A New Item Recommendation Procedure Based on Min-Max Distance

Min Kyu Jung, Moon Kyoung Jang, Hyea Kyeong Kim, and Jae Kyeong Kim

Abstract—**When new items are released, it is necessary to promote these items. In this situation, a recommender system specializing in new items help item providers find potential customers. This study aims to suggest a Min-Max distance-based preference boundary, and to develop a preference boundary-based recommender procedure applied 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 based on min-max distance, 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'. 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 Systems; 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, which filter 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 a 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 used as the sources of making recommendation.

Celma et al. [4] 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. [5] 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. [9] 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 generally rely on CB techniques. However, these systems have some crucial drawbacks [1, 3]. 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, recommended 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:

1) Customers' preferences and characteristics of items are represented as vectors in a feature space.
2) The preference boundary of each customer is defined by purchased or evaluated items, represented as vectors in

Manuscript received April 3, 2011. 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)

M. K. Jung, M. K. Jang, and H. K. Kim are with the School of Management, Kyung Hee University, 1 Hoeki-dong, Dongdaemoonku, Seoul 130-70, Korea.

J. K. Kim is with the School of Management, Kyung Hee University, 1 Hoeki-dong, Dongdaemoonku, Seoul 130-70, Korea (corresponding author to provide phone: +82-2-961-9355; fax: +82-2-961-0515; e-mail: jaek@khu.ac.kr).

feature space.

3) To prevent the overspecialization problem of CB methods, the scope of boundary of preferences is extended to include those of neighbors.
4) In the extended preference boundary, the range of preference boundary is not known exactly, so minimum range and maximum range are used in this research.
5) 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 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.

II. Methodology

A. Overall Procedure

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.

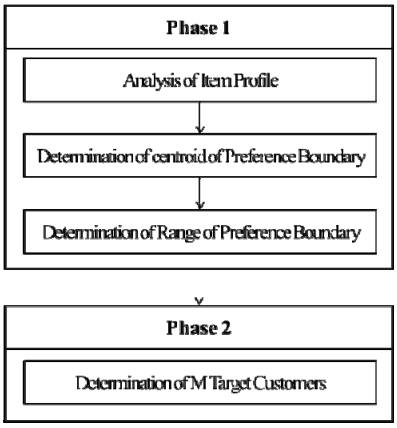


Figure 1. Overall Procedure

To determine the centroid of preference boundary of each customer, we suggest two algorithms: (1) SC which is using the centroid of a single customer only, (2) NC which is using centroids of a customer and his/her neighbors. SC is a method developed from contents-based approach, but NC are based on the concept of neighbors, which is come from CF. In order to determine the ranges of preference boundary, we use Min-Max distance. Here, maximum distance is referred to as the maximum preference boundary including all the purchased items of customer. Likewise, minimum distance distance is referred to as the minimum distance including only the centroid of customer. So preference boundary of each customer is decided by experiments, and the optimal ranges are between minimum distance and maximum distance.

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 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. The centroids of M target customers are closer to new item than those of other customers.

B. Representation of preference 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 $P^c = \{P_1, P_{2,} \cdots, P_L\}$. Each item is represented as vector $p_{ci} = \{p_{ci}^1, p_{ci}^2, \cdots, p_{ci}^k\}$ 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 $O_c = \{O_c^1, O_c^2, \cdots, O_c^k\}$ is the mean vector of all item vectors in customer *c*'s PIS:

$$O_c = \sum_{i=1}^{L} P_{ci} / L_c \quad (1)$$

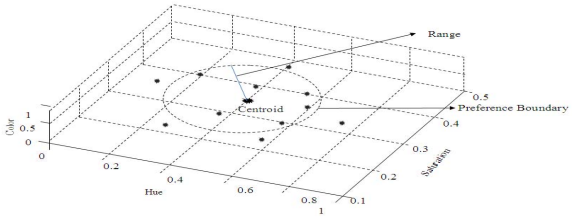


Figure 2. Representation of Preference Boundary

For an illustrative example, see Fig. 2, which shows the preference boundary composed of range and centroid vector of a personal information set consisting of 11 images over a 3-dimensional feature space. Each image is represented as collection of all possible visual features that describe its perceptual properties such as HSV (i.e. hue, saturation, and value of color) based color moment, shape and texture. Six images within the preference boundary are preferred by the target customer, and the other images outside of boundary are

not preferred.

C. Range of Preference Boundary

After determining the centroid of the preference boundary, we need to define the range of preference boundary. To represent the possible range of preference boundary, this research suggests to using minimum and maximum distances.

This distance should be calculated between the centroid of the customer and his/her purchased items in feature space. In this research, to represent a general range, we suggest a $\delta_i = (1-\alpha)\cdot Min + \alpha \cdot Max$, which is an equation between minimum distance and maximum distance. In this research, α value becomes any value between 0.0 and 1.0. However, as we stated before, the preference boundary of minimum distance includes almost centroid only, and that of maximum distance may include outliers. Both cases can not represent the preference of customers, so the preference boundaries of minimum distance and maximum distance exist in theory, but they are useless in reality. Therefore, we tested the SC and NC algorithms by varying α value from 0.2 and 0.8.

