An Extended Content-based Procedure to Solve a New Item Problem

Moon Kyoung Jang
School of Business Administration
Kyung Hee University
(moonk@khu.ac.kr)

Hyea Kyeong Kim
School of Business Administration
Kyung Hee University
(kimhk@khu.ac.kr)

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

Nowadays various new items are available, but limitation of searching effort makes it difficult for customers to search new items which they want to purchase. Therefore new item providers and customers need recommendation systems which recommend right items for right customers. In this research, we focus on the new item recommendation issue, and suggest preference boundary- based procedures which extend traditional content-based algorithm. We introduce the concept of preference boundary in a feature space to recommend new items. To find the preference boundary of a target customer, we suggest heuristic algorithms to find the centroid and the radius of preference boundary. To evaluate the performance of suggested procedures, we have conducted several experiments using real mobile transaction data and analyzed their results. Some discussions about our experimental results are also given with a further research area.

Received : November 2008 Revision : December 2008 Accepted : December 2008 Corresponding author : Jae Kyeong Kim

1. Introduction

The recommendation systems, one of the information techniques to recommend appropriate items for customers' preference, have been used in various fields. Customers can choose various items but, in other side, it is not easy find the items which they want to purchase among various items. Therefore item providers and customers need recommendation systems which recommend right items for right customers. Especially, as new items are frequently released nowadays, item providers and customers need the recommendation system which is specialized in recommending new items. Many ap-

* This work was supported by a grant from the Graduate Research Scholarship, Kyung Hee University Graudaute School, Seoul, Republic of Korea 2008.

proaches have been proposed for providing recommendations including collaborative filtering, content-based, demographic, knowledge-based and other technique, while two of the most prevalent approaches are collaborative filtering (CF) and content-based (CB) approach.

Traditional CF approaches have to rely on the transaction history of customers, so the currently released new items cannot be recommended, because these new items have no accessed records, or no rating from customers. This is ordinarily referred to as the 'New item ramp-up problem' (Avery and Zeckhauser, 1997; Jian et al., 2004). CB recommends items based on item source information, so it is hard to find the logical reason of serendipitous discovery (Balabanovic and Shoham, 1997; Good et al., 1999). Furthermore, if the customer has not enough purchase records or ratings, CB alone cannot be used for recommending new items. Consequently many researchers have suggested lots of hybrid approaches which combine CF approach, CB approach and other recommendation approaches to obtain better performance of recommending new items. Celma et al.(2005) have proposed the system using the FOAF and RSS vocabularies for recommending music to a customer. Cornelis et al.(2007) have proposed recommendation method using fuzzy logic techniques. Jian et al.(2004) have proposed recommendation method for new items based on indexing techniques. Kim et al.(2006) supposed the MOBICORS-music system for recommending new music in mobile web environment. Basically it follows the procedure of CF system, but it uses the data representation of CB system by using the content-based features of music. Based on this data representation, they argue that MOBICORS-music can solve the new item ramp-up problem and its performance is better than other CF systems. But their methods and experiences are performed for all items, not specified for the new item only.

In this research, we suggest a hybrid procedure for recommending new items based on the preference boundary in feature space. As a new item is expressed in the multi dimensional feature space, each customer is also represented as a cluster in this feature space because the customer is connected with the feature data of his/her purchased item. Thus for the recommendation of new items, the preference boundary of each customer is supposed to be the feature space comprised of the feature values of preferred items. New item recommendation procedures decide whether to recommend a new item to a target customer or not. The basic concept of our procedures is that if new items belong within the preference boundary, then it can be preferred by the target customer.

The preference boundary-based procedure is composed of two sub procedures, which are to determine the preference centroid, and to determine the range of preference boundary. The preference centroid is defined by the purchased items of the target customer. To determine the range of the preference boundary, we suggest three algorithms, which are based on t-distribution, Max-Min distances, and Average distances.

We performed a series of experiments to see how the accuracies of suggested algorithms differ with mobile image transaction data. In mobile business, new image item release is a frequent event, and its purchasing ratio is noticeably high (Korea IT Industry Promotion Agency, 2008), so the mobile image transaction data is used for our experiments. The experimental result is compared with that of others, and some discussions are given about the comparative analysis.

2. Backgrounds

2.1 Recommendation Systems

Recommendation System is applying statistical and knowledge discovery techniques to suggest products during a live customer interaction.

Two of the most prevalent approaches are content-based (CB) and collaborative filtering (CF) approach.

