2011 International Conference on Social Science and Humanity
IPEDR vol.5 (2011) © (2011) IACSIT Press, Singapore

Preference Boundary-based Approach for Recommending New Items

Min Kyu Jung
School of Management
Kyung Hee University
Seoul, Korea
e-mail: minkyuli@khu.ac.kr

Moon Kyoung Jang
School of Management
Kyung Hee University
Seoul, Korea
e-mail: jmoonk25@gmail.com

Hyea Kyeong Kim
School of Management
Kyung Hee University
Seoul, Korea
e-mail: kimhk@khu.ac.kr

Jae Kyeong Kim
School of Management
Kyung Hee University
Seoul, Korea
e-mail: jaek@khu.ac.kr

Abstract— **When new items are released, it is necessary to promote these items. In this situation, a recommender system specializing in new items can help item providers find potential customers. This study aims to develop a preference boundary-based procedure for recommending new items. The basic principle is that if a new item belongs within the preference boundary of a target customer, then it is evaluated to be preferred by the customer. The new item recommendation procedure is organized in the following two phases. The first phase defines each customer's preference boundary, and the second phase decides the target customer set for recommending new items. In this research, customer' preferences and item characteristics including new items are represented in a feature space. And the scope of boundary of the target customer's preference is extended to those of neighbors'. Furthermore, compared to existing recommender systems, the suggested procedure aims to find target customers for the new released items. Diverse algorithms are suggested for the procedure, and their effectiveness scores are measured and compared through a series of experiments with a real mobile image transaction data set. The experiment results are compared, and discussions about the results are also given with a further research opportunity.**

Keywords- Recommender System; Collaborative Filtering; Multimedia Content; Personalization

I. INTRODUCTION

Due to rapid growth of E-commerce, customers of a Web retailer are often overwhelmed with choices and flooded with promotional product information. A promising technology to overcome this information overload is recommender systems filtering out information that may be inapplicable to an individual or a group of individuals. Customers can browse various items, but it is not easy to fund the items that they want to purchase among many choices. Therefore item providers and customers need recommender systems that suggest right items to right customers.

In particular, when new items are introduced into the market, firms and customers can get benefits by promoting these items. In this context, it will be helpful to develop a recommender system specializing in new item recommendation. For example, in a mobile Web environment, new images are frequently supplied and their purchasing ratio to existing items is considerably high, so an image recommender system needs to evaluate new items effectively and efficiently form recommendation. However, there have been very few systems for recommending only new items [1]. That is because new items have no accessed records and no ratings from customers, which are the sources of making recommendation.

Celma et al. [3] have 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 users to select enormous information. Cornelis et al. [4] have proposed a hybrid recommendation algorithm which involves the fuzzy logic techniques, which combine the CB and CF contributions to the final recommendation.

Jian et al. [6] have proposed recommendation algorithms for new items based on indexing techniques. This method presents a different view of semantic knowledge into the recommendation process based on information retrieval techniques. Before the algorithm performs, it requires specifying a certain matching score of the customer transaction and the new item.

Previous systems for recommending new items rely on CB techniques. However, this system has some crucial drawbacks [1, 2]. Firstly, because most CB systems are based on feature analysis, they require a source of feature content information of all items under consideration. In other words, the applicability of CB systems is limited to areas in which feature values of items or textual descriptions are already available. Secondly, CB system can recommend to a customer with only items which have similar characters with the items which the customer rated high or purchased before.

This problem is known as the overspecialization problem. Therefore, recommendation item range can be narrow, because this system cannot catch the customer's potential preference. Lastly, in order to function effectively, CB systems require the customers already rated or purchased a sufficient number of items. As a result, this system is not enough to provide proper recommendations for new customers or new items.

This study aims to develop a hybrid recommender system for recommending new items. The basic idea of the suggested hybrid procedure is as follows. Customers' preferences and characteristics of items are represented as vectors in a feature space. To prevent the overspecialization problem of CB methods, the scope of boundary of preferences is extended to neighbors with similar ratings incorporating CF. If a new item belongs within the preference boundary, then, it is assumed to be preferred by the target customer.

The suggested preference boundary-based recommendation procedure is organized in two phases. The first phase defines each customer's preference boundary, and the second phase decides the target customer set for recommending new items.

