협업 필터링 시스템에서 Degree of Match를 이용한 성능향상

Using Degree of Match to Improve Prediction Quality in Collaborative Filtering Systems

손 재 봉 (Jaebong Sohn) 고려대학교 경영대학 경영학과 석사
서 용 무 (Yongmoo Suh) 고려대학교 경영대학 경영학과 교수

요 약

추천시스템은 사용자들의 관심을 끄는 아이템을 그들이 보다 쉽게 찾도록 도와주거나, 그들의 기호에 기반하여 의미 있는 아이템들을 제공한다. 지금까지 가장 성공적이었던 협업 필터링 기반 추천시스템은 다른 사용자들의 의견을 참조하여 추천을 원하는 사용자에게 추천을 한다. 즉, 아이템들에 대한 사용자 기호를 나타내는 다른 사용자들의 평가정보가 추천을 위한 정보원으로 사용된다. 이처럼, 협업 필터링 기반 추천시스템이 사용자들의 기호만을 이용하도록 설계되었지만, 다른 정보를 이용하면 추천시스템의 성능과 정확도를 높일 수 있을 것으로 사료되어, 본 논문에서는 유사 정도와 인구통계학 정보를 이용한 협업 필터링 기반 추천시스템을 제안한다.

이런 추천시스템에서는 평가정보가 계속적으로 누적되기 때문에, 추천시스템의 정확도를 유지할 수 있는 한, 사용하는 데이터의 양을 줄이는 게 중요하다. 본 논문에서는 유사 정도와 인구통계학 정보를, 사용할 데이터의 양을 줄이기 위한 기준으로 사용하여 자연스레 시스템의 성능을 향상시켰다. 본 논문에서는 실험을 통하여 유사 정도의 사용이 추천시스템의 정확도를 높여주었고, 특정 인구통계학 정보의 사용도 추천시스템의 정확도를 높였음을 보였다.

키워드 : 추천시스템, 협업 필터링, 인구통계학 정보, 스페어만 등급, 상관계수, 유사 정도

I. Introduction

In ACM President's letter on "Electronic Junk" in 1982, Peter Denning, one of the Presidents of ACM, noticed that the visibility of personal computers, individual workstations, and local networks had focused most of their attention on generating information - the process of producing documents and disseminating them. Therefore, we should pay more attention to receiving information - the process of controlling and filtering information that reaches the persons who use it (Peter *et al.*, 1982).

Recently, success of many new information services such as online education service, electronic mail, and the World Wide Web is crucially dependent on the availability of information filtering

technology. From past to present, even though the contents and sorts of information on the web are being changed, the demand for information filtering technology is being increased and is not limited to new information. Recently, we have seen explosive growth of sheer information. The huge volume of available information makes it difficult to find useful information that attracts our interest.

The most frequently-used approaches to solving the problems of *information overload and filtering* have been either content-based or collaborative filtering. Content-based filtering has been used mainly in the context of recommending items such as books, web pages, news, etc. for which informative content descriptors exist. Namely, content-based filtering uses features of items to make recommendation. In contrast, collaborative filtering (CF) system, one of the successful recommendation systems, helps users make choices based on the opinions of other like-minded users (Resnick *et al.*, 1994). CF systems essentially automate the process of "word-of-mouth" recommendations (Shardanand *et al.*, 1995). That is, they exploit correlations between ratings of each of a population of users and ratings of an active user to predict rates of new items (Basilico *et al.*, 2004). Predicting rates for an active user more accurately is a major challenge of CF systems (Basilico *et al.*, 2004). Actually, traditional CF algorithms consider *all* the users who have rated *at least one* same item as the active user. Because of this shortcoming, the performance of CF-based recommendation systems has been degraded and their prediction accuracy needs to be improved.

In this paper, to address the above problems of traditional CF systems, we propose *degree of match* (*DOM*) and demographic information as criteria for reducing the data volume, thereby improving both the accuracy and the performance of CF recommender systems. The basic idea of reducing data volume by *DOM* is that we will not use the ratings of *all* the users who have rated at least one same item as the active user. Instead, we will use the ratings of *selective users* who have rated *many items* that the active user has rated. Also, we will use the ratings of selective users who are similar to the active user in terms of demographic information.

The rest of the paper is organized as follows. We introduce classification and components of recommender systems in Sections Ⅱ and Ⅲ, respectively. Section Ⅳ explains CF-based recommendation systems using DOM and demographic information. Section Ⅴ describes the details of experiments and the results. Finally, the paper ends in Section Ⅵ with summary and conclusion.

Ⅱ. Literature Review: Recommender Systems

Oftentimes, we are at a loss whether to buy something or not. In that case, it would be helpful if someone gives recommendation regarding it. As such, when we come across a situation where we have to make an informed decision on a specific item or an artifact, we want to rely on external knowledge associated with it or on sources having such knowledge in order to make a better decision. Also, there could be many factors which influence a person when making decisions (Basu *et al.*, 1998).

So, recommender systems have come into being in order to help us make a good informed decision, as we do without such systems. That is, using the systems, we can either rely on information provided by others or consider various factors before making a decision. Those systems help users find items in which they have interest more easily given a huge

〈Table 1〉 Comparison of recommender systems (B2002)

	Background data	Input data	Process
Collaborative approach	Ratings from **U** of item in **I**	Ratings from a of items in **i**	Identify users in **U** similar to ***a***, and extrapolate from their ratings of **i**
Content-based approach	Features of items in **I**	a's ratings of items in **I**	Generate a classifier that fits ***a***'s rating behavior and use it on **i**
Demographic approach	Demographic information about **U** and their ratings of items in **I**	Demographic information about **u**	Identify users that are demographically similar to ***a***, and extrapolate from their ratings of **i**

amount of data. Though there have been diverse recommender systems, the most frequently-used approaches taken by the systems are content-based, collaborative filtering or demographic information-based.

