대한산업공학회/한국경영과학회 2002 춘계공동학술대회
한국과학기술원(KAIST) 2002년 5월 3일-4일

A Personalized Recommender based on Collaborative Filtering and Association Rule Mining

Jae Kyeong Kim [1*], Ji Hae Suh[2], Yoon Ho Cho [3], and Do Hyun Ahn [4]
1,2,4 School of Business Administration, KyungHee University#1, Hoeki-Dong, Dongdaemoon, Seoul, 130-701, South Korea

3 Department of Internet Information Systems, Dongyang Technical College62-160 Kochuk, Kuro, Seoul 152-714, Korea

Abstract

A recommendation system tracks past action of a group of users to make a recommendation to individual members of the group. The computer-mediated marketing and commerce have grown rapidly nowadays so the concerns about various recommendation procedures are increasing.

We introduce a recommendation methodology by which Korean department store suggests products and services to their customers. The suggested methodology is based on decision tree, product taxonomy, and association rule mining. Decision tree is to select target customers, who have high purchase possibility of recommended products. Product taxonomy and association rule mining are used to select proper products.

The validity of our recommendation methodology is discussed with the analysis of a real Korean department store.

1. Introduction

E-commerce has been growing rapidly keeping the pace with the web. However, its rapid growth rather 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) has been stressed from research as well as from practical affairs (Sarwar, et al., 2000; Mobasher, et 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. But, there are common problems on existing recommender systems. First 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). Second 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) Third is synonymy. In real scenario, different product names can refer to similar objects.

So, in this study, we propose a new methodology for personalized recommendations in the department store.

2. Backgrounds

2.1 Recommender System Using Collaborative Filtering

Collaborative filtering (CF) is the most successful recommender system technology, and it is used in many of the successful recommender systems on the web. CF systems recommend products to the target

customer based on the opinions of other like-minded customers (neighbor). These systems employ statistical techniques to find a set of customers known as neighbor who have a similar history of agreeing with the target user. When neighbor is formed, these systems use several algorithms to produce recommendations. On this study, we divide the entire process of CF-based recommendation procedure into sub-tasks namely, representation, neighbor formation, and recommendation generation.

2.2 Recommender System 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.

2.3 Combining Collaborative Filtering and Association Rule Mining

Recommender system which combines collaborative filtering and association rules is investigated. This study divides the customer group as similar purchase history groups to recommend customer preference product. Association rule is used to know product class association.

2.4 Combining Content-based and Collaborative filtering

Several hybrid approaches that combined content-based filtering and collaborative filtering have been proposed currently.

3. Methodology

3.1 Recommender Procedure

3.1.1. Definition of 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 prefer product to the customer. On E-Commerce environment, this study defines product recommendation as follows.

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. P and N is a figure more than 1, and I is a figure more than 0.

For instance, product recommendation can be presented Figure 1 in case of p=1, I=4, N=2.

Figure 1. Example of Product Recommendation

3.1.2 Product Taxonomy

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 *All* denotes the most general product class. For example, Figure 2 shows an example of such a taxonomy for a online women total product stores, where "Clothes", "Cosmetics", "Accessories", and "Leather" are classified into "Women's Product", and so on.

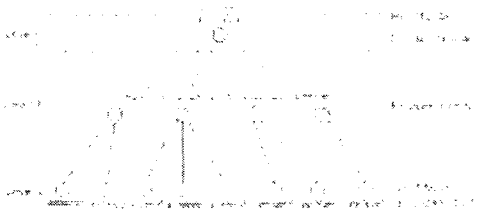


Figure 2. Example of Product Taxonomy

Product hierarchies play an important role in the knowledge discovery process since they represented department store. Mining association rules on different levels of the product taxonomy discovers more specific and concrete knowledge rather than at single level(Han & Fu, 1995). Furthermore, it is useful to have a taxonomy of the products being considered for data mining analysis because choosing the higher levels of the product taxonomy may lead to improve the results of the analysis (Berry & Linoff, 1997).

