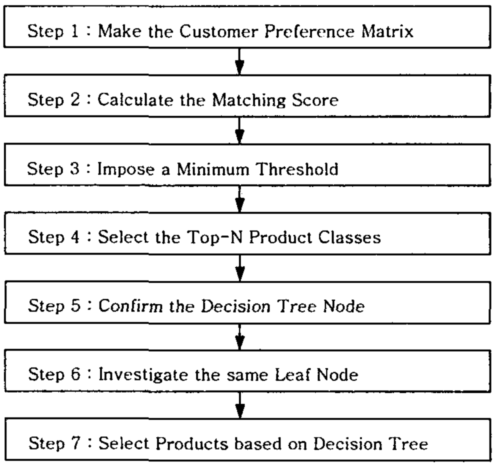
<Table 2> Customer Preference Matrix

CID	A	B	C	D	E	F
101	3	0	1	1	5	1
103	4	1	2	2	0	0
112	1	7	0	2	0	5
117	6	0	4	3	3	0

3.2.5 Product Recommendation

For the specific product recommendation, the procedure of product recommendation is shown in Figure 6.

Matching score is calculated based on product affinity matrix and customer preference matrix. Matching score is used to select recommending products. This study uses cosine coefficient as a matching score, $S_{ij} = \frac{P_i \cdot A_j}{\|P_i\| \cdot \|A_j\|}$, where, P_i is a row vector of the M*N customer preference matrix P, and A_j is a column vector of the N*N product affinity matrix.


Step 1 : Make the Customer Preference Matrix

Step 2 : Calculate the Matching Score

Step 3 : Impose a Minimum Threshold

Step 4 : Select the Top-N Product Classes

Step 5 : Confirm the Decision Tree Node

Step 6 : Investigate the same Leaf Node

Step 7 : Select Products based on Decision Tree

<Figure 6> The Procedure of Product Recommendation

Matching score obtained from the customer preference matrix, Table 2, and product affinity matrix, Table 1 is shown at Table 3. Matching score reflects the *similarity* between customer preference and product affinity. As an example, 0.51 implies the similarity measure about customer 101 and product A.

<Table 3> Matching Score Matrix

CID	A	B	C	D	E	F
101	0.51	0	0.44	0.26	0.65	0.31
103	0.91	0.2	0.83	0.73	0.12	0.22
112	0.09	0.78	0.07	0.24	0.48	0.79
117	0.91	0	0.85	0.69	0.29	0.2

A minimum threshold *t* is imposed on the matching score. To select top-N recommendation product classes, we have to give a heuristic limitation. Product classes with the matching score over a given threshold value is recommended to target customer. This results in Top-N product classes. For example, from the matching score table, if we give minimum threshold 0.5 to CID101, the top-N product classes is class A and E.

Product class unit is used from step 1 to step 4 of Figure 6, but there are many products in a same class. For recommending specific products among product class, we consider decision tree node led by step 2. Figure 7 shows the decision tree made at step 2.

If CID 103 is employee and the age is under 43, we investigate the customer in the same leaf node, represented by gray color at Figure 7. From that node, we consider customers who purchased product class A and E. In product class A and E, the most frequently purchased products or latest products are selected. The selected products are recommended to the target customer.