These representative approaches are compared in <Table 1>, where I represents the set of items over which recommendations might be made, U the set of users whose preferences are known, *a* is an active user for whom recommendations need to be generated, and *i* an unseen item for which we would like to predict *a*'s preference.

2.1 Content-Based Filtering

Based on the attributes or the features of the items that users already rated, content-based filtering systems try to recommend items similar to those an active user has liked in the past (Balabanovic *et al.*, 1997). A content-based filtering system works as follows: 1) it learns a profile of a user's interests, consisting of features of each item that he or she has rated (Burke, 2002); 2) it analyzes the contents (i.e., values of features) of the profile, given the values of features of a new item; 3) it infers which of the yet unseen items might be interesting to the active user (Yu *et al.*, 2003). Schafer, Konstan and Riedl regard content-based filtering as 'item-to-item correlation' (Burke, 2002). This approach has its roots in the IR (information retrieval) community, and thus employs many of the same techniques as is used in that community (Balabanovic *et al.*, 1997), such as neural networks, decision trees and vector-based representations. Since the content-based recommendation is appropriate to be used especially when there is rich content information on items, it has been used to recommend articles or web pages (Lim *et al.*, 2001).

2.2 Collaborative Filtering

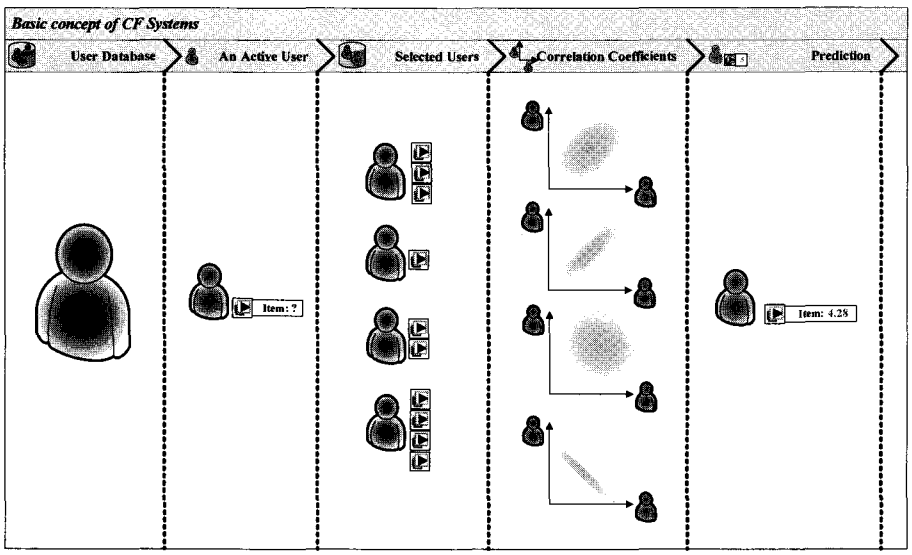
Collaborative filtering (CF) is the most famous and popular in the field of recommender systems. It is also knows as "word-of-mouth" (Shardanand *et al.*, 1995). In CF systems, ratings from other users on items such as movie and music are used to predict an active user's preference on a new item that has not been rated (Sun Lee, 2001). So, CF can be referred to as 'people-to-people correlation'. Two general classes of CF algorithms are used: memory-based and model-based. Memory-based algorithms operate over the entire user database composed of users' ratings on items. On the other hand, model-based algorithms use a part of user database when they produce a model and use another part when they estimate it (John *et al.*, 1998).

The concept of CF, which was coined by Goldberg *et al.* (Goldberg *et al.*, 1992), was descended from the study in the area of information filtering (Goldberg *et al.*, 1992). *Resnick* defined CF recommender systems as follows: it helped people make choices based on the opinions of other people (Sarwar *et al.*, 2000). CF utilizes user database which consists users' ratings on each item and predicts whether an active user will like or dislike (Khors *et al.*, 1999). There are many CF-based systems. *Tapestry* implemented a novel mechanism for CF in which users annotate documents before the documents are filtered but it was not automated (Goldberg *et al.*, 1992). *Ringo* was a social information filtering system for personalized music recommendations (Shardanand *et al.*, 1995). The first automated CF recommender system was introduced by *GroupLens* research project (Resnick *et al.*, 1994). *GroupLens* provided personalized predictions for *Usenet* news articles. It made use of Pearson correlation coefficient.

Basically, predictions in CF systems are made as is explained in <Figure 1>. First, when an active user needs a prediction on an unseen item, CF systems find users who already rated the item from the user database. Second, correlation coefficient between the active user and each one of the selected users is calculated using methods such as Pearson correlation coefficient, or Spearman correlation coefficient. Once the correlation coefficients are produced, recommendation on the unseen item is proposed for the active user using prediction formulas, explained later.