Diverse hybrid algorithms are suggested and their effectiveness measures are obtained and compared through a series of experiments with a real mobile image transaction data set. We will compare these algorithms and the CB method, and draw a final suggestion..

II. METHODOLOGY

The new item recommendation procedure is organized in the following two phases. The first phase defines each customer's preference boundary, and the second phase decides the target customer set for recommending new items.

Firstly, we present every item's profile in *K*-dimensional feature space. Individual customer's profile is built by merging his/her purchased items' profiles. Then, the preference boundary of each customer is defined at the feature space comprised of the feature values of his/her preferred items. In this research, the preference boundary is determined by two characters: (1) centroid which is customer's representative point of preference boundary and *K*-dimensional radiuses, and (2) ranges in the feature space based on his/her purchased item set. To determine the centroid of preference boundary of each customer, we suggest three algorithms: (1) SC which is using the centroid of a single customer only, (2) BC which is using the centroid of a (dummy) big customer that is composed of a customer and his/her neighbors, and (3) NC which is using centroids of a customer and his/her neighbors. SC is a method developed from contents-based approach, but BC and NC are based on the concept of neighbors, which are come from CF. In order to determine the ranges of preference boundary, we use t-distribution or normal distribution.

In the second phase, we find target customers for recommending new items. When recommending new items, it is important to decide target customers who would purchase the recommended items. As a new item is also represented in *K*-dimensional feature space, the basic principle of suggested procedures is that if the new item belongs within the preference boundary of a customer, then it can be preferred by the customer. To decide the M numbers of target customers among the customers who include a suggested new item in their preference boundaries, we use Euclidean distance which is the distance between the centroid of customer's preference boundary and that of new item.

A. Representation of Preference Boundary

In general, purchased items by a customer include information about the customer's preference on items. The *personal information set (PIS)* of a customer C consists of items that customer has purchased. *PIS* is represented as $P^c = \{p_1, p_2, \ldots, p_L\}$. Each item is represented as vector $p_{ci} = \{p_{ci}^1, p_{ci}^2, \ldots, p_{ci}^k\}$ of features a in the *K*-dimensional feature space that describe its properties such as price, color, and brand. In the proposed method, a customer's actual preference is represented as a *preference boundary*, which is defined by the *centroid* and the *range* of his PIS in the *K*-dimensional feature space.

The centroid vector $O_c = \{O_c^1, O_c^2, \ldots, O_c^k\}$ is the mean vector of all item vectors in customer *c*'s PIS:

$$O_c = \frac{\sum_{i=1}^{L_c} p_{ci}}{L_c}. \quad (1)$$

B. Range of Preference Boundary

The range value δ_i obtained from the distribution of items in the PIS and a *t*-value of Student's *t*-distribution at the 90%, 95%, or 99% confidence level:

$$\delta_i = TINV(\lambda, L_c - 1). \quad (2)$$

Where denotes the error (i.e., 0.1, 0.05, or 0.01) and the size of customer *c*'s PIS. Note values 0.10, 0.05 and 0.01 represent 1σ, 2σ, and 3σ of Student's *t*-distribution, respectively. Figure 1 shows the preference boundary of customer *c*'s PIS consisting of 11 items over a 3-dimensional feature space.

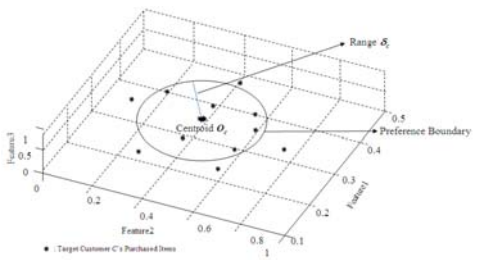


Figure 1. Preference Boundary Using Single Customer

C. Neighbor Formation

Since each item is represented as a vector in the *k*-dimensional feature space, we can obtain the neighbor set using the Euclidean distance function as the similarity measure [7]. The distance function *d(c,a)* between the target customer *c* and a potential neighbor *a*, is calculated as

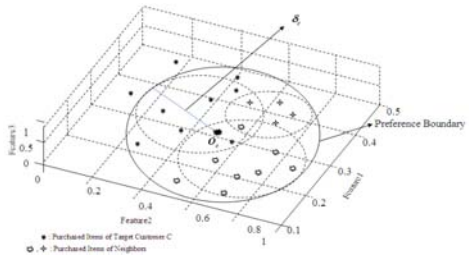
$$d(c,a)=\sqrt{\frac{\sum_{k=1}^{K}(O_c^k-O_a^k)^2}{K}}. \quad (3)$$

where O_c^k and O_a^k are *k*th feature value of centroid vector O_c and O_a, respectively. A shorter distance indicates more similarity. A target customer's neighborhood includes top-*K* most similar customers.