CB approach gives recommendations to customers by automatically comparing representations of content contained in an item which have rated by the customer to representations of content contained in an item to be recommended. The best character of CB approach lies in its independence on information provided by other customers. In CB recommendation system, each customer is supposed to be treated independently and the system demands a profile of the customer's requirements and preferences in order to be able to recommend an item list. The customer profile contains preference information about content of items. Using these items as a basic information, the system searches similar items, which are suggested as recommendation list.

CF approach generates recommendations by employing overlap of preference ratings to integrate the opinions of 'like-minded' customers. Typically, a customer profile in a CF-based recommendation system is made up of a vector of items and their ratings. That is CF approach is an alternative information estimation approach based on the valuations of people. The main idea of CF algorithm is to give item recommendations or predictions based on the opinions of other like-minded customers. It works by building a preference database for items by customers. It tries to automate the 'word of mouth' recommendations. In other words, this algorithm is very similar to the procedure in which we ask relatives, friends, and coworkers for recommendations. It illustrates customers whose inclination is similar to that of a given customer.

CF algorithm has been known to the most successful recommendation method that has been used in many different applications. Nevertheless, CF based recommendation systems have fundamental weaknesses such as sparsity problem i.e. short of item rating, gray sheep problem i.e. customers whose preference is unusually different than others, a new item ramp-up problem i.e. a new item that has not had enough ratings cannot be easily recommended. Especially, the new item ramp-up problem is one of the serious problems of CF algorithm. Although many researchers have been studied hybrid algorithms to overcome these limitations of CF algorithm, little of the existing proposed algorithms can be a suitable solution for the new item recommendations.

2.2 A New Item Recommendation

To solve a new item ramp-up problem, several researches have been studied. The new item is added to a system currently, so it cannot be found (Jian et al., 2004) or the new item is having just a few customers' ratings (Burke, 2002). Because there are no accessed records, no ratings from customers and still no connections, all of which make it hard for customer to find out the new items.

Most of previous approaches for recommendation system have to rely on the usage history of customers to focus on the current request of customers. So it's an unavoidable that they are not suitable to recommend new items. There is not much research for recommending new items. Kim et al.(2006) supposed the MOBICORS-music (MOBIle COntents Recommendation System for Music) system for recommending new music item in mobile web environment. MOBICORS-music is a hybrid system. Basically it follows the procedure of CF system, but it uses the data representation of CB system by using the content-based features of music. Based on this data representation, they argue that MOBICORS-music can solve the new item ramp-up problem and its performance is better than other CF systems. They introduced the concept of preference boundary using t-distribution. But their methods and experiences are performed for all items, not specified for the new item only. Celma et al.(2005) has proposed the system that uses the Friend of a Friend (FOAF) and RDF Site Summary (RSS) vocabularies for recommending music to a user, depending on the user's musical preference and listening habits. This system, however, needs an additional effort to get individual preference of customers. Cornelis et al.(2007) have proposed recommendation method using fuzzy logic techniques which reflect the rating information and unknown information in the domain. However it applied only in the context of trade exhibition recommendation for e-government, not evaluated on real transaction data. Jian et al.(2004) have proposed recommendation method for new items based on indexing techniques. This method presents a different view of a semantic knowledge into the recommendation process based on information retrieval techniques, but this system additionally needs the description of individual preference of customers.

3. Methodology

3.1 Overall Procedure

A general recommendation algorithm uses bi-matrix of customer-item as a customer profile, but our research uses the multi dimensional feature space presented by customer's purchased items as the customer profile to recommend new items. An item is presented as items' feature vector $X_i = \{X_{i1}, X_{i2}, \cdots X_{ik}\}$ in the k-dimensional feature space.

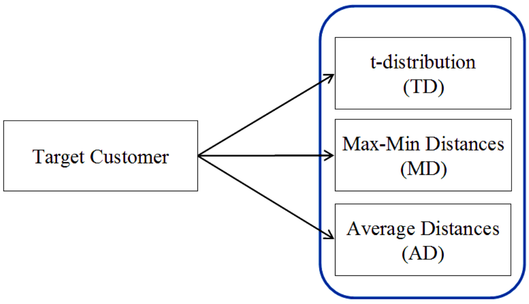
The new item recommendation procedure consists of the following four steps. The first step is selecting new items which are recommended for customers. The new items are just released and need to promote the sales.

The second step is building item profiles. We find the feature values of not only new items but also purchased items. The feature values of purchased data are used to find the preference boundary of target customer, and the feature values of the new items are used to decide whether to recommend the new item or not.