2.3 Demographic Information-Based Filtering

Demographic recommendations first segment the users based on personal features and then make recommendation based on the information in segmentations. The benefit of demographic recommendations is that it may not require a history of user ratings needed by CF systems (Burke, 2002). Other than that, demographic recommendation systems are



〈Figure 1〉 Basic Concept of Collaborative Filtering Systems

similar to CF systems. However, they analyze similarities between users based on their demographic data, such as age, gender, and occupation (Krulwich *et al.*, 1997). One of the most crucial advantages of the demographic recommendation is the ability to recommend crossing genres or beyond the boundary of specific artifacts. For example, it is possible that listeners who enjoy cool jazz also enjoy classical music, but a content-based recommender trained on the preferences of a cool jazz would not be able to recommend classical music since none of the features such as performers, instruments, repertoire associated with cool jazz would be shared in the different categories (Burke, 2002).

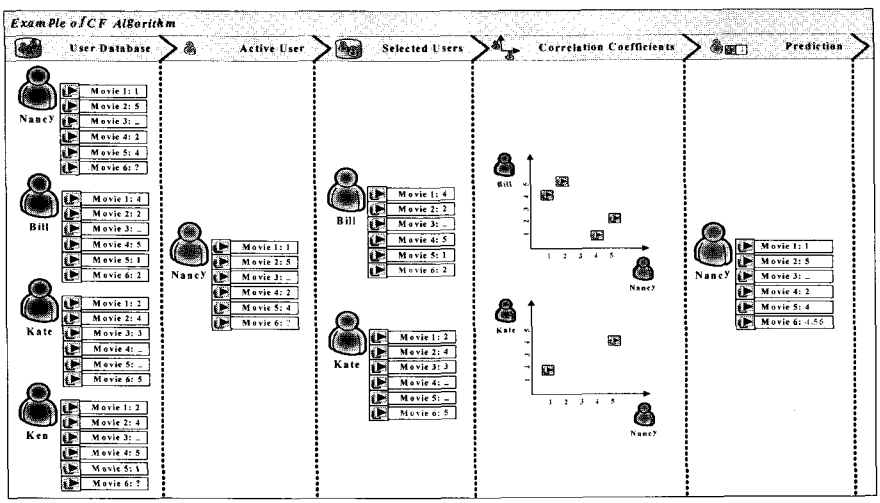
Ⅲ. Major Components of CF Systems

The purpose of CF systems is to predict the utility of an unseen item to an active user based on other users' opinions. Components that are needed by the CF systems are those such as *user database*, a method to calculate *correlation coefficient* between users, and *prediction formula*.

We will explain briefly how CF systems works, using a simplified case, as is depicted in <Figure 2>. Suppose Nancy is an active user who wants to know whether she herself will like movie 6 or not and we have three other users in our database, Bill, Kate and Ken. CF systems work as follows. First, users, Bill and Kate, who rated movie 6 are selected from the user database. Then, correlations between Nancy and Bill and between Nancy and Kate are calculated using correlation coefficients such as Pearson correlation coefficient. Finally, prediction (i.e., 4.56 in our case) on movie 6 is calculated, using a prediction formula which is explained later. Now each component of CF systems will be explained in detail.

3.1 User Database

User database contains each user's preferences on specific items. User's preferences are represented as ratings and are the only source for CF systems to recommend. Therefore, the user database can be seen as a matrix ($v_{i,j}$), where $v_{i,j}$ corresponds to the user i's vote (i.e., rate) on item j.



〈Figure 2〉 Example of Traditional CF Systems

3.2 Correlation Coefficients

Simply, prediction in CF systems is made in two steps. First, correlation coefficient between an active user and a selected user needs to be computed. Second, the ratings of users need to be combined and transformed into a predicted rating for an active user using a prediction formula (Bergholz, 2003). In the following, we explain the various methods to calculate correlation coefficients between users.

3.2.1 *Pearson* Correlation Coefficient

The first CF system introduced by *GroupLens* Research Project (Resnick *et al.*, 1994) uses *Pearson* formula to calculate the correlation coefficient between users. *Pearson* correlation coefficient, ranging from -1 to 1, indicates how much an active user tends to agree to other user on items that they both rated. The *Pearson* formula is defined below in equation (1):

$$R_{a,u} = \frac{\sum_{i=1}^{m}(r_{a,i} - \bar{r_a})(r_{u,i} - \bar{r_u})}{\sqrt{\sum_{i=1}^{m}(r_{a,i} - \bar{r_a})^2 \sum_{i=1}^{m}(r_{u,i} - \bar{r_u})^2}} \quad (1)$$

, where $\bar{r}_{a,i}$ is the rating given to item i by user a; and $\bar{r}_a$ is the mean rating given by user a and is calculated as in equation (2); and m is the total number of items (Rojsattarat *et al.*, 2003). If Ii is the set of items on which user i has voted, then we can define the mean vote for user i as in equation (2) (John *et al.*, 1998):

$$\bar{v}_i = \frac{1}{|I_i|}\sum_{j \in I_i} v_{i,j} \quad (2)$$

Variations of *Pearson* correlation coefficients are also used in CF systems. Pearson 2.5 (P25) defined in equation (3) replaces $\bar{r}_a$ and $\bar{r}_u$ in the Pearson formulas with the average possible rating (Bergholz, 2003). The average possible rating in *MovieLens* dataset is 2.5 because users' opinions were rated using a 5 point *Likert* scale (1="very bad movie" to 5="very good movie").

$$R_{a,u} = \frac{\sum_{i=1}^{m}(r_{a,i} - 2.5)(r_{u,i} - 2.5)}{\sqrt{\sum_{i=1}^{m}(r_{a,i} - 2.5)^2 \sum_{i=1}^{m}(r_{u,i} - 2.5)^2}} \quad (3)$$

However, the *Pearson* correlation coefficient is derived from a linear regression model that relies on a set of assumptions regarding the data, namely that the relationship between two users' ratings must be linear, and the errors must be independent and have a probability distribution with mean 0 and constant variance for every setting of the independent variable. When these assumptions are not satisfied, *Pearson* correlation coefficient becomes a much less accurate indicator of similarity. It is common for these model assumptions to be violated in CF data (Jonathan *et al.*, 1999). Because Pearson correlation coefficient uses the actual values of the observations, it is much affected by outliers. On the other hand, *Spearman* correlation coefficient which is explained in the following section uses only the ranks of the observations. For that reason, the coefficient is less affected by outliers than Pearson correlation coefficient. In the case of *MovieLens* dataset, *Pearson* correlation coefficient was not suitable for calculating similarity between an active user and a selected user.