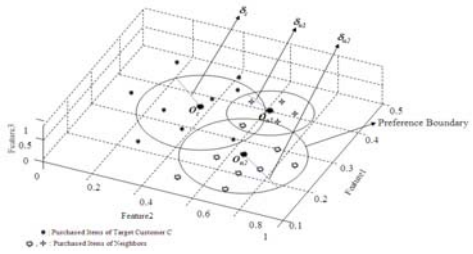
D. Phase 1: Defining Each Customer's Preference Boundary

As items are represented as points in *k* dimensional feature space, neighbors are found by calculating the distance between customer *c* and other customers. PIS of customers are represented as a cluster in feature space, so cluster distance function is used to calculate the distance between centroids of customer *c* and those of other customers [7]. The customer is assumed to have similar preference with the customer *c* if the distance is very close.

The idea is to use similar customers, neighbors derived from CF to determine the representative of customer *c*. Euclidean distance function is used as a cluster distance function in this research, because this function is simple, easy to calculate and generally used [5].



(a) BC



(b) NC

Figure 2. Preference Boundary Using BC and NC

In this research, three algorithms are suggested to define customer's preference boundary. Contrast to the first algorithm, SC, the other two algorithms BC and NC are related with CF. SC is a method developed from the typical CB approach, while BC and NC are based on the concept of neighbors that comes from CF. Thus, BC and TC are hybrid methods. Figure 1 shows an example of the TC method; Figure 2 shows examples of the BC and NC methods.

The Phase 1 also includes checking if a new item is within the preference boundary of a customer. When the TC or BC method is used in the second step, a new item I is evaluated to be within the preference boundary of customer *c*' whose simple or big customer centroid is O if the distance between the centroid vector and the new item vector, after normalized by the standard deviations of *c*'s PIS, is within the range δ_c:

$$\sqrt{\frac{\sum_{k=1}^{K}[(O_c^k-O_a^k)]/s_c^{j\,2}}{K}}\le\delta_c. \quad (3)$$

Where s is the standard deviation vector of *c*'s PIS:

$$s_c^j=\sqrt{\frac{\sum_{i=1}^{L_c}p_{ci}^k-O_c^k}{L_c}} \quad (4)$$

When the NC method is used, a new item *I* is evaluated to be within the preference boundary if (1) holds for the customer or any of his neighbors.

E. Phase 2: Finding Target Customers to Recommend New Items

When the preference boundary of each customer is generated, the next step is to find target customers for recommending the new item. The basic principle is that if the new item belongs within the preference boundary of a customer, then it can be preferred by the customer. So one way to find target customers is to find all the customers who include the new item within his/her preference boundary. Considering the cost of marketing activity such as campaign, we need to restrict the number of target customers. To decide M target customers among the customers who include a suggested new item in their preference boundaries, we use Euclidean distance between the centroid of customer's preference boundary and new item. This is choosing top M customers whose centroids are closer to the suggested new item.

III. EXPERIMENT

To evaluate the suggested algorithms, we carry out experiments with the intent to answer a main question:

- How do the approaches to determine the preference boundary affect overall performance of the recommender system for new items?

For this purpose, we develop two CF-based hybrid methods, BC and NC, which use neighbors to determine

preference boundaries, and tested whether the approaches improve the recommendation quality or not. The results of BC and NC are compared with that of SC, which uses a target customer's purchase information set only. Furthermore, to determine the range of preference boundary, or the radius, we use normal distribution and t-distribution, and several experiments are performed at confidential levels of 90%, 95%, and 99% respectively

A. The Data Set

For our experiments, we use character images and real transaction data in mobile commerce. The data set is provided by one of leading content distributers in Korea. The data set contains 8,776 image products, 1,921 customers, and their 55,321 transactions during the period between June 1, 2004 and August 31, 2004.