The third step builds the preference boundary of customers. This step has two sub-steps. The first sub-step is to find the representative point of the target customer's preference boundary, the centroid, based on the set of customer's purchased item set. The second sub-step is to define the range of preference boundary, the radius, so we suggest three methods using t-distribution, Max-Min Distances, and Average distance. We give a full detail of these sub-steps in next sections.

The final step is to decide the target customers, whose preference boundaries include the new item. The fact that a new item is in the preference boundary of the target customer means the target customer will like it, so the new item is recommended to the customer.

<Figure 1> illustrates three algorithms to determine the range of preference boundary of target customer; t-distribution (TD), Max-Min Distances (MD), and Average Distances (AD).



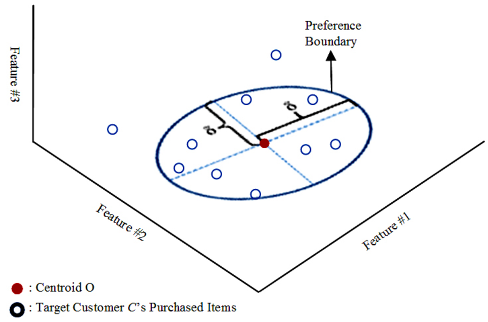
<Figure 1> Three Algorithms for Recommending New Items

Preference boundary is defined in the feature space based on the purchase information of a target customer. The basic concept of defining the preference boundary is that each customer is assumed to have a constant preference when they purchase items. So if a new item is in the target customer's preference boundary, the target customer is assumed to prefer that item than other items. However knowing the exact preference boundary of the target customer is not easy, so we propose an algorithm for the preference boundary based on his/her past purchase history. This step is to define the data scope when the algorithm builds the preference boundary. Some notations used in our research and their meanings are as follows :

- C : Subscript for the target customer
- I : Subscript for new items
- O : Subscript for the centroid
- k : Subscript for the dimension in item feature spaces
- δ : Subscript for the deviation of the preference boundary
- n : Subscript for neighbors of the target customer

3.2 Determination of the Centroid of Preference Boundary

Preference boundary is defined in the feature space based on the purchase information of a target customer. Algorithm for finding the centroid of tar-



<Figure 2> Finding the Centroid of Target Customer

get customer is to organize his/her preference boundary based on his/her purchase history only. Preference boundary of target customer C is determined by the centroid O and the deviation d of each features using C's purchase history, $\{O_c^k - \delta_c^k,\ O_c^k + \delta_c^k\}$ for any k in feature space. So a new item I is recommended to target customer C when $O_c^k - \delta_c^k \leq I_c^k \leq O_c^k + \delta_c^k$ for all k.

A proper value of the deviation δ is not known exactly so it is determined by several experiments. The result of the deviation variation is illustrated in next section. The preference boundary of target customer in three dimensional feature spaces is shown in <Figure 2> and the algorithm for finding the centroid of target customer is summarized in <Figure 3>.

Algorithm : Centroid of Target Customer

Input :

T_date : Launching date of items

Output:

A centroid of the target customer, O_c

Method :

(1) **for each** target customer *C* **do**,

(2) Find centroids for features based on items which are purchased in the training period by the target customer;

(3) **endfor**

(4) return O_c;

<Figure 3> The Algorithm for the Centroid of Target Customer

3.3 Determination of the Range of Preference Boundary

This step is to determine the range of preference boundary range of each customer. Preference boundary is determined as the customer's preference range in feature space using purchase history. This research suggests three algorithms using t-distribution (TD), Max-Min distance (MD), and Average distance (AD). These algorithms are based on the t-distribution of statistics (Kim et al., 2006) and clustering algorithms in data mining (Han and Kamber, 2006).

3.3.1 Preference Boundary by t-distribution

Based on customer C's preference boundary, a new item I is recommended when $O_c^k - \delta_c^k \leq I_c^k \leq O_c^k + \delta_c^k$, *for all k*. To determine the deviation d of each feature, we can use the standard deviation function of normal distribution function. However the standard deviation function assumes that the number of items is over 30. But when we analyzed some real transaction data, the number of new items is usually smaller than 30. Therefore we decide to use t-distribution in this research. The confidential level of t-distribution is set up considering the recommending items or domain environment. So a new item I is recommended when,

$$O_c^k - t(\frac{\alpha}{2}, df) \leq I_c^k \leq O_c^k + t(\frac{\alpha}{2}, df),$$
for all k.