3.2.2 *Spearman Rank-Order Correlation Coefficient*

Spearman rank-order correlation coefficient is similar to *Pearson* correlation coefficient but it is free from model assumptions, computing a measure of correlation between ranks instead of between rating values. When there are no ties of ranks in the dataset, the *Spearman* correlation coefficient is de-

fined as in equation (4):

$$Spearman\ \rho = 1 - \frac{6}{n^3 - n}(\Sigma d_i^2) \quad (4)$$

, where d_i is the rank difference of item i between two users and n is the number of items. Because the ratings of *MovieLens* dataset were ranged from 1 to 5, the ties among items occurred frequently.

3.2.3 Modified *Spearman* Rank-Order Correlation Coefficient

Jonathan's study showed that *Spearman* rank-order correlation performs similarly to *Pearson* correlation coefficient. However, he noticed the large number of tied rankings results in a degradation of the accuracy of *Spearman* correlations coefficient (Jonathan *et al.*, 1999). According to *Callan*'s study, correlation coefficient that ignores the effects of ties can give misleading result (Callan *et al.*, 1999). For that reason, *Callan* (Callan *et al.*, 2001) proposed *Spearman* rank-order correlation coefficient considering both complete ordering and effect of ties of rank. Prior to computing the *Spearman* rank-order correlation coefficient, the original ratings are first transformed into ranks. If the rank correlation coefficient is 1, 0 or -1, it means that the ranks between two users are identical, unrelated, or in reverse order, respectively. The exact correlation coefficient is computed as in equation (5):

$$\rho = \frac{1 - \frac{6}{n^3 - n}(\Sigma d_i^2 + \frac{1}{12}\Sigma(f_k^3 - f_k) + \frac{1}{12}\Sigma(g_m^3 - g_m))}{\sqrt{(1 - \frac{\Sigma(f_k^3 - f_k)}{n^3 - n})} \times \sqrt{(1 - \frac{\Sigma(g_m^3 - g_m)}{n^3 - n})}} \quad (5)$$

, where f_k is the number of ties in the k-th group given by a user, g_m is the number of ties in the m-th group given by another user and d_i is the rank difference of item i, n is the number of items.

3.3 Prediction Formulas

Having calculated the correlation coefficients, CF systems made a prediction for an active user by using the equation (6) given below.

$$Ui_{pred} = \bar{U} + \frac{\sum_{J \in raters} (J_i - \bar{J}) r_{UJ}}{\sum_J |r_{UJ}|} \quad (6)$$

, where U is an active user, i is an unseen item, J is a user who rated item i, and J_i is the J's rating of item i and $\bar{J}$ is the mean voting of J. (Resnick *et al.*, 1994).

The combination of both *Pearson* and Significance Weighting is also used in a variation of the *Pearson* method. The idea is that user correlation based on a high number of commonly rated items should be more important than that based on a lower number. Four or more commonly rated items count as significance 1, three as 0.75, two as 0.5 and one as 0.25 in (Bergholz, 2003). The prediction formula given below is adapted in (Bergholz, 2003):

$$P_{a,j} = \bar{v}_a + \frac{\sum_{i=1}^{n} s_{a,i}\, r(a,i)(v_{i,j} - \bar{v}_i)}{\sum_{i=1} s_{a,i}\, r(a,i)} \quad (7)$$

, where $s_{a,i}$ denotes the significance of the correlation between users a and i, $v_{i,j}$ indicates user i's rating on item j, and $\bar{v}$ is the mean rating, and $r(a, i)$ represents correlation coefficient between user a and i.

Ⅳ. Collaborative Filtering using Selected Dataset

Because World Wide Web technologies are growing so fast, lots of data could be collected and stored in the type of databases or files. If we use all the

available data when we make predictions based on such data, the prediction will take too much time and the result of the prediction may not be exact.

That is, the prediction will be useless. Therefore, in order for the recommender system to have practical value, it is important to use the available data selectively. Here, we propose two ways of reducing the size of the dataset. One is to use a new concept of *degree of match* and the other is to use some of demographic information which shows positive correlation with the domain item.

4.1 Dataset Selection using *Degree of Match (DOM)*

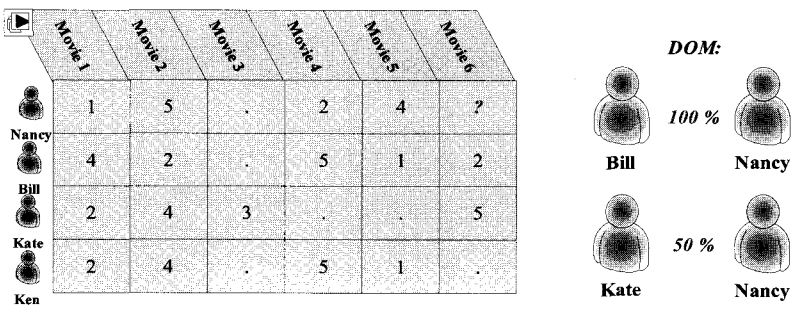
When correlation coefficient is calculated between two users, all users who rated the item whose rating is to be predicted for an active user are utilized in the process of prediction. In other words, if users have rated at least three items, including the unseen item among all the items, prediction formula could provide recommendation for an active user. What if there are so many such users who rated the item whose rating is to be predicted? It is natural to think of selecting more reliable users out of those users based on some criterion.