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, "Clothes" is a non-leaf node and a parent of "4229800" which is a leaf node, and at the same time a descendant of "Women's Product",

a root node. 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 three levels, referred to as level 0 (for root), 1 and 2.

Before recommending product, product class level is decided by product taxonomy. Product class level is an analyzing unit when product association is performed.

3.2 Overall Procedure of Recommender System

3.2.1 The Outline of Recommender System

This recommender system uses decision tree technique to select target customer who has a high possibility of purchase. Recommendation product is acquired by association rule mining. Also, to recommend product efficiently information of each customer's preference is used.

3.2.2 The Process of Product Recommendation

Figure 3 shows general product recommendation process. Product recommendation process will be explained as follows.

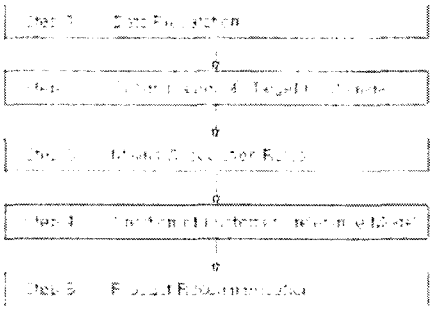


Figure 3. The Procedure of Product Recommendation

3.3 Specification of Procedure

3.3.1 Data Preparation

First, collection of data is needed to perform product recommendation. Data source is prepared by customer data, product data, and sales data from the database.

(1) Customer Data

Customer data has demographic data like age, sex, academic career, marriage, and job, and psychology data like life style, personality. From this, customization service can provide to the customer.

(2) Product Data

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.

(3) Sales Data

Sales data is composed of customer ID, transaction ID, purchased date, and purchased product. Customer behavior and customer value can be detected by purchase 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.

3.3.2 Determination of Target Customers

Product recommender system selects not all customers but high purchase possibility customer, target customer. Target customer is 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. A decision tree is a representation of a decision procedure for determining a class label to associate with a given example. At each internal node of the tree, there is a test (question), and a branch corresponding to each of the possible outcomes of the test. At each leaf node, there is a class label (answer). Traversing a path from the root to a leaf is much like playing a game of twenty questions.

Decision trees have a great many uses, particularly for solving problems that can be cast in terms of producing a single answer in the form of a class name. For example, one can build a decision tree that could be used to answer a question such as `Does this patient have hepatitis?' The answer may be as simple as `yes' or `no'. Based on answers to the questions at the decision nodes, one can find the appropriate leaf and the answer it contains.

Decision trees are constructed from examples that are already labeled. For example, if one has established for a variety of patients with varying attributes which of them do and do not have hepatitis, then these examples can guide the tree construction process.

On this study, Decision tree used to classify the customer. What attributes does the target customer have? To know the attribute of customer, procedure goes through as follows. Figure 4 will show the procedure of decision tree for the target customer.

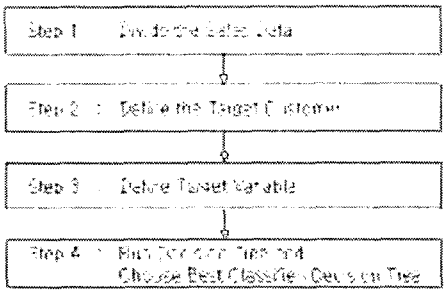


Figure 4. The Procedure of Select Target Customer

3.3.3 Mining Association Rules

One of the reasons behind maintaining any database is to enable the user to find interesting patterns and trends in the data.

To mine association rules, Figure 10 shows the procedure of association rule mining.

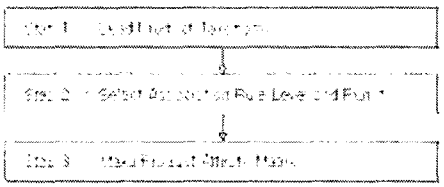


Figure 5. The Procedure of Association Rule Mining

First, product taxonomy is needed. The reason about necessity of taxonomy is to determine the association rule level. On this study, product class level 1 is selected as an association rule unit. Figure 6 shows association rule mining unit.