Algorithm : Preference Boundary by t-distribution
Input :
T_date : Launching date of items
xsDev : the range of standard deviation
Output :
The recommended new item set for target customer
Method :
(1) *new_item_set* = {};
(2) **for each** target customer *C* **do**,
(3) **repeat**
(4) Find the standard deviation (*SDEV*) of target customer
(5) Calculate the distance (*CI_DIST*) between new item *I* and the centroid O_c of target customer
(6) Recommend when *CI_DIST* <= *xsDev* * *SDEV*
(7) **until** new items from I_1 to I_{ni}
(8) **endfor**
(9) return *new_item_set*;

<Figure 4> The Algorithm for Preference Boundary by t-distribution

df means the degree of freedom and it is calculated by the number of sample size - 1. $t(\frac{\alpha}{2}, df)$ means t value which has the possibility $\frac{\alpha}{2}$ of t-distribution. In this paper, we perform experiments

Algorithm : Preference Boundary by Max-Min Distances
Input :
T_date : Launching date of items
α
Output :
The recommended new item set for target customer
Method :
(1) *new_item_set* = {};
(2) **for each** target customer *C* **do**,
(3) **repeat**
(4) Find the minimum distance (*minD*) and the maximum distance (*MaxD*) from the transaction data to the centroid O_c of target Customer;
(5) Calculate the distance (*CI_DIST*) between new item *I* and the centroid O_c of target customer;
(6) Recommend when *CI_DIST* <= *α* * *minD* + (1-*α*) * *MaxD*;
(7) **until** new items from I_1 to I_{ni};
(8) **endfor**
(9) return *new_item_set*;

<Figure 5> The Algorithm for Preference Boundary by Max-Min Distances

in which we varied the α value from 0.01, 0.05 to 0.10. It represented 1σ, 2σ, and 3σ respectively. It means that we conduct experiments with the confidence level of 90%, 95%, and 99%. The algorithm for t-distribution is summarized in <Figure 4>.

3.3.2 Preference Boundary by Max-Min Distances

In data mining, four widely used measures for distance between clusters are minimum distance, maximum distance, mean distance and average distance (Han and Kamber, 2006). The algorithm suggested in this section is using the minimum and the maximum distance at the same time. This distance is calculated between the centroid of the customer and his/her purchased items in feature space. A new item I is recommended when

$$O_c^k - \{\alpha \cdot Min + (1-\alpha) \cdot Max\} \le I_c^k$$
$$\le O_c^k + \{\alpha \cdot Min + (1-\alpha) \cdot Max\}, \textit{ for all } k.$$

α value is between 0 and 1 in an increment of 0.1. The algorithm for Max-Min Distances is summarized in <Figure 5>.

3.3.3 Preference Boundary by Average Distances

The above minimum and maximum distances represent two extremes in measuring the distance in preference boundaries. They tend to be overly sensitive to outliers or noisy data. The use of average distance compromises the minimum and maximum distances and overcomes the outlier sensitivity problem. This algorithm defines radius of the preference boundary as an average distance between the centroid of the target customer and transaction data. A new item I is recommended when $O_c^k - d_{avg}^k \le I_c^k \le O_c^k + d_{avg}^k$, *for all k* The algorithm for Average Distances is summarized in <Figure 6>.

Algorithm : Preference Boundary by Average Distances

Input :

T_date : Launching date of items

Output :

The recommended new item set for target customer

Method :

(1) *new_item_set* = {};
(2) **for each** target customer, *C* **do**,
(3) **repeat** while new items from I_1 to I_{ni};
(4) Find the average distance (d_{avg}) between the centroid O_c of target customer and transaction data;
(5) Calculate the distance (*CI_DIST*) between new item I and the centroid of target customer;
(6) Recommend when *CI_DIST* <= d_{avg};
(7) **until** new items from I_1 to I_{ni};
(8) **endfor;**
(8) return *new_item_set*;

<Figure 6> The Algorithm for Preference Boundary by Average Distances

4. Experimental Evaluation

4.1 Data Preparation

For our experiments, we use customer, image and transaction data as basic data set from the S telecom company, which is one of the largest telecom companies in Korea. Based on these basic data sets, we suggest training data set, test data set, target customers' data set, and new item data set from an experimental design.

- ***Transaction Data*** : S company server has collected the transaction data during the period between 1st June 2004 and 31st August 2004. We obtained a transaction data set in the form of <customer_id, download_datetime, image_id> which show customers' purchased images. We know that the customer's preference is represented by features of more often purchased image items. This data set contains transactions of 1,921 customers and 8,876 images. In total, the data set contains 55,284 records. <Table 1> provides an example of transaction data set.