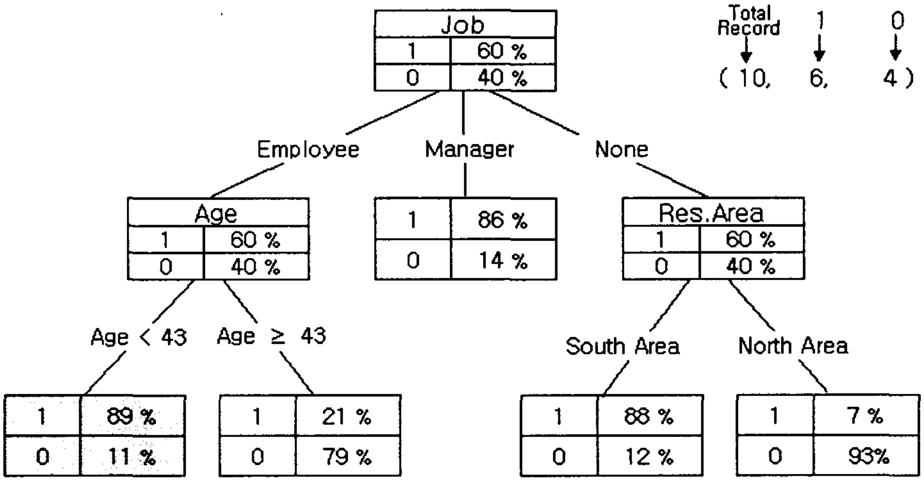
4. Experimental Evaluation

4.1 Experimental Step

4.1.1 Data Preparation

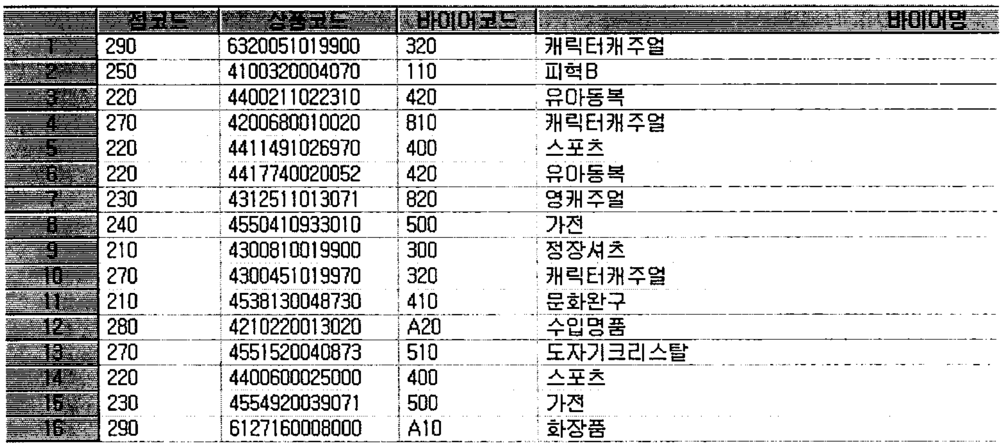
This study uses product data, sales data, and customer data of Korean *H* department store. The store has about 50,000 customers and over 190,000 products. But the number of customers who purchased only women's products is 1883 and the number of women's products is 1051.

Table 4 shows some part of the slightly modified data table of H department store.