Figure 6. Association Rule Mining Unit

Second, after selecting association rule mining unit, mining association rule as an product class unit. Product class set that customer 'm' purchase is defined as $Purset_m$. Product class set that associated with purchasing product class is defined as $AssoSet_m$. Conf(s) is a confidence of association rule leading from product class s. If there are many rules that have lots of results about product class 's' and customer m, the most high confidence is selected. Figure 7 shows discovery of association rule. Figure 8 is an example of $AssoSet_m$ about each customer m.

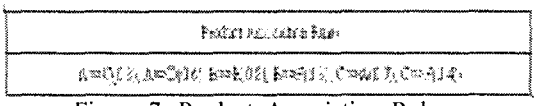


Figure 7. Product Association Rules

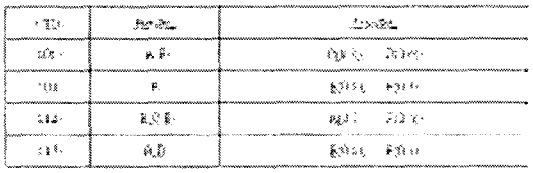


Figure 8. The Example of AssoSetm about Each Customer 'm'

Third, product affinity matrix is made based on product association rule of each customer. Figure 9 shows an example of product affinity matrix about product association rule.

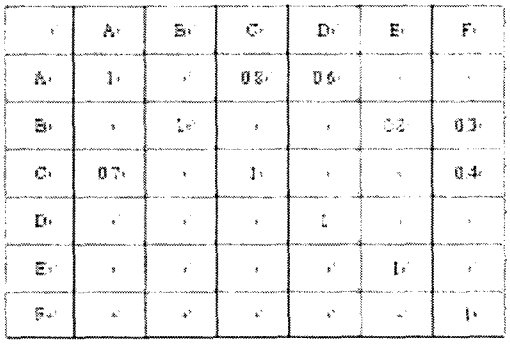


Figure 9. Product Affinity Matrix

3.3.4 Creation of Customer Preference Model

This study suggests a customer preference model which measures customer's preference toward products through analysis of store data. From the sales data and customer data, purchase data is led. From the purchase data, the purchase of the product-completion of a transaction is known and customer preference matrix is made. Figure 10 shows an example of customer preference matrix.

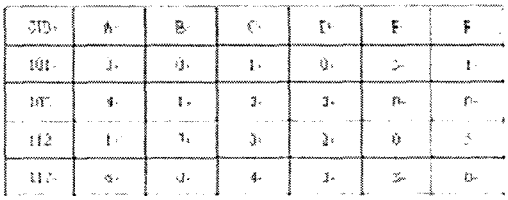


Figure 10. Customer Preference Matrix

3.3.5 Product Recommendation

For the specific product recommendation, this study goes as this procedure. Figure 11 shows the procedure of product recommendation.

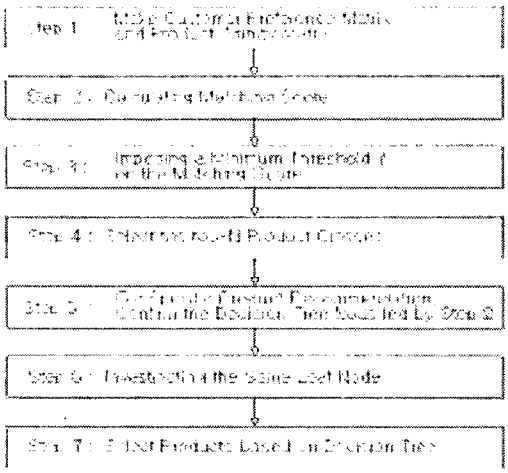
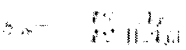


Figure 11. The Procedure of Product Recommendation

First, bring the customer preference matrix and product affinity matrix. Second, calculate a matching score. This study uses matching score for exact recommendation. Matching score is calculated based on product affinity matrix and customer preference matrix.