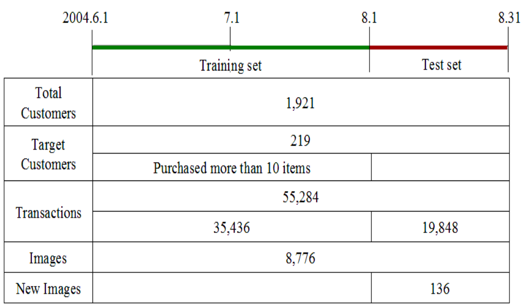
<Table 1> Example of Transaction Data Set

	customer_id	download_date	image_id
1	826	20040601	2562
2	833	20040601	2812
3	826	20040601	2709
4	856	20040601	1105
5	833	20040601	2170
6	826	20040601	2551
7	826	20040601	2640
8	851	20040601	452
9	852	20040601	2858
10	846	20040601	2093
11	856	20040601	2132
12	851	20040601	1750
13	851	20040601	2373
14	856	20040601	1333
15	856	20040601	1333
16	826	20040601	2559
17	833	20040601	2815
18	856	20040601	2400
19	862	20040601	2838
20	887	20040601	2865
21	869	20040601	2839

- ***Image Data*** : S Telecom Company has dealt with 8,776 image items. <Table 2> shows items sold in S Telecom Company. <Table 3> shows the features of image characters. This shows the stored indexing data which are the features of S telecom image data extracted using Meanshift segmentation method (Comaniciu et al., 1997). Using these feature data, we can present images in vector space which have nine dimensions.

4.2 Experimental Design

We spilt the period between 1st June 2004 and 31st July, and the period between 1st August and



<Figure 7> Experimental Design

31st August 2004 as a training set and a test set, respectively. The training set is used to make the preference boundary, and the test set is used to evaluate the performance of the suggested hybrid recommendation systems. As the target customers, we selected 219 customers who have purchased more than ten images in training period. Finally, the training set consists of 6,499 transaction records, and the test set consists of 2,514 transaction records created by the target customers. 136 items are selected as new

<Table 2> Example of Image Data Set

	image_id	character_no	character_title	service_start_date	category
1	1	76511	소유진♡사랑은솜사탕	20020404142405	소유진
2	2	81936	류진◉그리움이더해~!	20020424175113	류진
3	3	84331	최정윤-♥내마음♥..	20020508174151	최정윤
4	4	87298	보아-늘기다릴께~	20020523143030	보아
5	5	87306	보아-NO.1	20020523143619	보아
6	6	88385	강성연♥섹시꽃요정	20020530092717	강성연
7	7	89505	최강희♥사랑의정의	20020604100509	최강희
8	8	91195	성인전용♥로딩중	20020614175002	잼버거◆인기짱
9	9	93581	장혁-내꿈	20020624134315	조이포토샵
10	10	94302	민지혜★사랑스러운걸	20020629091720	민지혜
11	11	95511	플라이-우정변치말자	20020705110248	RytotheSky
12	12	97429	민지혜♥네가그리워	20020718095005	민지혜
13	13	100181	짱이뻐요보아	20020805091723	보아
14	14	100966	김민희♥순수한첫사랑	20020811092052	김민희

<Table 3> Example of Image Feature Set

	character_no	f1	f2	f3	f4	f5	f6	f7	f8
1	100181	0.11	0.15	0.75	0.239	0.304	0.59	0.028	0.022
2	100966	0.34	0.21	0.341	0.67	0.478	0.794	0.014	0.016
3	101788	0.125	0.449	0.525	0.202	0.423	0.592	0.04	0.014
4	102072	0.546	0.469	0.354	0.653	0.768	0.68	0.013	0.011
5	104702	0.411	0.407	0.665	0.569	0.734	0.612	0.013	0.012
6	106698	0.391	0.247	0.488	0.597	0.538	0.833	0.015	0.012
7	107449	0.358	0.276	0.425	0.693	0.47	0.647	0.011	0.014
8	110954	0.329	0.178	0.45	0.601	0.341	0.715	0.014	0.014
9	110958	0.27	0.47	0.334	0.336	0.468	0.447	0.029	0.012
10	111209	0.525	0.192	0.628	0.852	0.446	0.717	0.01	0.019
11	111441	0.32	0.31	0.542	0.639	0.639	0.686	0.015	0.017
12	111775	0.3	0.112	0.438	0.575	0.619	0.592	0.011	0.024

items, which are released after Aug, 1, 2004. <Figure 7> shows the overall description of experimental design.