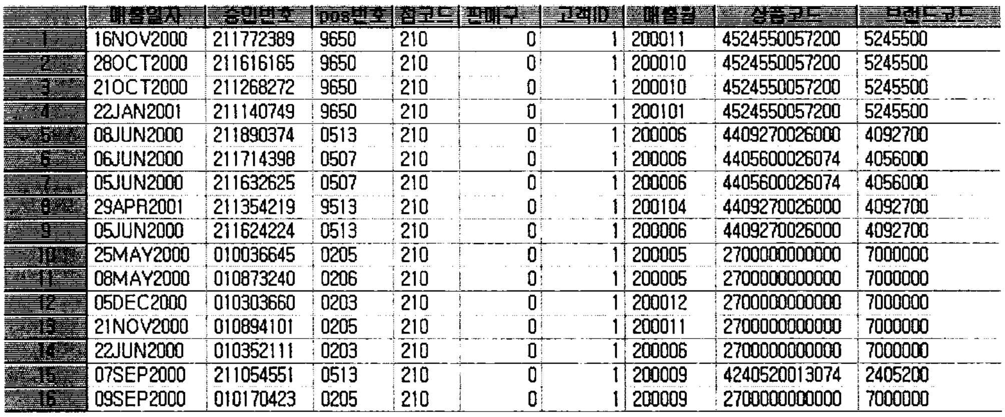
<Figure 7> A Decision Tree

<Table 4> Data Table of H department store



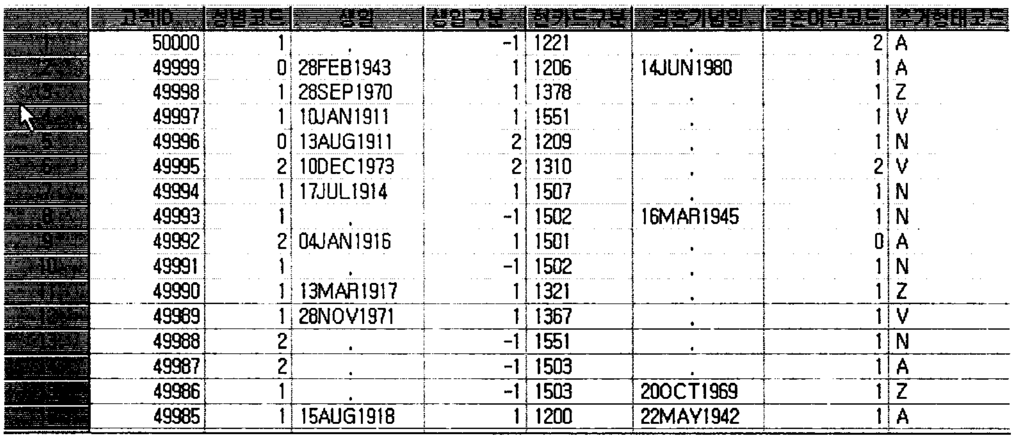
	점코드	상품코드	바이어코드	바이어명
1	290	6320051019900	320	캐릭터캐주얼
2	250	4100320004070	110	피혁B
3	220	4400211022310	420	유아동복
4	270	4200680010020	810	캐릭터캐주얼
5	220	4411491026970	400	스포츠
6	220	4417740020052	420	유아동복
7	230	4312511013071	820	영캐주얼
8	240	4550410933010	500	가전
9	210	4300810019900	300	정장셔츠
10	270	4300451019970	320	캐릭터캐주얼
11	210	4538130048730	410	문화완구
12	280	4210220013020	A20	수입명품
13	270	4551520040873	510	도자기크리스탈
14	220	4400600025000	400	스포츠
15	230	4554920039071	500	가전
16	290	6127160008000	A10	화장품

(a) Product Data Table



	매출일자	승인번호	pos번호	점코드	판매구	고객ID	매출월	상품코드	브랜드코드
1	16NOV2000	211772389	9650	210	0	1	200011	4524550057200	5245500
2	28OCT2000	211616165	9650	210	0	1	200010	4524550057200	5245500
3	21OCT2000	211268272	9650	210	0	1	200010	4524550057200	5245500
4	22JAN2001	211140749	9650	210	0	1	200101	4524550057200	5245500
5	08JUN2000	211890374	0513	210	0	1	200006	4409270026000	4092700
6	06JUN2000	211714398	0507	210	0	1	200006	4405600026074	4056000
7	05JUN2000	211632625	0507	210	0	1	200006	4405600026074	4056000
8	29APR2001	211354219	9513	210	0	1	200104	4409270026000	4092700
9	05JUN2000	211624224	0513	210	0	1	200006	4409270026000	4092700
10	25MAY2000	010036645	0205	210	0	1	200005	2700000000000	7000000
11	08MAY2000	010873240	0206	210	0	1	200005	2700000000000	7000000
12	05DEC2000	010303660	0203	210	0	1	200012	2700000000000	7000000
13	21NOV2000	010894101	0205	210	0	1	200011	2700000000000	7000000
14	22JUN2000	010352111	0203	210	0	1	200006	2700000000000	7000000
15	07SEP2000	211054551	0513	210	0	1	200009	4240520013074	2405200
16	09SEP2000	010170423	0205	210	0	1	200009	2700000000000	7000000

(b) Sales Data Table



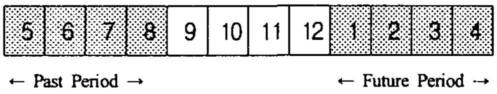
	고객ID	성별코드	생일	생일구분	현카드구분	결혼기념일	결혼여부코드	주거형태코드
1	50000	1	.	-1	1221	.	2	A
2	49999	0	28FEB1943	1	1206	14JUN1980	1	A
3	49998	1	28SEP1970	1	1378	.	1	Z
4	49997	1	10JAN1911	1	1551	.	1	V
5	49996	0	13AUG1911	2	1209	.	1	N
6	49995	2	10DEC1973	2	1310	.	2	V
7	49994	1	17JUL1914	1	1507	.	1	N
8	49993	1	.	-1	1502	16MAR1945	1	N
9	49992	2	04JAN1916	1	1501	.	0	A
10	49991	1	.	-1	1502	.	1	N
11	49990	1	13MAR1917	1	1321	.	1	Z
12	49989	1	28NOV1971	1	1367	.	1	V
13	49988	2	.	-1	1551	.	1	N
14	49987	2	.	-1	1503	.	1	A
15	49986	1	.	-1	1503	20OCT1969	1	Z
16	49985	1	15AUG1918	1	1200	22MAY1942	1	A

(c) Customer Data Table

From these tables, this study leads product taxonomy and then analyzes association rules based on product class unit.