To characterize images, we perform the preprocessing task to extract visual features to characterize images. In this research, we use color moment —hue, saturation, and value (HSV) of color— over other choices of features such as shape of texture, because color moment is the most generally used feature and HSV represents human color perception more uniformly than others [8]. We obtain the bitmap format files whose images are represented by 256 colors. For all pixels in images, we translate the values of three-color channels (RGB or red, green, and blue) into HSV values. Then, the mean, standard deviation, and skewness for HSV values are calculated to present images as vectors in a 9-dimensional feature space.

We divide the period into two: (1) one between 1st June and 31st July to obtain a training data set, and (2) the other between 1st August and 31st August to obtain a test data set. The training data set consists of 35,436 transaction records, and the test data set consists of 19,848 transaction records created by the target customers. The training set is used to determine the preference boundaries of customers, and the test set is used to evaluate the effectiveness of the suggested algorithms.

As potential target customers, we select 219 who have purchased more than 10 images in the training period. New images are released after 1st August 2004, and purchased more than 10 times by customers during the test period. There were 136 new images satisfying these criteria. Fig. 3 shows the overall description of experimental data.

B. Measures and Experimental Environment

Precision, Recall, and $F1$ are often used to measure effectiveness of recommender systems, or information retrieval systems in general [4, 7, 9].

Figure 3. Data set Design

Recall measures the fraction of purchased items that are recommended; precision measures the fraction of recommended items that are purchased. $F1$ combines both recall and precision into a single measure:

$$F1 = \frac{2 \times recall \times precision}{recall + precision} \quad (5)$$

We used the $F1$ measure to evaluate recommendation effectiveness.

A system to perform our experiments is implemented using Visual Basic 6.0, and ADO components. The system consists of two parts: one for image data pre-processing, and the other for experiment execution and result analysis. MS-SQL Server 2000 is used to store and process all the data necessary for our experiments. We our experiments on a Window XP computer with 3.24GB RAM and an Intel Core 2 Quad CPU having 2.40GHz clock speed.

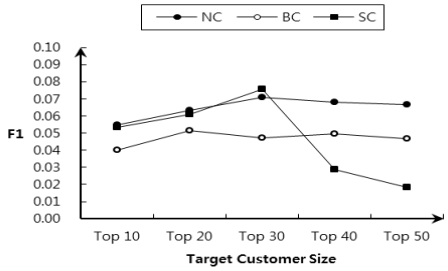
C. Recults and Discussion

Since the quality of CF or CF-based hybrid algorithms varies with the neighborhood's size, we perform an initial experiment to determine the optimal size. In our view, a neighborhood size of 10 is reasonable and its results are reported in the rest of the paper. That is because the use of other neighbor set sizes between 5 and 35, does not make significant difference in observed behaviors of recommender systems.

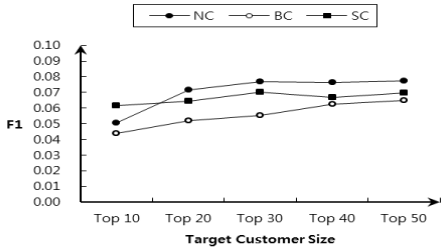
The number of target customers to whom new items are recommended effects on the quality of recommendation. It depends on the application area, the number of items, the number of customers, and so on. The total number of candidate target customers is 219, so recommending a new item to large number of customers is rather impractical. Therefore, we consider target customer set sizes of up to 50.

Fig. 6 shows our results of SC, BC, and NC with three different values of standard deviations 1σ, 2σ, and 3σ for Student's for Student's *t*-distribution, respectively. We observe from Fig. 7 that NC works better than SC at all standard deviations for Student's *t*-distribution except top 10 target customers. It indicates that very small number of target customers obtained from neighbors causes poor reflection of the preference of single customer, which leads to lower quality of recommendations. A target customer set of size between 30 and 50 seems a good choice, and we consider 30 a cost effective choice.