4.3 Evaluation Metrics

Recommendation system research has various measures for evaluating the effectiveness and efficiency of recommendation system. The main aim of this research is to compare with suggested recommendation algorithms, and find out successful recommendation algorithm which has better quality compared other algorithms. To evaluate the performance of each algorithm new items we compare the recommending new item list purchased in test period with the recommended new item list by suggested algorithms.

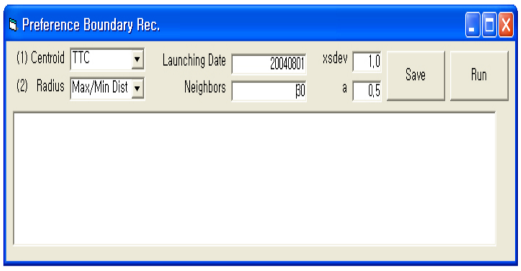
Recall and Precision have been widely used to test recommendation quality in recommendation systems (Billsus and Pazzani, 1998; Lin et al., 2002; Cho and Kim, 2004). Recall is defined as the ratio of the number of items in both the purchased item list and the recommended item list to the number of items in the purchased item list. Recall means how many of items in the customer's purchased item list are recommended properly. Precision is defined as the ratio of the number of items in both the purchased item list and the real recommended item list to the number of items in the recommended item list. Precision means how many of the recommended items belong to the real customer's purchased item list.

These measures are clear to evaluate and intuitively attractive, however they are in agitation since increasing the size of recommendation set leads to an increase in recall but at the same time a decrease in precision. So a combination metric, F1 metric is widely used (Sarwar et al., 2000; Cho and Kim, 2004).

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

4.4 Experimental Environments

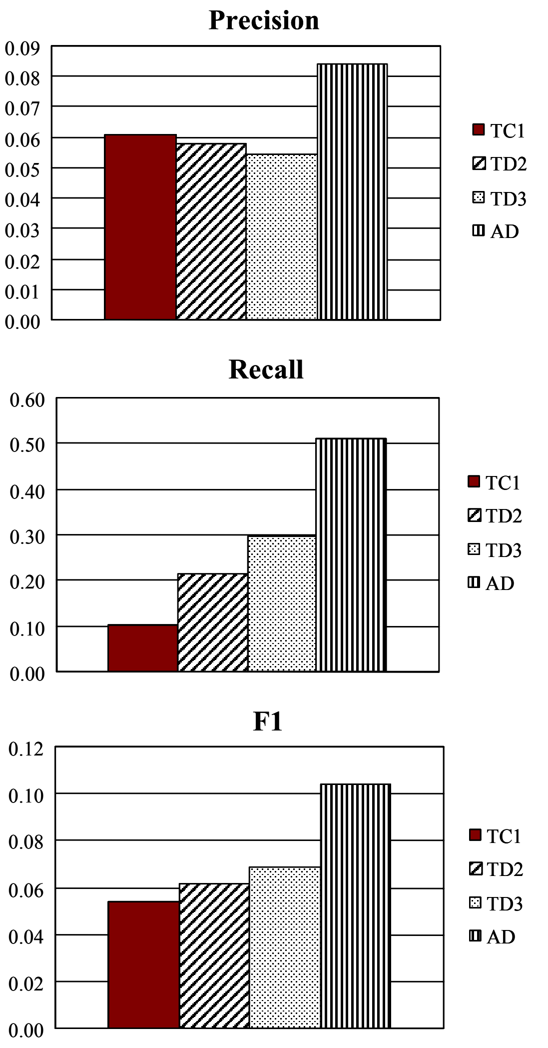
We test the suggested algorithms in Visual basic 6.0 and ADO components, and MS-SQL Server 2000. The system is running on window XP-based PC with Intel Core 2 Quad CPU having a speed 2.40GHz and 3.24GB RAM. <Figure 8> shows our system's interface.



<Figure 8> System Interface

4.5 Experiment Results

In this section, we present a detailed experimental evaluation of the different algorithms. First, we perform experiments in which we varied the range of standard deviation and calculate precision, recall and F1 metric of the algorithm for preference boundary by t-distribution (TD). <Figure 9>



<Figure 9> Result of TD and AD

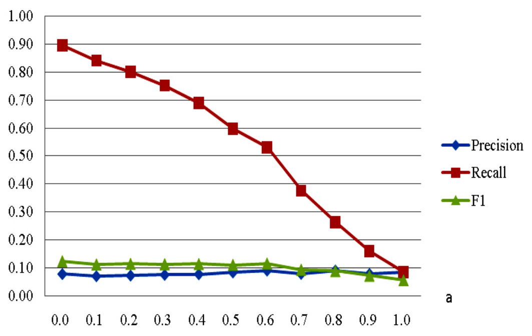
shows our experimental result of TD. TD1 means using t-distribution with standard deviation 1σ. TD2 means using t-distribution with standard deviation 2σ. And TD3 means using t-distribution using standard deviation 3σ.