4.1.2 Determination of Target Customer

First, sales data are divided. The period of *H department store*'s sales data is one year (from May 2000 to April 2001). So, we divide sales data as past period (from May 2000 to August 2000) and future period (from Jan 2001 to April 2001). Figure 8 shows the period division.

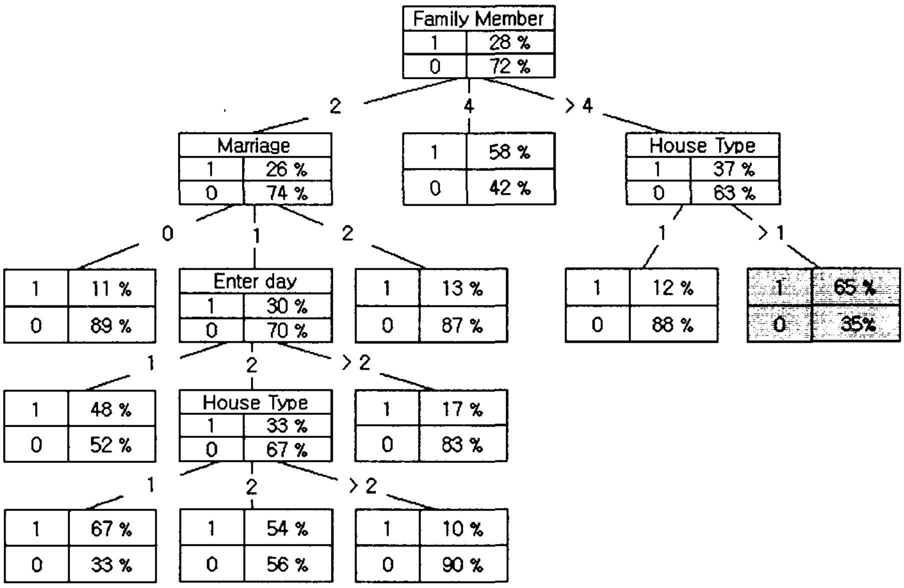


<Figure 8> Division of Period

Second, the target customer is determined. As mentioned before, we define the target customer who purchased women's products more than one during the past period and purchase other product classes during the future period.

Third, to decide the target customer, the target variable is setup to be "1" to the customers who purchase women's products more than one from the different classes during the past period.

Fourth, the decision tree is made and the best-classified decision tree is chosen. We made a decision tree based on *H department store* customers who purchased not only women's product but also all the products in the department store during the definition period. The reason that made a decision tree based on the customers who purchase all the products is to prevent the



<Figure 9> Result of Decision Tree

overfitting. From this decision tree, we use the customer field related with the purchase of only women's product instead of precious customer field.

We made decision tree from 1% of department store customers (500 customers). 500 customers" transaction data is 17,807. The input variables of decision tree are sex, marriage, house type, hobby, auto-payment, job, family number, recent visit day, and the day of joining the department store card. Figure 9 shows the result of decision tree.