So, as such a criterion we propose a new concept, *degree of match* (DOM) between an active user and a selected user. DOM is defined to be the percentage of items rated by a selected user among items rated by the active user. That is, DOM can be used to select users among the users who have rated the item whose rating is to be predicted. Our assumption is that the higher the DOM between an active user and a user is, the more accurate the prediction based on the user's opinion will be.

In CF systems, ratings on unseen items are computed using the ratings of others who rated the unseen item. <Figure 3> explains the DOM concept. Nancy who is an active user and wants to know how much she will like movie 6 can make use of the opinions of Bill and Kate just because they rated item 6. In this case, DOM between Nancy and Bill is 100% and DOM between Nancy and Kate is 50%. Prediction made based on Bill's opinions would be better than that made based Kate's opinion. If we set DOM to be a high value, a much less number of users will be selected and the prediction based on the information of those users will be more accurate.

4.2 Dataset Selection using Demographic Information

As is mentioned earlier, another way of reducing the size of the dataset is to use some of demographic



〈Figure 3〉 Example of *Degree of Match* (DOM)

information which shows positive correlation with the item of interest, say, movie in our case.

Therefore, we first need to find such demographic information. Suppose occupation is known to be positively correlated with movie. Then, only those users whose occupations are the same as the active user are selected from the original dataset. Dataset thus reduced will be much less in size, and therefore the performance of the recommender system using the reduced dataset will be improved a lot.

V. Experiments and Results

We conducted an experiment, where we implemented a new CF algorithm which differs from the traditional CF algorithm in two aspects: 1) Our CF algorithm adopts modified Spearman correlation coefficient instead of the Pearson correlation coefficient; 2) We reduced the dataset size based on the concept of DOM and based on demographic information. All functions related to CF algorithm were written in Structured Query Language (SQL).

5.1 Data Preprocessing

MovieLens dataset collected by the *GroupLens* Research Project was used for the experiment. The dataset contained 1,000,209 anonymous ratings of 3,706 movies rated by 6,040 *MovieLens* users (www.movielens.org). 10,000 users were selected for prediction at random without replacement. *MovieLens* dataset provides three files which are movie.dat, users.dat and ratings.dat in the form of text file. The file movie.dat contains MovieID, Title and Genre of movie. The attributes of users.dat are UserID, Zip-code, Occupation and demographic information such as age and gender. And the last file, ratings.dat, consists of UserID, MovieID, Rating and Timestamp. We designed a relational model for transferring the above three flat files to a relational database. All functions needed to make prediction were implemented in commercial relational database system, *SQL Server 2000.*

5.2 Experiment Design

We implemented CF algorithm that uses either *degree of match* or demographic information. And we adopted the *Spearman* rank-order correlation coefficient which considers both complete ordering and the effect of ties in ranks, instead of using either *Pearson* correlation coefficient or pure *Spearman* correlation coefficient.

In the experiments, we adopted modified CF algorithm which contains the processes of selecting users based on both *DOM* and demographic information. The flow divides CF algorithm into three steps as follows.

1) Selecting an active user: In user database, an active user is selected among users.
2) Calculating correlation coefficients: Users whose ratings are to be used in making a prediction are selected based on both *DOM* and demographic information. Then, correlation coefficient between an active user and each user among the selected users is calculated using modified *Spearman* correlation coefficient.
3) Recommendation: recommendation on a new item for an active user is proposed using prediction formula based on correlation coefficients calculated in step 2.

The purpose of the experiment is to verify the improvement of CF system in prediction accuracy and performance by adopting *degree of match* and demographic information. Since reducing the scope

of information by *degree of match* and demographic information results in the less computation time, it is natural that the performance of the CF algorithm will be better. So our concern in this experiment can be summarized in three hypotheses as follows:

H1: Prediction based on the ratings of users with high DOM will be more accurate than that based on ratings of users with low DOM

H2: Prediction based on the ratings of users selected using demographic information will be more accurate than other wise.

H3: Combination of demographic information and degree of match provides better prediction than otherwise.

5.3 Results of Experiments

We compare the performance of our approaches to pure CF systems. To measure performance of our approach, we use Mean Average Error (MAE) to evaluate the accuracy of CF by comparing the prediction values against actual user ratings for the items (Good *et al.*, 1999).