As looking into the result, we find that the range of standard deviation does affect the quality of recommendation. The precision decreases and the recall increases as increasing the standard deviation value (from 1σ to 3σ). As a result, F1 metric increases as increasing the standard deviation. So TD3 results in the best algorithm among those for preference boundary by t-distribution.

<Figure 9> also shows the results of AD, the algorithm for preference boundary by average distance.

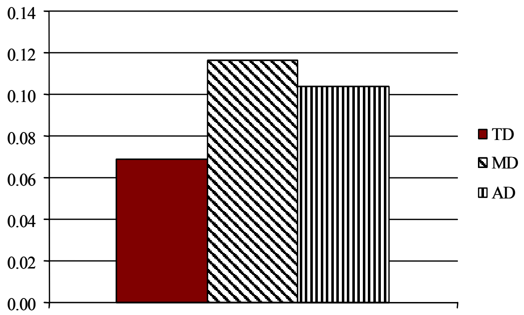
To test the performance of Max-Min Distances (MD), we varied the value of α from 0 to 1.0 in $\alpha \cdot Min + (1-\alpha) \cdot Max$. If α is 0.0, then the preference boundary is the maximal set of feature values of target customer's purchase item set. In contrast, if α is 1.0, then the preference boundary is the minimal set of feature values. <Figure 10> shows the experimental result of MD. As looking into the result, the recall decreases and precision increases as increasing the value of α. This is similar to the experimental result of TD. When α is 0.6, F1 metric has the largest value.

Base on these results, the comparison of three algorithms are summarized in <Figure 11>. We can



<Figure 10> Result of MD

see that using MD is most effective to recommend new items to customers. And with our mobile commerce data, the value of α is 0.6.



<Figure 11> Comparison of three algorithms

5. Conclusion

We suggest three algorithms to recommend new items, and evaluate the suggested algorithms with real transaction data of mobile image. The contribution of this research is summarized as follows.

The research suggests preference boundary-based procedures to solve the new item ramp-up problem, which is a well known problem of CF. The suggested procedures are heuristic algorithms, but their logical basis is adopted from statistics (t-distribution), and data mining (Max-Min Distances, and Average Distances). Compared to existing studies, three algorithms are run with only the purchase history of customers and feature values of items. So our suggested procedures are easy to be applied to other automatic domains such as ubiquitous computing environment.

With real transaction data of mobile images, we perform experiments to see the result of three algorithms. While our experiments give promising result to solve the new item ramp-up problem, these results are based on the particular mobile image transaction data set. So it is required to evaluate our algorithms in more detail using other data sets from diverse commerce environments. And it will be an interesting research area to suggest other algorithms related with our suggested ones. For instance, in this research we only use target customer's purchased data to make the preference boundary, but we can also use the purchased data of the neighbors who have similar preference with the target customer.

References

Avery, C. and R. Zeckhause, "Recommendation systems for evaluating computer messages", *Communications of the ACM*, Vol.40, No.3(1997), 88～89.

Balabanovic, M. and Y. Shoham, "Fab : Content-based, Collaborative Recommendation", *Communications of the ACM*, Vol.40, No.3(1997), 66～72.

Billsus, D. and M. J. Pazzani, "Learning collaborative information filters", Proceedings of the Fifteenth International Conference on Machine Learning, 1998, 46～54.

Burke, R., "Hybrid Recommendation Systems : Survey and Experiments", *User Modeling and User-Adapted Interaction*, Vol.12, No.4(2002), 331～370.

Celma, O., M. Ramirez, and P. Herrera, "Foafing the music : A music recommendation system based on RSS feeds and user preferences", 6th

International Conference on Music Information Retrieval, 2005, 464～457.

Cornelis, C., J. Lu, X. Guo, and G. Zhang, "One -and-only item recommendation with fuzzy logic techniques", *Information Sciences*, Vol. 177, No.22(2007), 4906～4921.

Cho, Y. H. and J. K. Kim, "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce", *Expert Systems with Applications*, Vol.26, No.2 (2004), 233～246.

Comaniciu, D. and P. Meer, "Robust Analysis of Feature Spaces : Color Image Segmentation", Proceedings of the 1997 Conference on CVPR, 1997, 750～755.