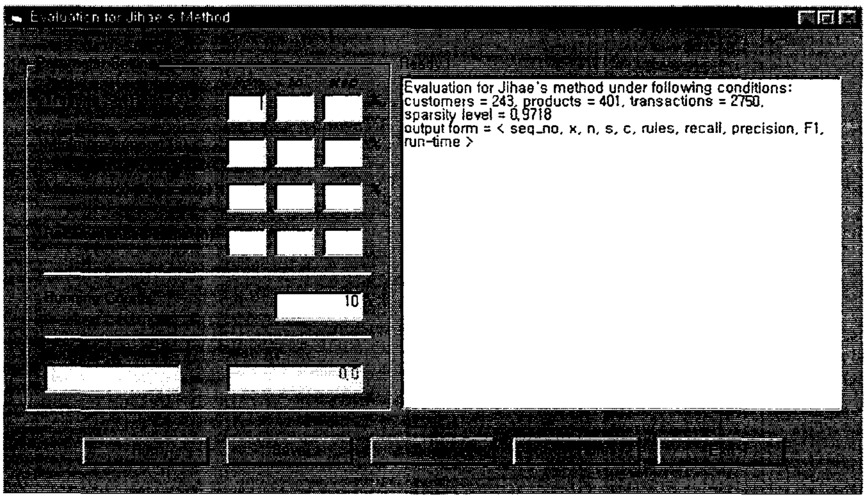
From the decision tree, we can explain the characteristics of target customers, e.g. the customer of right bottom node that has more than 4 family members and the house type is apartment.

4.1.3. Association Rule Mining, Customer Preference Model, and Product Recommendation

On the association rule mining step, we could know the probability which products are purchased together, and we also know the association between two or more items. Customer preference model measures customer's preference toward products. Finally, we could recommend products based on association rule mining and customer preference table. To do this experiment more efficiently, we made a system using Visual Basic. Figure 10 shows a user interface screen of the system. On this screen, there are spaces to give value of training ratio, minimum support, minimum confidence and recommended items.

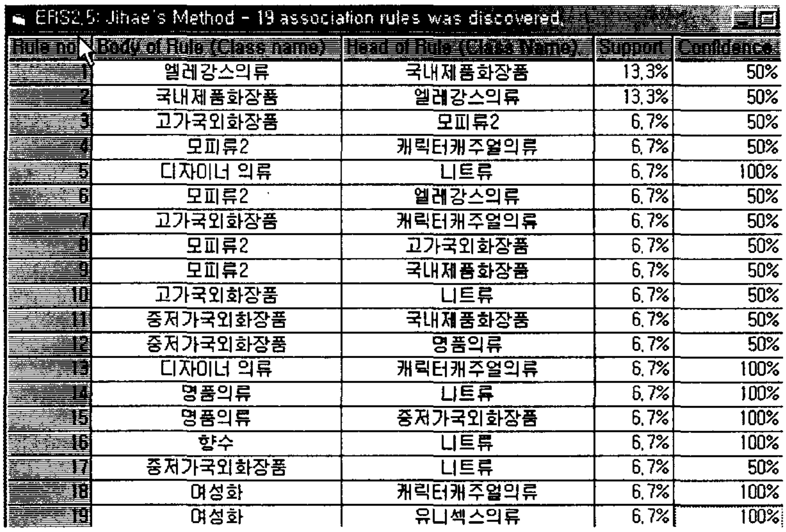
Figure 11 shows an association rule mining screen of products, *e.g.* if a customer purchased a perfume, he/she would purchase knit cloth together, because the confidence of this rule is 100%.

Figure 12 shows an example screen of personalized recommendation. There are products which customer already purchased is in the upper cell. Recommended products are in the middle cell.



```
Evaluation for Jihae's method under following conditions:
customers = 243, products = 401, transactions = 2750,
sparsity level = 0.9718
output form = < seq_no, x, n, s, c, rules, recall, precision, F1,
run-time >
```

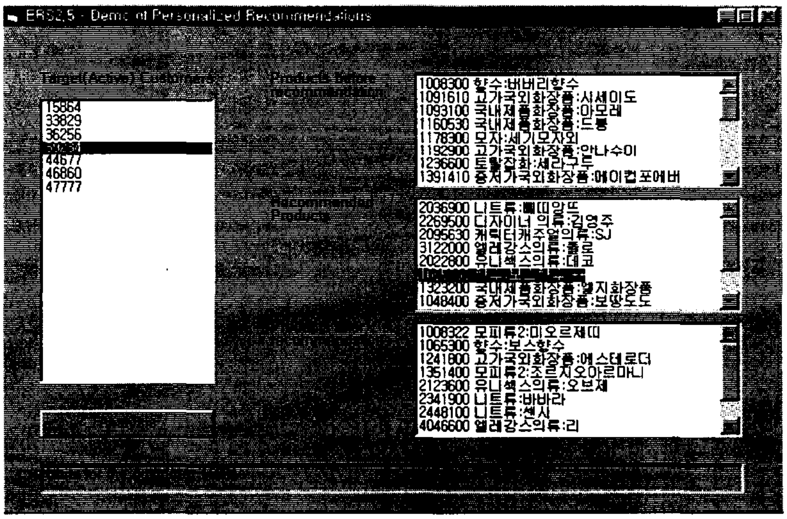
<Figure 10> Recommender System



ERS2.5: Jihae's Method - 19 association rules was discovered.