This study uses cosine coefficient as a matching score,



where

Pi is row vector of the M*N customer preference matrix P.

Ai is column vector of the N*N product affinity.

Figure 12 shows the customer preference matrix and product affinity matrix.

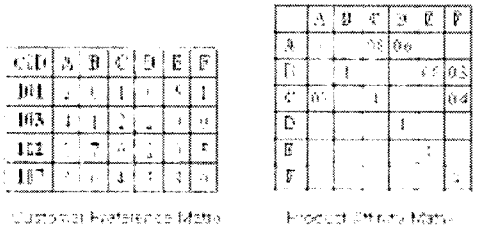


Figure 12.Customer Preference Matrix and Product Affinity Matrix

Matching score obtained from the customer preference matrix and product affinity matrix is shown at Figure 13. Matching score reflects the *similarity* between customer preference and product affinity. From this result, a minimum threshold t is imposed on the matching score. It means that, to select top-N classes, we have to give limitation. Giving minimum threshold is arbitrary based on result of matching score.

CID	A	B	C	D	E	F
101	0.51	0	0.44	0.30	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

Matching Score

$$0.51 = \frac{(5*1)+(1*0.7)}{\sqrt{5^2+0^2+1^2+0^2+5^2+1^2}+\sqrt{1^2+0.7^2}}$$

Figure 13. Matching Score Matrix and the Example of Matching Score

Fourth, top-N product classes are selected. 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. Fifth, we select product as a product class unit from step 1 to step 4 but there are many products in a same class.

For recommending specific product among product class consider decision tree node led by step 2. Sixth, decision tree node is investigated. Figure 14 shows the decision tree made by step 2.

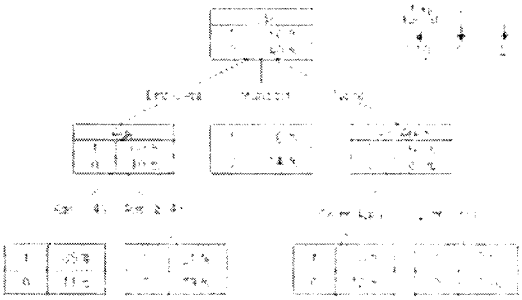


Figure 14. Decision Tree

If the attribute of CID 103 is employee and the age is under 43, investigate same leaf node. Represented by gray color at Figure 19. From that node, it is found the person who purchased product class A and E. Seventh, it is selected the most frequently purchased products or latest products in product class A and E. Then recommend the product to the target customer.

4. Experiments

4.1 Experimental Platform

In this chapter, a case application to the method proposed in chapter 3 is illustrated. This case is based on the real H department store. This store has 50,000 customers and 190,000 product items. This

study uses sales data, customer data and product data of *H department* store and the period is from May 2000 to April 2001. Before beginning the experiment, this data leads product taxonomy based on sales data. This study's product taxonomy has 4 levels. Level 0 is total women product, level 1 is consisted of women clothing, cosmetic, accessary, and leather. Level 2 is consisted of 26 products. Level 3 is consisted of 1051 products.

From this taxonomy, this study uses association rule mining as a class unit. As mentioned before we will focus two different experiments. On the first experiment, we evaluate the result based on the customer by differentiating total purchased products counts. On the second experiment, we evaluate the result comparing target customers and base customers.

To evaluate *Top-N* recommendation we use two metrics widely used in the information retrieval(IR) community namely recall and precision. However, we slightly modify the definition of recall and precision as our experiment is different from standard IR. We divide the products into two sets, the *test* set and *topN* set. Products that appear in both sets are members of the *hit set*. We now define recall and precision as the follows. Recall in the context of the recommender system is defined as the ratio of hit set size and to the test set size.

$$Recall = \frac{|test \cap topN|}{|test|}$$

Precision is defined as the ratio of hit set size to the top-N set size.

$$Precision = \frac{|test \cap topN|}{N}$$

These two measures are, however, often conflicting in nature. For instance, increasing the number N tends to increase recall but decrease precision. The fact that both are critical for the quality judgment leads us to use a combination of the two. In particular, we use the standard F1 metric (Yang et. al. 1999) that gives equal weight to them both and is computed as follows,

$$F1 = \frac{2 * Recall * Precision}{(Recall + Precision)}$$

We compute F1 for each individual customer and calculate the average value to use as our metric.