Good, N., J. B. Schafer, J. A. Konstan, A. Borchers, B. M. Sarwar, J. Herlocker, and J. T. Riedl, "Combining Collaborative Filtering with Personal Agents for Better Recommendations", Proceedings of AAAI'99, 1999, 439～446.

Han, J. and M. Kamber, *Data Mining : Concept and Techniques*, 2nd ed., Morgan Kaufmann Publishers, 2006.

Jian, C., Y. Jian, and H. Jin, "Recommendation of New Items Based on Indexing Techniques", Proceedings of the Third International Conference on Machine Learning and Cybernetics, 2004, 1168～1172.

Kim, J. K., Y. H. Cho, M. Y. Kang, and H. K. Kim, "A Hybrid Multimedia Contents Recommendation Procedure for a New Item Problem in M-commerce", *Journal of Korea Intelligent Information Systems Society*, Vol.12, No.2 (2006), 1～15.

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

Sarwar, B. M., G. Karypis, J. A. Konstan, and J. T. Riedl, "Analysis of Recommendation Algorithms for E-Commerce", Proceedings of the ACM Conference on Electronic Commerce, 2000, 158～167.

The Korea IT Industry Promotion Agency Homepage, www.software.or.kr/index.html.

Abstract

신상품 추천을 위한 확장된 내용기반 추천방법

장문경* · 김혜경* · 김재경*

현재 다양한 신상품의 잦은 출시로 인해 고객들은 자신이 원하는 신상품을 찾는데 어려움을 겪고 있다. 또한 기업들은 신상품을 구매할 가능성이 높은 고객을 찾는데 많은 노력을 기울이고 있는 상황에서 고객의 선호에 부합하는 신상품을 찾도록 도와주는 추천시스템에 대한 요구가 대두되고 있다. 본 연구는 신상품 추천을 위해 상품 특성을 추출하여 다차원 속성 공간에 표현하고 이를 바탕으로 선호영역(Preference Boundary)를 제시하였다. 다시 말해 고객들이 과거 구매한 상품의 속성을 바탕으로 고객의 선호 영역을 형성하고, 신상품의 속성이 선호 영역 내에 위치하면 추천이 이루어지는 방법을 제시하였다. 선호 영역을 형성하는 과정은 크게 선호영역의 중심점을 구하는 단계와 선호영역의 범위를 구하는 단계로 구성되는데, 이 연구에서는 선호영역의 범위를 구하는 단계로 t-분포를 이용하는 방법, 중심점과 구매 상품과의 가장 먼 거리와 가까운 거리를 이용하는 방법, 그리고 중심점과 구매 상품들 간의 평균 거리를 이용하는 방법을 제시하였다. 제시된 방법들의 성능을 평가하기 위해 신상품 출시와 구매가 잦은 모바일 이미지 거래 데이터를 이용하여 실험을 진행하였다. 이 논문에서 제시한 각 방법들의 성능을 비교해본 결과 목표 고객의 중심점과 구매 상품과의 가장 먼 거리와 가까운 거리를 이용하는 방법으로 각 상품별 선호영역의 적정한 범위를 구하였을 때, 신상품 추천의 정확도가 향상되는 것으로 분석되었다.

Keywords : 신상품추천, 멀티미디어 컨텐츠, 내용기반필터링

※ 본 논문은 2008년 한국지능정보시스템학회 추계학술대회에서 우수논문상을 수상하였음.
* 경희대학교 경영대학, 경영연구원

저 자 소 개

장문경
경희대학교 경영학과에서 학사를 취득하고, 롯데정보통신에서 근무하였다. 현재 동 대학원에서 MIS 전공 석사과정에 재학 중이다. 관심분야로는 CRM, 추천시스템, 데이터마이닝 등이다.

김혜경
경희대학교 물리학과에서 학사, 경영학과에서 MIS 전공으로 석사 학위를 취득하고, 현재 동 대학원 박사과정에 재학 중이며 2009년 2월 졸업예정이다. 주요 관심분야는 상품추천시스템, 복잡계 시스템 등이다. Expert Systems : The International Journal of Knowledge Engineering and Neural Networks, Expert Systems With Applications, Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence 등에 논문을 게재하였다.

김재경
서울대학교에서 산업공학 학사, 한국과학기술원에서 경영정보시스템 전공으로 석사 및 박사학위를 취득하였으며 현재 경희대학교 경영대학 교수로 재직하고 있다. 미국 미네소타 주립대학교 및 텍사스 주립대학교(달라스)에서 교환교수를 역임하였다. 주요 관심분야로는 비즈니스 인텔리전스, 추천시스템, 유비쿼터스 서비스 등이다.