Rule no	Body of Rule (Class name)	Head of Rule (Class Name)	Support	Confidence
1	엘레강스의류	국내제품화장품	13,3%	50%
2	국내제품화장품	엘레강스의류	13,3%	50%
3	고가국외화장품	모피류2	6,7%	50%
4	모피류2	캐릭터캐주얼의류	6,7%	50%
5	디자이너 의류	니트류	6,7%	100%
6	모피류2	엘레강스의류	6,7%	50%
7	고가국외화장품	캐릭터캐주얼의류	6,7%	50%
8	모피류2	고가국외화장품	6,7%	50%
9	모피류2	국내제품화장품	6,7%	50%
10	고가국외화장품	니트류	6,7%	50%
11	중저가국외화장품	국내제품화장품	6,7%	50%
12	중저가국외화장품	명품의류	6,7%	50%
13	디자이너 의류	캐릭터캐주얼의류	6,7%	100%
14	명품의류	니트류	6,7%	100%
15	명품의류	중저가국외화장품	6,7%	100%
16	향수	니트류	6,7%	100%
17	중저가국외화장품	니트류	6,7%	50%
18	여성화	캐릭터캐주얼의류	6,7%	100%
19	여성화	유니섹스의류	6,7%	100%

<Figure 11> Association Rule Mining



ERS2.5 - Demo of Personalized Recommendations

15864
33829
36256
44677
46860
47777

1008300 향수:버버리향수
1091610 고가국외화장품:시세이도
1093100 국내제품화장품:아모레
1160530 국내제품화장품:드봉
1178300 모자:세기모자외
1192900 고가국외화장품:안나수이
1236600 토탈잡화:세라구두
1391410 중저가국외화장품:메이컵포에버

2036900 니트류:빼떼앙뜨
2269500 디자이너 의류:김영주
2095630 캐릭터캐주얼의류:SJ
3122000 엘레강스의류:홀로
2022800 유니섹스의류:데코
1323200 국내제품화장품:쌀지화장품
1048400 중저가국외화장품:보땅도도

1008322 모피류2:미오르제띠
1065300 향수:보스향수
1241800 고가국외화장품:에스테로더
1351400 모피류2:조르지오아르마니
2123600 유니섹스의류:오브제
2341900 니트류:바바라
2448100 니트류:센시
4046600 엘레강스의류:리

<Figure 12> Personalized Recommendation

And in the bottom cell, it shows purchased products from the recommended products.

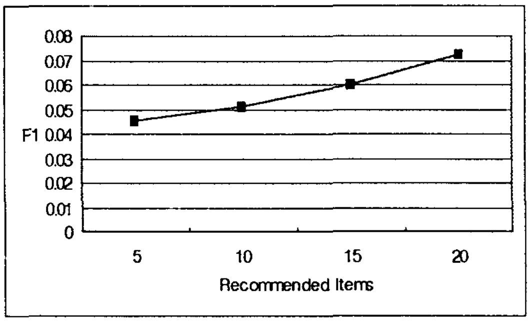
4.2 Evaluation Metrics

With the training set and the test set, our methodology works on the training set first, and then it generates a set of recommended products, called recommendation set, for a given customer. To evaluate the quality of the recommendation set, *recall* and *precision* have been widely used in the recommender system community (Sarwar, et al., 2000). Recall is defined as the ratio of the number of products in both test set and recommendation set to the number of products in test set. Precision is defined as the ratio of the number of products in both test set and recommendation set to the number of products in recommendation set. Recall means how many of all the products in the actual customer purchase list are recommended correctly whereas precision means how many of the recommended products belong to actual customer purchase list. These measures are simple to compute and intuitively appealing, but they are often in conflict since increasing the size of recommendation set tends to increase recall but at the same time decrease precision (Sarwar, et al., 2000). Hence, a widely used combination metric called *F1 metric* (Rijsbergen, 1979) that gives equal weight to both recall and precision is employed for our evaluation, and computed as follows:

$$F1 = \frac{2 \times recall \times precision}{recall + precision} .$$

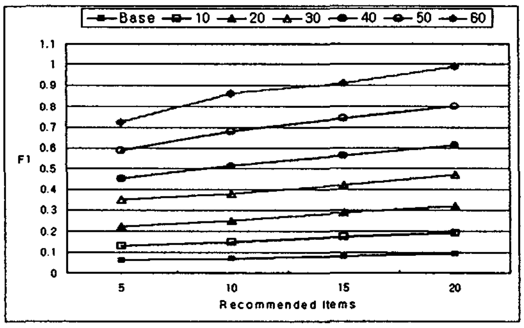
4.3 Results

Figure 13 shows the experimental results based on the number of recommended items. When we recommend the products from 5 to 20 to the customer, we could know how does F1 metric change. The more products are recommended, the higher F1 metric value becomes. It's somewhat a natural result, because if a person recommended more products, they will have more opportunity to select among the recommended products.



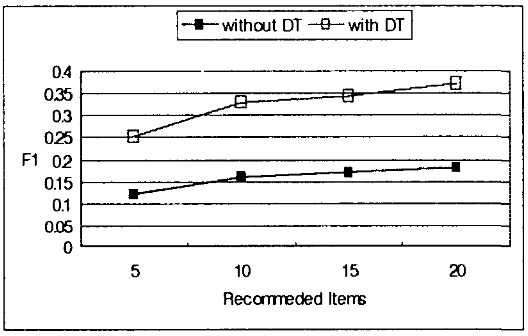
<Figure 13> The Change of Recommended Items

Next experiment is performed on the target customer data led from decision tree. Target customer is a customer who has a higher purchase possibility of recommended product. Figure 14 shows F1 metric based on changing recommended product counts and customers. We divide customers who purchased products more than 10 as categories; 10, 20, ... , 60. So, the total customer group is 7. The result shows that customers who purchased more products are likely to buy recommended products.



<Figure 14> The Change Recommended Items and Target Customers

The comparison between target customers and base customers results in Figure 15. From this, we could know the F1 metric of each target customer is evidently higher than base customers. Figure 15 shows only the results of customers who purchased products more than 60, but other results shows the same result. This implies that product recommendation based on the past product history and demographic information results a higher value of F1 metric. This shows an evidence that target marketing is more useful than mass marketing.



<Figure 15> The Comparison of Base Customers and Target Customers

5. Conclusion

This study suggests a hybrid recommendation procedure and experiments with sales data, customer data, and product data of H Department store in Korea. To determine target customer we use decision tree algorithm, and to select proper products, we use collaborative filtering, content-based filtering and association rule mining. Also, this study applied the procedure to the real department store and evaluates the procedure. From the result, we prove that our procedure is better than existing collaborative filtering based recommender system.

But, this study uses only off-line customer data, product data and sales data. For the better and more product recommendation, using web-log data will result in good performance. Furthermore, the implementation of recommender system is necessary to use in the real field.

References

Adomavicius, G. & Tuzhilin, A., "Expert-driven validation of rule-based user models in personalization applications", *Data Mining and Knowledge Discovery*, 5(2001), 33-58.

Agrawal, R., Imielinski, T., & Swami, A., "Mining assocaition between sets of items in massive database", In *International Proceedings of the ACM-SIGMOD International Conference On Management of Data,* (1993), 207-216.

Basu, C., Hirsh, H., & Cohen, W., "Recommendation as classification: Using social and content-based information in recommendation", *In Proceedings of the Fifteenth National Conference on Artificial Intelligence*, (1998), 714-720.

Balabanovic, M., & Shoham, Y., "Fab: Content-Based, Collaborative Recommendation", *Communications of the ACM,* 40 (1997), 66-70.

Berry, J. A., & Linoff, G., *Mastering Data Mining,* The Art and Science of Customer Relationship Management, New York: Wiley, 2000.