D. Range of Preference Boundary

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

$$d(c,a) = \sqrt{\sum_{k=1}^{K} (O_c^k - O_a^k)^2 \Big/ K} \qquad (3)$$

where O_c^k and O_a^k are *k*th feature value of centroid vector O_c and O_c^k, respectively, *k* is the total vector number of features. The similarity between a target customer c and another customer a, *sim(c, a)* is calculated using the Weighted Centroid Euclidean distance function:

$$sim(c,a) = \underset{b\in H}{Max}[d(c,b)] - d(c,a) \Big/ \underset{b\in H}{Max}[d(c,b)] - \underset{b\in H}{Min}[d(c,b)] \qquad (4)$$

Where *b* implies any customer in neighbor set H, and *d(c, a)* is *a* distance function between the target customer *c* and other customer a, and *Max*[*d(c, b)*] and *Min*[*d(c, b)*] denote the maximum and minimum distance between two customers *c* and *b*, respectively.

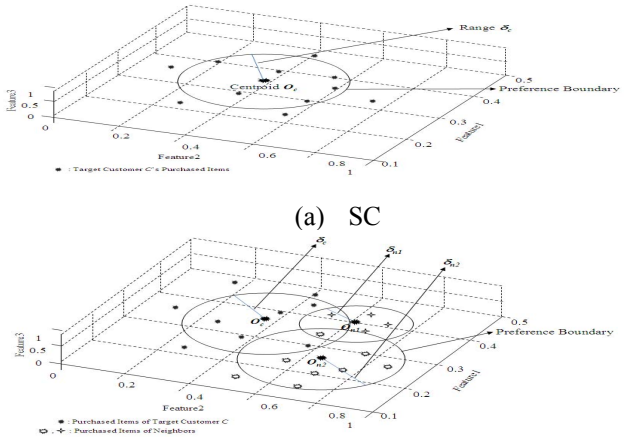
E. Phase 1: Defining Each Customer's Preference Boundary

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 in cluster[7]. The customer is assumed to have similar preference with the customer *c* if the distance is very close.

In this research, two algorithms are suggested to define customer's preference boundary. SC is a method developed from typical CB approach, while NC is based on the concept of neighbors that comes from CF. Thus, NC is thought to be a hybrid method. Fig. 3(a) and Fig. 3(b) show examples of the SC and NC method respectively.

The algorithm SC is to organize the preference boundary of individual customer c based on his/her purchase history only. Preference boundary of a customer c is determined by the centroid O_c and the range δ_c of each feature using c's purchase history. Therefore, the preference boundary of c is represented as $\{O_c - \delta_c, O_c + \delta_c\}$ for any k in feature space. Here, O_c^k and δ_c^k represent the centroid and the range of customer c's kth feature, respectively. So a new item I is recommended to the customer c when

$$O_c - \delta_c \le I^k \le, O_c + \delta_c \ \ for\ all\ k \qquad (5)$$



(a) SC

(b) NC

Figure 3. Preference Boundary Using TC and NC

Here, I^k represents the kth feature value of the new item I. A proper value of δ_c^k is not known exactly, so the optimal value is determined by experiments. The preference boundary of an individual customer in three dimensional feature spaces is shown in Fig. 3(a).

The algorithm NC makes the preference boundary of customer c based on the customer c's and his/her neighbors' purchase history. Therefore, a new item I is recommended to customer c, if

$$O_c - \delta_c \le I^k \le O_c + \delta_c \ or\ O_n - \delta_n \le I^k \le O_n + \delta_n \quad for\ all\ k\ in\ feature\ space\ and\ all\ c's\ neighbors \qquad (6)$$

The preference boundary of customer c and his/her neighbors are shown in Fig. 3(b).

F. 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 developed a CF-based hybrid method, NC, which use neighbors to determine preference boundaries, and tested whether the approaches improve the recommendation quality or not. The result of NC is 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, this research suggests using minimum and maximum distances. These distances are calculated between the centroid of the target customer and all his/her purchased items in feature space. Using the minimum/maximum distance implies very small/large preference area, so in real life both extreme cases may not be used, especially minimum distance is never used. So this research suggests using $(1-\alpha)\cdot Min + \alpha \cdot Max$ as a range of preference boundary, where Min/Max is a minimum/maximum distance, and α is between 0.2 and 0.8.

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 [11]. 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. 4 shows the overall description of experimental data.