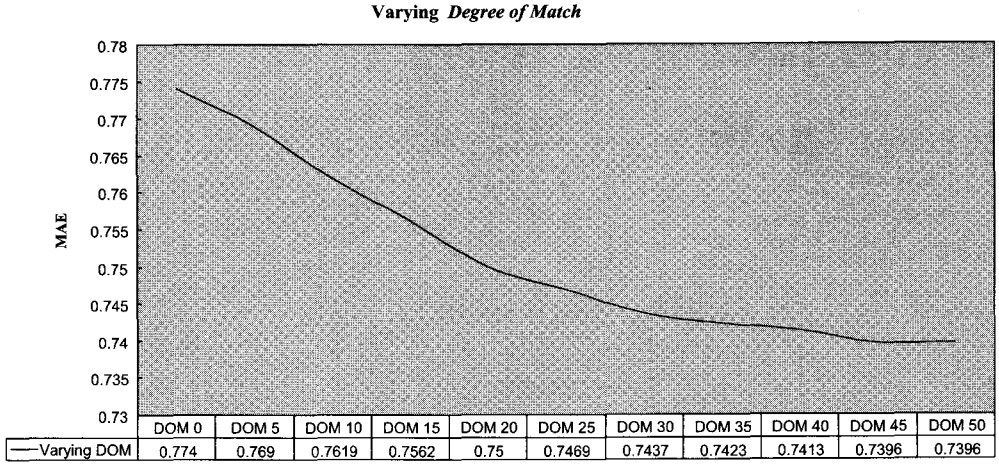
5.3.1 Test Result of Hypothesis H1

<Figure 4> tells us that the higher *DOM* between users is, the lower MAE is produced, as we expected. The numerical difference in MAE between *DOM* = 0 and *DOM* = 50 was 0.034. And also we could confirm that the MAE was decreased gradually as the *DOM* increases. Therefore, hypothesis H1 can be accepted.

5.3.2 Test Result of Hypothesis H2

<Figure 5> shows that age or gender can be used to select users whose ratings are to be used in making predictions. However, it also shows that occupation is not good demographic information that can use used to select users when making predictions.

Therefore, we can conclude regarding hypothesis H2 that *some* of the demographic information can



〈Figure 4〉 MAE based on Varying Degree of Match (DOM)[1)]

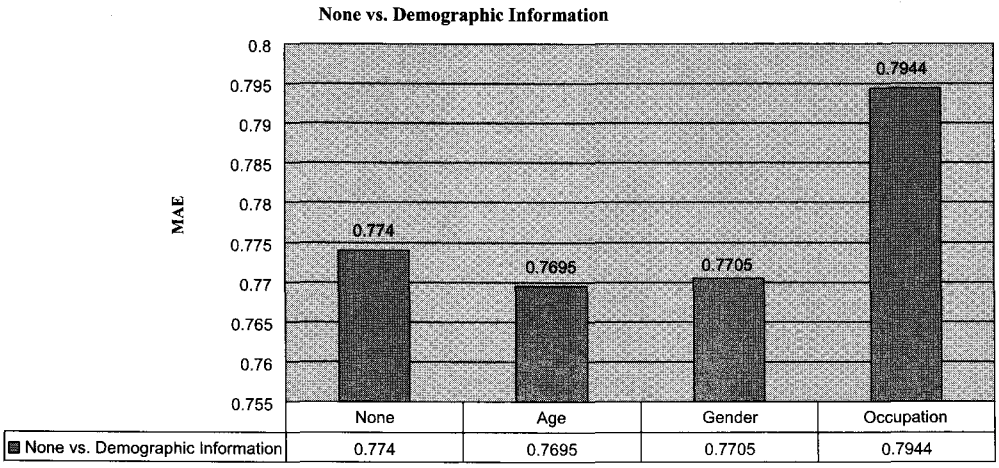
1) The results of paired-samples *t* test identify significance at 5% level between *DOM 0* and from *DOM 5* to *DOM 50*.

be used to reduce the dataset size for better prediction. The kind of demographic information that can be utilized when making a prediction seems to be different for different domain of items.

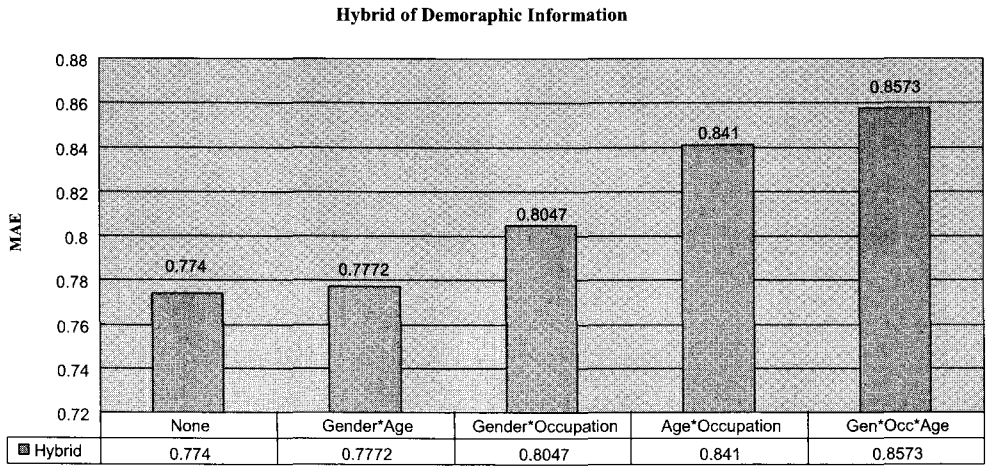
We also compared the MAE difference between when no demographic information is used and when *combination* of demographic information is used. As we can see in <Figure 6>, when more than two kinds of demographic information are used to select users, the result proves to be worse than otherwise. Hypothesis H2 can thus be *partially accepted* according to the experimental results.