Berson, A., Smith, K., & Thearing, K., *Building Data Mining Applications for CRM,* New York: McGraw-Hill, 2000.

Billsus, D., & Pazzani, M. J., "Learning collaborative information filters", *In Proceedings of the Fifteenth International Conference on Machine Learning*, (1998), 46-54.

Changchien, S. W. & Lu, T., "Mining Association rules procedure to support on-line recommendation by customers and products fragmentation", *Expert Systems with Applications*, 20 (2001), 325-335.

Cho, Y.H., Kim, J.K., and Kim, S.H., "A Personalized Recommender System based on Web Usage Mining and Decision Tree Induction", *Expert Systems With Applications*, 23 (2002), 329-342.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., & Sartin, M., "Combining content-based and collaborative filters in an online newspaper", *In ACM SIGIR"99 Workshop on Recommender Systems*, Berkeley, CA, 1999.

Han. J. & Fu, Y., "Mining multiple-level association rules in large databases", *IEEE Transactions on knowledge and data engineering*, 11 (1999), 798-804.

Han, J. & Kamber, M., Data mining: *concepts and techniques*, Morgan Kaufmann Publishers (2001).

Kim, S. H., Shin, S. W., & Kim J.H., "Personalized recommendations for retailing in internet commerce: a multistrategy filtering approach" In *International Conference on Electronic Commerce 2000* (2000), 103-111.

Lawrence, R. D., Almasi, G. S., Kotlyar, V., Viveros, M. S., & Duri, S.S., "Personalization of supermarket product recommendations", *Data Mining and Knowledge Discovery*, 5 (2001), 11-32.

Lin, W., Alvarez, S. A., & Ruiz, C., "Efficient Adaptive-Support Association Rule Mining for Recommender Systems", *Data Mining and Knowledge Discovery*, 6 (2002), 83-105.

Mobasher, B., Cooley, R., & Srivastava, J., "Automatic personalization based on web usage mining" *Communications of the ACM*, 43 (2000), 142-151.

Rijsbergen, C.J. *Information Retrieval, 2nd edn.* London: Butter-worth, 1979.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J., "Analysis of recommendation algorithms for e-commerce", In *Proceedings of ACM E-Commerce 2000 Conference* (2000), 158-167.

Sarwar, B.M., Konstan, J.A., Borchers, A., Herlocker, J., Miller, B. & Riedl, J., "Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System" *In Proceedings of the CSCW-98*, (1998), 345-354.

Schafer, J. B., Konstan, J. A., & Riedl, J. "E-commerce recommendation applications", *Data Mining and Knowledge Discovery*, 5 (2001), 115-153.

Yuan, S. & Chang, W., "Mixed-initiative synthesized learning approach for web-based CRM", *Expert Systems with Applications*, 20 (2001), 187-200.

요약

협업 필터링 기법을 활용한 개인화된 상품 추천 방법론 개발에 관한 연구

김재경*
서지혜*
안도현*
조윤호**

본 연구에서는 기존 협업 필터링의 문제점을 해결할 수 있는 효율적인 상품추천 방법론을 제시하고자 한다. 연구에서 제시하는 상품추천 방법론은 기존 협업 필터링 알고리즘의 데이터 희박성 문제 및 동의어 문제를 극복하기 위하여 판매 데이터로 구성된 제품 계층도(Product Taxonomy)를 이용하며, 이 계층도를 기반으로 한 연관 규칙(association rule)과 의사결정 나무를 사용한다. 본 연구에서는 제시한 방법론을 단계별로 설명하였을 뿐만 아니라, 실제 H 백화점 데이터를 이용하여 적용하였다. 다양한 경우에 대하여 실험을 한 결과, 기존의 협업 필터링 알고리즘이 갖고있는 문제점을 상당히 해결하였음을 제시하였다. 이 연구에서 제시한 상품 추천 방법론은 현재 기업이 직면한 경쟁환경 하에서 고객이 과연 누구이며, 고객이 진정 무엇을 원하고 있는지를 파악하는데 도움을 줄 것이며, 고객관계관리(CRM)를 효율적으로 구현하는 방법론으로 사용될 것으로 기대된다.

* 경희대학교 경영학부
** 동양공업전문대학 인터넷정보과