4.3 Experimental Steps

This study shows the experiment procedure in accordance with methodology steps.

4.3.1 Data Preparation

This study use product data, sales data, and customer data of *H department store*. *H department store* has 50,000 customer and 190,000 Products. But the person who purchased only women's product is 1883 and the women's product is 1051. Figure 15, 16, and 17 show the table of data for analysis.

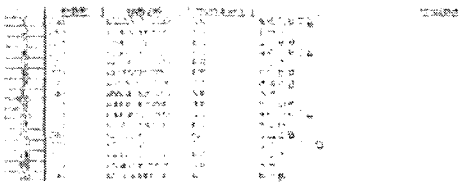


Figure 15. Product Data Table

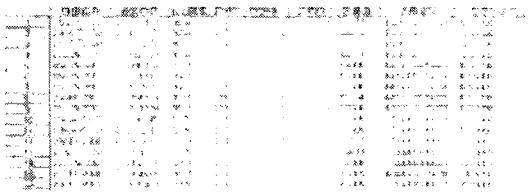


Figure 16. Sales Data Table

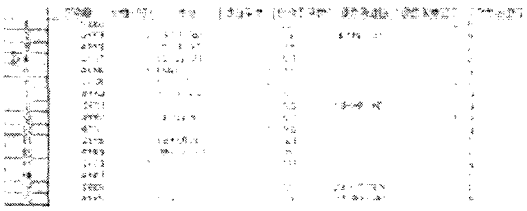


Figure 17. Customer Data Table

4.3.2 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 is from May 2000 to August 2000 and future period is from Jan. 2001 to April 2001. Figure 18 shows the period division

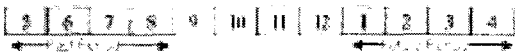


Figure 18. Division of Period

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 classes products during the future period. To lead the target customer we made the program using Visual Basic.

Third, the target variable is defined. On this study target variable is represented as "1". Number 1 is customer who purchases women's products more than one from the different classes during the past

period and future period. Figure 19 shows the variables of customer table.

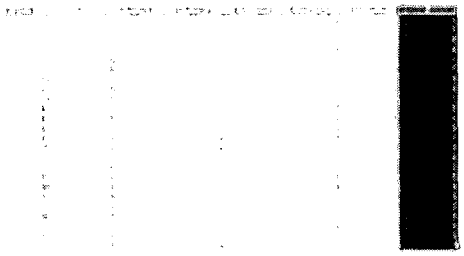


Figure 19. Variables of Target Customers

Column Y is the target variable of customer table. 1 is target variable and 0 is other cases.

Fourth, a decision tree is made and 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 product 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 overfitting. And then, from this decision tree we use the customer field who purchased only women's product instead of precious customer field. So, the customer with the same attribute value of all the product customer in the "1" node is our target customer. We also know the attribute of target customer based on decision tree nodes.

We made decision tree based on 1% of department store customer (500 customers). 500 customer's 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. And then we fix data as exact input data.

4.3.3. 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 program using Visual Basic. Figure 20 shows a user interface screen of this program. On this screen, There are spaces to give value of training ratio, minimum support, minimum confidence and recommended items.