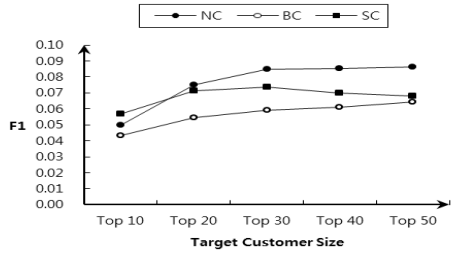
Further, we can see that BC is worse than NC and TC. This is explained by the fact that extending preference boundary using big (dummy) customer result in worse performance than his/her own preference boundary. Therefore, we find that NC is more effective to extend the preference boundary than BC.



(a) Preference Boundary Range(1σ)



(b) Preference Boundary Range(2σ)



(c) Preference Boundary Range(3σ)

Figure 4. Evaluation of BC, and NC compared to SC

Fig. 6 also shows the effect of the size of preference boundary. In general, a wider preference boundary increases the recommendation quality of NC. Our experiments, however, reveal that when the preference boundary range is 1σ, there are radical changes of F1 values of SC. In Fig. 6(a), the F1 of SC decreases considerably when M is over 30. The reason may be thought of customers who have the new items in their preference boundary also decreases, so the resulting F1 decreases rapidly after 30. When the standard deviation is 2σ, and 3σ, the number of new items in preference boundary may be to be sufficient to result in stable F1 of SC.

IV. CONCLUSION

CF and CB are the best known recommendation algorithms, but they are not enough for new item recommendation. Lack of new item recommendation is known to be one of deficiencies of CF-based recommender systems. Although previous studies have developed CB-based methods to address this problem; The SC method in this paper is a typical CB-based method. We proposed two hybrid methods (combining CB and CF-based techniques). Both BC and NC methods use not only a target customer's data, but also his neighbors' data when obtaining preference boundary. When determining the centroid, the BC method uses both the target customer's data and his neighbors' data; but the NC method, instead, keeps the target customer's and neighbors' original centroids. Among them, we found that the NC method performed the best; but the other hybrid method (BC) performed even worse than the CB-based method (SC)

Accordingly, it is needed to expand into validation with data from other domains (e.g. department store transactions). Algorithms of the procedures also have rooms for further improvement and variations. For instance, when defining preference boundaries or determining neighbors, we use simple Euclidean distance, but a certain measure of degree of similarity can be used to have a more flexible algorithm. In addition, when defining preference boundaries, we use Student's *t*-distribution, but other criterion such as average or max-min distance can be used. We are currently considering all these possibilities of improving the algorithms.

ACKNOWLEDGMENT

This work was supported by the Knowledge service Ubiquitous Sensing Network(USN) Industrial Strategic technology development program, 10035426, Personalization marketer for an intelligent exhibit marketing funded by the Ministry of Knowledge Economy(MKE, Korea).

REFERENCES

[1] Adomavicius G, Tuzhilin A "Toward the nest generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Transactios on Knowledge and Data Engineering , Vol 17, no.6,2005,pp.734-749

[2] Burke R "Hybrid Recommendation Systems: Survey and Experiments. User Modeling and User-Adapted Interaction," Vol 12, no.4: 2002, pp.331-370

[3] Celma O, Ramirez M, Herrera P. "Foafing the music: A music recommendation system based on RSS feeds and user preferences. in ISMIR: 2005, pp.464-457

[4] Cornelis C, Lu J, Guo X "One-and-only item recommendation with fuzzy logic techniques," Information Sciences ,Vol 177, no.22: 2007, pp.4906-4921

[5] Ishikawa Y, Subramanya R, Faloutsos C "MindReader: Querying databases through multiple examples," Proceedings of the 24rd International Conference on Very Large Data Bases: 1998, pp.218-227

[6] Jian C, Jian Y, Jin H "Recommendation of New Items Based on Indexing Techniques," Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, no.26-29: 2004, pp.1168-1172

[7] Han J, Kamber M "Data Mining: Concept and Techniques," 2nded. Morgan Kaufmann Publishers, 2006.

[8] Porkaew K, Chakrabarti K, Mehrotra S "Query Refinement for Multimedia Similarity Retrieval in MARS," In Proceedings Of the 7th ACM Multimedia Conference: 1999, pp.235-238

[9] Sarwar B, Konstan J, Herlocker J, Miller B "Using filtering agents to improve prediction quality in the GroupLens research collaborative filtering system," In Proceedings of the 1998 ACM conference on Computer supported cooperative work: 1998, pp.345-354