B. Measures and Experimental Environment

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

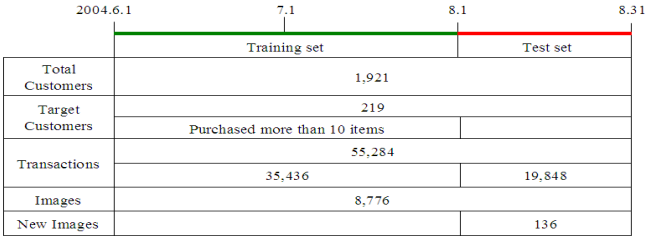


Figure 4. Data Set Design

Recall and Precision have been widely used to test recommendation quality in recommender systems [2, 6, 10]. 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 all items in the real customer 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 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 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 [6, 10]. It is written the following equation:

$$F1 = 2 \times recall \times precision \,/ recall + precision \quad (6)$$

Most of existing recommender systems, first, decide target customer, and then choose items to recommend for him/her. But in this paper, new item is determined first, and then the problem is to choose target customers to recommend the new item. Therefore, many researchers use these metrics for each of target customers and use the average value to evaluate the performance of suggested algorithms[6, 10], but in this research, we calculate precision, recall and F1 for each new item and use the average value to evaluate the performance of algorithms. We compute the values of each metric based on

the item not on the customer.

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. Results and Discussion

Since the quality of CB or CF-based hybrid algorithms varies with the neighborhood's size, we perform an initial experiment to determine the optimal size. Respecting the whole customer set, 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.

To determine the sensitivity of target customer size, M, we performed an experiment where we varied the number of M from 10 to 50 with an increment of 10. Fig. 5 shows our experimental results of SC (using single customer to determine the centroid of preference boundary) at 6 different values of α which is varied from 0.2 (near minimum range) to 0.8 (near maximum rage) with an increment of 0.1. The results show that the size of the target customer does affect the quality of new item recommendations. Our experiment shows that 50 is determined as optimal target customer size for the rest experiments, but the differences are not much.

As looking into Fig. 6, the F1 of the SC is increasing rapidly as increasing the value of α. because the SC is based on distance function between the centroid and customer's purchased items. The preference boundary range increases as increasing the value of α. So SC is effective in wide range of preference.

The preference boundary range of NC is composed of target customer' boundary range and his/her neighbor customers' boundary ranges. If the value of α is higher than 0.3, the total preference boundary range of NC becomes too large. So the F1 becomes worse from that point. However, at the small value of α like 0.2 or 0.3 the F1 value of NC is higher than SC, where CF works better than CB.

In summary, SC is based on target customer's purchased items, so wider preference range makes the F1 value better. Contrast to SC, NC is composed of target customer's purchased items and neighbors' purchased items, so wider preference range of target customer and his/her neighbor makes the F1 value worse.

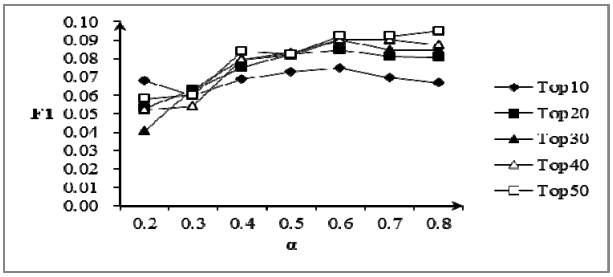


Figure 5. Effectiveness of SC by M Target Customer

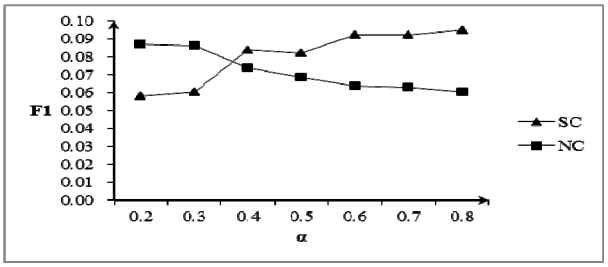


Figure 6. Evaluation of NC and SC

IV. Conclusion

CF is one of the well-known recommendation algorithms, but it is 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, it is not enough. The SC method in this paper is a typical CB-based method. We proposed a hybrid method combining CB and CF-based techniques, NC method, which uses not only a target customer's data, but also his/her neighbors' data when obtaining preference boundary. When determining the centroid, the NC method keeps the target customer's and neighbors' original centroids. Among them, we found that the NC method performed better than the SC in narrow area of preference boundary; but SC performed even better than the NC in wide area of preference boundary.

Accordingly, it is needed to expand the 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 other measure of similarity can be used to derive a more flexible algorithm.

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] Billsus D, Pazzani M "Learning collaborative information filters", In Proceedings of the Fifteenth International Conference on Machine learning, 1998, pp.46-54
[3] Burke R "Hybrid Recommendation Systems: Survey and Experiments", User Modeling and User-Adapted Interaction, Vol 12, no.4, 2002, pp.331-370
[4] 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
[5] 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
[6] Cho Y H, Kim J K "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce", Expert Systems with Applications 26, 2004, pp.233-246
[7] Han J, Kamber M, Data Mining: Concept and Techniques, 2nd ed., Morgan Kaufmann Publishers, 2006.
[8] 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
[9] 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
[10] Lin W, Alvarez S, Ruiz C "Efficient Adaptive-Support Association Rule Mining for Recommendation systems", Data Mining and Knowledge Discovery, 6, 2002, pp.83-105
[11] 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
[12] Sarwar B, Karypis G, Konstan J, Riedl J "Item-based collaborative filtering recommendation algorithms", In Proceedings of the 10th international conference on World Wide Web, 2001, pp.285-295