6D-6

# A New Framework of Hybrid Recommender System using Multi-Clustering

Sutheera Puntheeranurak†
Graduate School of Engineering
Tokai University†

Hidekazu Tsuji‡
School of Information Technology and Electronics
Tokai University‡

## 1. Introduction

In a world where the number of choices can be overwhelming, recommender systems help users find and evaluate items of interest. They connect users with items to consume by associating the content of recommended items or the opinions of other individuals with the consuming user's actions or opinions. These systems use a variety of techniques to help users identify the items that best fit their tastes or needs. Based on the underlying technology, recommender systems can be broadly categorized as content-based filtering or collaborative filtering.

A content-based filtering system recommends items based on their similarity to what a given user has liked in the past. Therefore reliable and automatic determining of the results of content analysis and user preferences is required for this system. A collaborative filtering system recommends items that other users who are similar in preferences to the user in question have liked. The performance of collaborative filtering relies on the amount of available opinions on the considered objects and it therefore performs poorly when few or no opinions are known.

In order to build better performing filtering systems both techniques can be combined, called Hybrid system. However, despite all of these advances, the current generation of recommender system still requires further improvements to make recommendation methods more effective and applicable to an even broader range of real-life applications.

## 2. The Proposed Hybrid Recommender System Framework

Figure 1 illustrates overall framework which consists of two processing: offline mining and online mining. First, all users' rating and user's information are recorded by the WWW server and stored into the rating database and user profile. The user-ratings matrix is a matrix of users versus items, where each cell is the rating given by a user to an item. We will refer to each row of this matrix as a user-ratings vector. The user-ratings matrix is very sparse, since most items have not been rated by most users. We apply clustering algorithm to group the items and the users, and use the result which is represented by the fuzzy set, to create a group-rating matrix. We then compute the similarity of both of group-rating matrixes by using Pearson correlation-based algorithm. At last, the total similarity is the linear combination of the above.

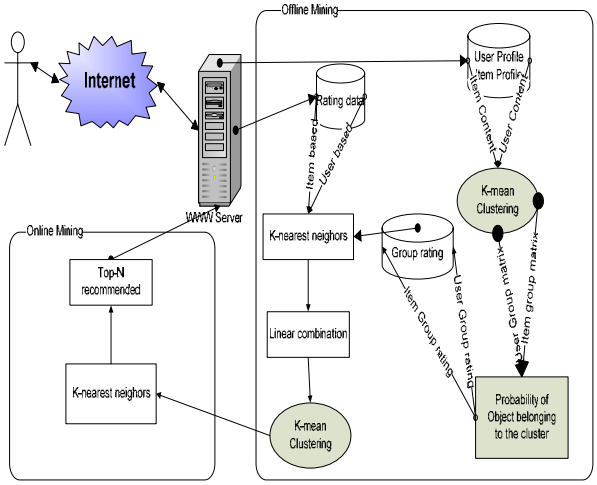


**Figure 1.** The proposed hybrid recommender system framework.

After completing offline mining, we do K-means clustering again to find a group of liked minded users within which a memory-based search is carried out. This group is small in size compared to the original set, thus making the technique scalable.

### 2.1 Group rating matrix

The goal of grouping ratings is to group the items and users into several cliques and provides content-based information for collaborative similarity calculation. Each item has its own attributes, such as movie items, which may have genre, actor, actress, director, published date, and synopsis as its attributes. And each user has its own attributes which may have age, occupation and liked genre, etc. Thus, we can group the items and users based on those attributes.

K-mean Clustering algorithm is a simple and fast clustering method, which has been popular used. So we apply it to group the items and users with some adjustments. The difference is that we apply the fuzzy set theory to represent the affiliation between an item and a cluster.

Firstly, items are grouped into a given number of clusters. After completion of grouping, the probability of one item *j* to be assigned to a certain cluster is calculated as follows.

$$P(j,k) = \frac{Ec(j,k)}{MaxEc\ (j,k)} \quad (1)$$

where *P(j,k)* means the probability of object *j* to be assigned to cluster *k*; The *Ec(j,k)* means the function

to calculate Euclidean distance between an object *j* and cluster *k*; *MaxEc(j,k)* means the maximum Euclidean distance between an object *j* and cluster *k*.