5.3.3 Test Result of Hypothesis H3

In <Figure 7>, we can see that when demographic information is used to select users whose ratings are

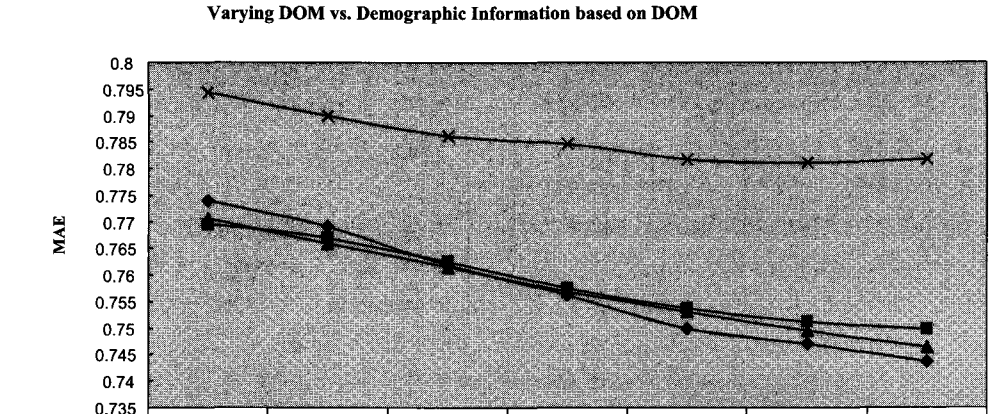


〈Figure 5〉 MAE according to Demographic Information[2)]



〈Figure 6〉 None vs. Hybrid of Demographic Information

2) The results of paired-samples *t* test identify signiicance at 10% level between *None* and *Age* and between *None* and *Gender*.



	DOM 0	DOM 5	DOM 10	DOM 15	DOM 20	DOM 25	DOM 30
Varying	0.774	0.769	0.7619	0.7562	0.75	0.7469	0.7437
Age	0.7695	0.7669	0.7626	0.7576	0.7538	0.7512	0.7497
Gender	0.7705	0.7659	0.7615	0.7568	0.7531	0.7496	0.7464
Occupation	0.7944	0.7898	0.7861	0.7846	0.7818	0.781	0.7818

〈Figure 7〉 MAE Comparison between DOM and Combination of DOM and Demographic Information

to be used to make predictions MAE decreases by degrees as *DOM* increases as in the case of hypothesis H1. (The MAE difference between *DOM* 0 and DOM 30 based on gender was 0.0241.) However, a close look at the figure reveals that 1) MAE obtained after occupation is used to select users is bigger than MAE obtained without using demographic information; 2) MAE obtained after age or gender is used to select users is smaller than MAE obtained without using demographic information until *DOM* is 10; 3) When the *DOM* exceeds 10, combination of *DOM* and demographic information results in worse prediction accuracy. Therefore, Hypothesis H3 is not accepted.

VI. Conclusions and Further Works

Problems with the CF-based recommendation systems include data sparsity and first-rater problems. First-rater problem seems to have no solutions, and data sparsity problem is no longer a problem because a large amount of data is available nowadays. Instead, it would be better to think of how we can improve the performance of the system as well as the prediction accuracy, by reducing the size of the dataset to be used since it is another problem of CF-based recommender systems that it takes long time to make a recommendation because of the too much available information.

In the study, in order to select more useful information out of too much available information, we proposed a new concept of degree of match (*DOM*) and the use of demographic information such as age, gender and occupation. The size of the user database could be reduced a lot by selecting users based on the *DOM* concept and on the demographic information. In addition, the user dataset thus obtained has more useful than the whole set of users, which means the prediction made from the reduced dataset of users could be more accurate than that from the original dataset. So, we made three hypotheses about

the prediction and tested them by conducting experiments. Conclusions we have drawn from the experiments are:

1) Prediction based on the ratings of users with high *DOM* will be more accurate than that based on ratings of users with low *DOM*. By adjusting the *DOM* value to an appropriate value we can get the right size of data set, which could be dependent on the domain of items.
2) Prediction based on the ratings of users selected using *some* demographic information such as age or gender in our domain of movie will be more accurate than other wise.
3) Use of both demographic information and *DOM* provides worse prediction than otherwise, to the contrary of our expectation.

In the experiments, we adopted *Spearman* rank correlation coefficient instead of Pearson correlation coefficient. *Spearman* correlation coefficient is the right choice because the assumptions for the use of *Pearson* correlation coefficient are not satisfied by our dataset.

We have used MAE to compare the traditional CF-based recommendation system with our recommendation system which use the reduced data set based on *DOM* and demographic information. We may have to examine whether such a difference in MAE is statistically effective. Also, other metric may be investigated to compare them, such as the number of instances showing decreased accuracy or increased accuracy as *DOM* increases.

References

Burke, R., "Hybrid Recommender Systems: Survey and Experiments", *User Modeling and User-Adapted Interaction*, Vol.12, 2002, pp. 331-370.

Bergholz, A., *Coping with Sparsity in a Recommender System*, Springer-Verlag Berlin Heidelberg, 2003.

Basilico, J. and T. Hofmann, "Unifying Collaborative and Content-Based Filtering," Proceedings of the 21th international Conference on Machine Learning, 2004, p. 9.

Basu, C., H. Hirsh, and W. Cohen, "Recommendation as classification: Using social and content-based information in recommendation", Preceedings of the International Conference on Artificial Intelligence, 1998.

Balabanovic, M. and Y. Shoham, "Fab: Content-based, collaborative recommendation", *Communications of the ACM*, Vol.40, 1997, pp. 66-72.

Callan, J. and M. Connell, "Query-Based Sampling of Text Databases", *ACM Transactions on Information Systems*, Vol.19, 2001, pp. 97-130.

Callan, J., M. Connell, and A. Du, "Automatic discovery of language models for text databases", Proceedings of the 1999 ACM International Conference on Management of Data, 1999, pp. 479-490.