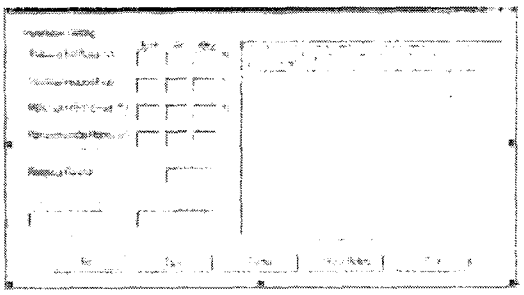


Figure 20. Recommender System

Figure 21 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. And in the bottom cell, it shows which products are purchased from the recommended products

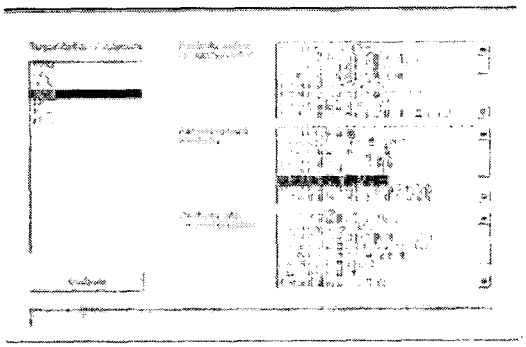


Figure 21. Personalized Recommendation

4.4 Results

From this experiment, we use target customer data led from decision tree. Target customer is a customer who has high purchase possibility of recommended product. Figure 22 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.

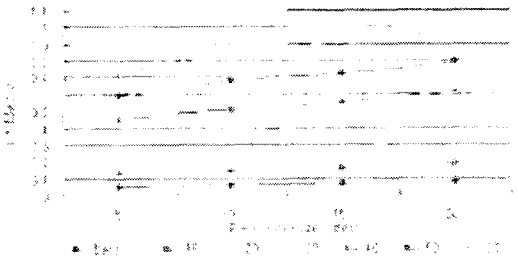


Figure 22. The Change of F1 Metric Value based on the Change of Recommended Items and Target Customers

5. Conclusions

This study uses hybrid algorithm, in other word, we use sales data, customer data, and product data from *H Department store* in Korea. And then to determine target customer we use decision tree algorithm. To select proper products, we use collaborative filtering, content-based filtering and association rule mining. From this we can perform an efficient recommendation. Also, this study applied to the real shopping mall and evaluates result based on it. From the result, we prove that our algorithm is better than existing recommender system.

But, this study uses only customer data, product data and sales data, but for the better and more accurate research using web-log data will be good.

Also, comparing with other algorithm will be an interesting research area. This study suggests a recommender procedure evaluates the methodology. However the implementation of recommender system is necessary to use in the real field.

References

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

Basu, C., Hirsh, H., & Cohen, W. (1998). Recommendation as classification: using social and content-based information in recommendation. In *Proceedings of the 1998 Workshop on Recommender Systems*, 11-15, AAAI Press.

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

Berry, J. A., & Linoff, G. (1997). *Data mining techniques: for marketing, sales, and customer support*, New York: Wiley.

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

Han. J. & Fu, Y. (1999). Mining multiple-level association rules in large databases. IEEE *Transactions on knowledge and data engineering*, 11 (5), 798-804.

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

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

Lawrence, R. D., Almasi, G. S., Kotlyar, V., Viveros, M. S., & Duri, S.S. (2001). Personalization of supermarket product recommendations. *Data Mining and Knowledge Discovery*, 5 (1-2), 11-32.

Lin, W., Alvarez, S. A., & Ruiz, C. (2000). Collaborative recommendation via adaptive association rule mining. In *WEBKDD'2000*, Boston, MA.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, 43 (8), 142-151.

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

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based Collaborative Filtering Recommendation Algorithm. In *Proceedings of The Tenth International World Wide Web Conference*, 285-295.

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, 5 (1-2), 115-153.

Popescul, A., Ungar, L. A., Penneck, D. M., & Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments, accepted to *UAI 2001 Conference*, Seatle, WA, August 2001.

Srikant, R. & Agrawal, R. (1995). Mining generalized association rules. In *Proceedings of the International Conference On Very Large Data Bases.*

Vucetic, S. & Obradovic, Z. (2000). A regression-based approach for scaling-up personalized recommender systems in e-commenrce. In *WEBKDD'2000*, Boston, MA.

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