**2.2 Similarity computation**

After grouping the items and users, we get a new item group rating matrix *IG* and a new user group rating matrix *UG*. We can use the item-based collaborative algorithm (IC) to calculate the similarity by using Pearson correlation-based similarity that is the most common measure for calculating the similarity. To make the correlation computation accurate we must first isolate the co-rated cases (i.e., cases where the users rated both *i* and *j*)

$$sim(i,j)=\frac{\sum_{u\in U}\left(R_{u,i}-\overline{R}_i\right)\left(R_{u,j}-\overline{R}_j\right)}{\sqrt{\sum_{u\in U}\left(R_{u,i}-\overline{R}_i\right)^2}\sqrt{\sum_{u\in U}\left(R_{u,j}-\overline{R}_j\right)^2}} \quad (2)$$

where $R_{u,i}$, $R_{u,j}$ is the rating given to item *i and j* by user $u$; $\overline{R}_{u,i}, \overline{R}_{u,j}$ is the mean rating given by user; and $U$ is the total number of items.

After completion of calculating the item-based collaborative similarity, we get $sim_{IG}(i,j)$ the similarity of the new item group rating matrix. Thus we apply to calculate the user-based collaborative algorithm (UC). In UC, clustering is based on the attributes of user profiles and clustering result is treated as items. We apply Equation 2 to calculate $sim_{UG}(i,j)$the similarity of new user group rating matrix, *i and j* mean the user and *u* mean the item, instead the original meaning.

In additional, we use Pearson correlation-based to calculate the similarity from item-rating matrix and user-rating matrix. We call $sim_I(i,j)$ and $sim_U(i,j)$ respectively. At last, the total user similarity is linear combination between $sim_{IG}(i,j)$ and $sim_I(i,j)$ Another is the total item similarity is linear combination between $sim_{UG}(i,j)$ and $sim_U(i,j)$.

$$simI(i,j)=sim_I(i,j)\times(1-c)+sim_{IG}(i,j)\times c$$

where *simI(i,j)* means the similarity between item *i* and *j*; *c* means the combination coefficient; $sim_I(i,j)$ means that the similarity between item *i* and *j*, which is calculated from item-rating matrix; $sim_{IG}(i,j)$ means that the similarity between item *i* and *j*, which is calculated from item group rating matrix. Then make a linear combination of the similarity in user group rating matrix and the user-rating matrix as the above.

**2.3 Collaborative prediction**

The final step of offline mining process is to make a collaborative prediction. Prediction for an item is then computed by performing a weighted average of deviations from the neighbor's mean. The general formula for a prediction on item i of user u as follows:

$$P_{u,i}=\overline{R}_i+\frac{\sum_{k=1}^{n}\left(R_{u,k}-\overline{R}_k\right)\times simU(i,k)}{\sum_{k=1}^{n}\left|simU(i,k)\right|}$$

where $P_{u,i}$ represents the prediction for the user *u* on item *i* ; *n* means the total neighbors of item *i* ; $R_{u,k}$ means the user *u* rating on item *i* ; $\overline{R}_i$ is the average ratings on item *i* ; *simU(i,k)* means the similarity between item *i* and its neighbor *k*; $\overline{R}_k$ means the average ratings on item *k*.

The score that the user *u* rated the item *j* can be predicted as $P_{u,j}$:

$$P_{u,j}=\overline{R}_j+\frac{\sum_{k=1}^{n}\left(R_{u,k}-\overline{R}_k\right)\times simI(j,k)}{\sum_{k=1}^{n}\left|simI(j,k)\right|}$$

**3. Conclusions**

We have proposed a new approach to combine collaborative and content-based filtering techniques. In offline mining, the collaborative sub-system provides two types of clusters, item-based and user-based, to overcome a shortage of ratings. Besides, in our approach, based on the information from user group rating matrix and item group rating matrix, we can make predictions for the new item and new user. Thus we can solve the cold start problem, sparsity, and scalability.

**4. References**

[1] B.M. Sarwar, G. Karypis, J.A. Konstan, and J. Rield, "Item-based Collaborative Filtering Recommendation Algorithms," *In Proc. 10th Int. WWW Conf.*, 2001, pp.285-295.

[2] A. Jameson, J. Konstan, and J. Riedl, *AI techniques for personalized recommendation. Tutorial notes*, AAAI-02, Edmonton, Canada, 2002.

[3] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin," Combining Content-Based and Collaborative Filters in and Online Newspaper," *Proc. ACM SIGIR'99 Workshop Recommender Systems: Algorithms and Evaluation*, Aug. 1999.

[4] J. Han and M. Kamber, *Data mining: Concepts and Techniques*, Morgan-Kaufman, New York, 2001

[5] G. Salton and C. Buckley, "Term-weight approaches in automatic retrieval," *Information Processing and Management*, vol. 24, no. 5, 1988, pp.513-523.