Goldberg, D., D. Nichols, Brian M. Oki, and D. Terry, "Using collaborative filtering to weave an information Tapestry", *Communications of the ACM*, Vol.35, 1992, pp. 61-71.

Good, N., J. Schafer, A. J. Konstan, A. Borchers, B. Sarwar, J. Herlocker, and J. Riedl, "Combining Collaborative Filtering with Personal Agents for Better Recommendations", Proceedings of the American Association for Artificial Intelligence, 1999.

John, S. B., D. Heckerman, C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", Technical Report, MSR-TR 98-12, Microsoft Research, Microsoft Corpora-

tion, 1998.

Jonathan, L. H., Joseph A. Konstan, A. Borchers, J. Riedl, "An algorithmic framework for performing collaborative filtering," Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, Berkeley, California, United States, 1999, pp. 230-237.

Khors, A. and B. Merialdo, "Clustering for Collaborative Filtering Applications", Proceedings of CIMCA 1999. IOS Press, 1999.

Krulwich, B., "Lifestyle Finder: Intelligent user profiling using large-scale demographic data", *Artificial Intelligence Magazine*, Vol.18, No.2, 1997.

Lim, M. and J. Kim, "An Adaptive Recommendation System with a Coordinator Agent", In Web Intelligence: Research and Development, LNAI 2198, 2001.

Peter, J. D., "ACM president's letter: electronic junk", *Communication of the ACM*, Vol.25, 1982, pp. 163-165.

Rojsattarat, E. and N. Soonthornphisaj, "Hybrid Recommendation: Combining Content-Based Prediction and Collaborative Filtering", Proceedings of the Intelligent Data Engineering and Automated Learning, 4th International Conference, IDEAL 2003, Hong Kong, China, March 21-23, 2003, pp. 337-344.

Resnick, P., N. Lacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews", Proceedings of the 1994 Computer Supported Collaborative Work Conference, 1994.

Shardanand, U., "Social information filtering: Algorithms for automating "Word of Mouth", Proceedings of Human Factors in Computing Systems ACM CHI, 1995, pp. 210-217.

Sun Lee, W., "Collaborative Learning for Recommender System", Proceedings of the 18th international Conference on Machine Learning, 2001, pp. 314-321.

Salton, G. and M. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

Sarwar, B., G. Karypis, J. Konstan, and J. Riedl, "Analysis of Recommendation Algorithms for E-Commerce", Proceedings of the 2nd ACM conference on Electronic Commerce, 2000, pp.158-167.

Yu, K., A. Schwaighofer, V. Tresp, W.-Y. Ma, H.J. Zhang, "Collaborative Ensemble Learning: Combining Collaborative and Content-Based Information Filtering via Hierarchical Bayes," Proceedings of UAI 2003, Morgan Kaufman, 2003.

Information Systems Review
Volume 8 Number 2
August 2006

Using Degree of Match to Improve Prediction Quality in Collaborative Filtering Systems

Jaebong Sohn* · Yongmoo Suh*

Abstract

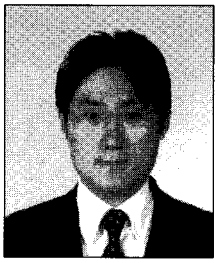
Recommender systems help users find their interesting items more easily or provide users with meaningful items based on their preferences. Collaborative filtering (CF) recommender systems, the most successful recommender system, use opinions of users to recommend for an active user who needs recommendation. That is, ratings which users have voted on items to indicate preference on them are the source for making recommendation. Although CF systems are designed only to use users' preferences as the source of recommendation, use of some available information is believed to increase both the performance and the accuracy of CF systems.

In this paper, we propose a CF recommender system which utilizes both degree of match and demographic information (e.g., occupation, gender, age) to increase the performance and the accuracy. Since more and more information is accumulated in CF systems, it is important to reduce the data volume while maintaining the same or the higher level of accuracy. We used both degree of match and demographic information as criteria for reducing the data volume, thereby naturally enhancing the performance. It is shown that using degree of match improves the prediction accuracy too in CF systems and also that using *some* demographic information also results in better accuracy.

Keywords: Recommender System, Collaborative Filtering, Demographic Information, Spearman Rank Correlation Coefficient, Degree of Match

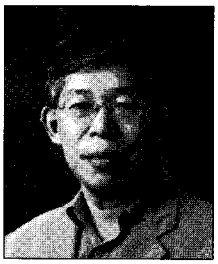
* Department of Business, College of Business Administration, Korea University

◐ 저 자 소 개 ◑



손 재 봉 (soulway@hanmail.net)

건국대학교 경영학과에서 경영정보학을 전공하고 고려대학교 일반대학원 경영학과에서 경영학 석사학위를 취득하였다. 현재 SPSS Korea에서 Consultant로 재직 중이며 주요 관심분야는 database, data mining, recommender system 등이다.



서 용 무 (ymsuh@korea.ac.kr)

서울대학교 사범대학 수학과, 한국과학원 전산학과를 졸업하고, 한국과학기술연구소 전산센터에서 연구원으로 재직 시 도미하여, University of Texas (at Austin)에서 전산학석사, 경영정보학박사를 취득한 후, 세종대학교, 건국대학교를 거쳐, 현재 고려대학교 경영대학에 재직하고 있다. 주요 관심분야는 collaboration technology, ontology, data mining 등이다.

논문접수일 : 2006년 1월 18일 **게재확정일 : 2006년 5월 15일**
1차 수정일 : 2006년 4월 11일