Journal of the Korean Data & Information Science Society
2010, **21**(1), 185–194

한국데이터정보과학회지

The research of new algorithm to improve prediction accuracy of recommender system in electronic commerce[†]

Sun Ok Kim[1]

[1]School of Information Communication & Broadcasting Engineering, Halla University

Received 21 November 2009, revised 15 January 2010, accepted 19 January 2010

Abstract

In recommender systems which are used widely at e-commerce, collaborative filtering needs the information of user-ratings and neighbor user-ratings. These are an important value for recommendation in recommender systems. We investigate the information of rating in NBCFA (neighbor Based Collaborative Filtering Algorithm), we suggest new algorithm that improve prediction accuracy of recommender system. After we analyze relations between two variable and Error Value (EV), we suggest new algorithm and apply it to fitted line. This fitted line uses Least Squares Method (LSM) in Exploratory Data Analysis (EDA). To compute the prediction value of new algorithm, the fitted line is applied to experimental data with fitted function. In order to confirm prediction accuracy of new algorithm, we applied new algorithm to increased sparsity data and total data. As a result of study, the prediction accuracy of recommender system in the new algorithm was more improved than current algorithm.

Keywords: Collaborative filtering, EDA, EV, LSM, NBCFA, recommender system.

1. Introduction

The Internet users come to get a various information on the Internet. It takes long time to get necessary information on the Internet. Hence, the Internet users need the method to get a suitable information. In recommender systems, the Internet users are able to get a suitable information. Also, recommender systems become important tool on the Internet and has been continuously studied to improve.

The commercial and successful algorithm of recommender systems is Neighbor Based Collaborative Filtering (NBCF). This filtering method provide suitable information to user in accordance with user's neighbor who has similarity. The user-ratings of user's neighbor has relationship with performance of recommender systems. It is studied continuously.

[†] This work was supported by the Korea Research Foundation Grant funded by the Korean Government (KRF-2008-531-D00025)

[1] Professor, School of Information Communication & Broadcasting Engineering, Halla University, Wonju, Gangwon 220-712, Korea. E-mail: sokim@halla.ac.kr

Kim *et al.* (2008) use item information that user prefer at NBCF, and defines a number of data sparsity of item, and analyze user-ratings preference of group user's neighbor. In accordance with comparison of prediction value by algorithm, he has studied to improve prediction accuracy of user-ratings.

Lee (2006) has studied to select new neighbor in accordance with preference item and information of user's neighbor. The recommender systems improved by this selection.

Konstan *et al.* (1997) suggest 50 co-rated items as weighting factor. If two users have less than 50 co-rated items, They multiply their correlation by a factor n/50, where n is the number of co-rated items. If the number of overlapping items is greater than 50, then They leave the correlation unchanged. They found that the accuracy of recommender systems improved.

Lee *et al.* (2006) suggest 150 co-rated items as similarity weighting factor. They have studied that the prediction accuracy improved only if weighting factor of similarities is n/150.

Melville *et al.* (2002) have created user matrix with scarcity data in user-rating by Contents Based Collaborative Filtering. They have proved that the accuracy of recommender systems improved with NBCF.

This paper is organized as following sections. Section 2 introduce recommender systems. Section 3 analyze two variables to suggest the new algorithm and study relations between two variable and Error Value (EV). Section 4 define new algorithm and compute MAE (Mean Absolute Error) by applying it to data. Finally, Section 5 discusses conclusion.

2. Recommend algorithm

2.1. NBCF (Neighbor Based Collaborative Filtering) algorithm

The general algorithm of Collaborative Filtering that is used in recommendation with neighbor information is NBCF (Kim, 2008; Kim *et al.*, 2009; Pazzani *et al.*, 1999). This method is applied to recommend movie at GroupLens, University of Minnesota. The neighbor-based collaborative filtering should select neighbors of target user to recommendation. There are continuously studying of various method to select neighbor.

Kim *et al.* (2007) studied performance improvement of recommender systems with clustering. Lee *et al.* (2007) contributes to improve recommender systems using neighbor selection method with demographics data. This paper select user that gave user-ratings in recommend item of target user as neighbors.

We apply preference relations of user-ratings of selected neighbor and user-ratings to recommender systems. This preference relations use interrelation of target user and neighbors.

Similarity between users is measured as the Pearson correlation between their ratings, defined below (Pazzani, 1999; Resnick *et al.*, 1994).

$$c_{u,j} = \frac{\sum_{x=1}^{m}(r_{u,x} - \bar{r_u})(r_{j,x} - \bar{r_j})}{\sqrt{\sum_{x=1}^{m}(r_{u,x} - \bar{r_u})^2}\sqrt{\sum_{x=1}^{m}(r_{j,x} - \bar{r_u})^2}}. \tag{2.1}$$

Where $c_{u,j}$ is the value of preference relations between target user u and user's neighbor j ; $r_{u,x}$ is rating given to item x by target user u ; $r_{j,x}$ is rating given to of item x by user's neighbor j ; $\bar{r}_u$ is the mean rating given by target user u ; $\bar{r}_j$ is the mean rating given by neighbor j user of target user u and m is the total number of items. x is item.

The prediction value of item use target user's mean, neighbors' user-ratings, neighbors' mean and a weighted combination of the selected neighbor's and target user's ratings.

We compute a prediction value using of NBCFA, defined below (Kanstan *et al.*, 1997; Resnick *et al.*, 1994).

$$p_{u,x} = \bar{r_u} + \frac{\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}}{\sum_{j=1}^{k}|c_{u,j}|}. \tag{2.2}$$

Where $p_{u,x}$ is the prediction value for the target user u for item x ; j is the number of users in the neighbor; $\bar{r}_u$ is the mean rating given by target user u ; r_j is rating given to of item x by user's neighbor j ; $\bar{r}_j$ is the mean rating given by neighbor j of target user u and $c_{u,j}$ is the similarity between users, u and j, is computed using the Pearson correlation coefficient.

2.2. Performance measurement of NBCFA

The recommender systems use the MAE to evaluate prediction accuracy of user-ratings. Then user-ratings that target user evaluated and the calculated by NBCF are used to calculate the MAE.

Below equation is computed to get prediction accuracy of user-ratings (Kim, 2008; Kim *et al.*, 2009; Kim *et al.*, 2008; Lee *et al.*, 2007).

$$MAE = |r_{u,x} - p_{u,x}|. \tag{2.3}$$

Here, $r_{u,x}$ is user-ratings of target user u on item x and $p_{u,x}$ is prediction of target user u on item x. The prediction use algorithm of neighbor based collaborative filtering. This paper analyze relations between two variables that used at NBCF and EV (Error Value) and apply to the new algorithm. EV (Error Value) means error of rating value and prediction value. The definition is as following.

$$EV_{u,x} = r_{u,x} - p_{u,x}. \tag{2.4}$$

Here, $r_{u,x}$ is user-ratings of target user u on item x and $p_{u,x}$ is prediction of target user u on item x .

This paper calculate personal the EV of item and apply to experimental data.

3. Motivation

3.1. The structure of dataset

This paper use MovieLens 100K dataset at GroupLens. 100K dataset is user-ratings that 943 users entered between 1 to 5 points for 1682 movies. Also, each dataset contains age, sex, occupations and zip code. The dataset structure is defined below.

The dataset compose with three data sets that is named data1, data2 and data3. The data contains 100,000 dataset that are random classified by 80% training set and 20% test set. The data2 contains 100,000 training set and100,000 test set. The data3 contain 70% training set that removes 30% data in 100% dataset and 20% test set. The training set is

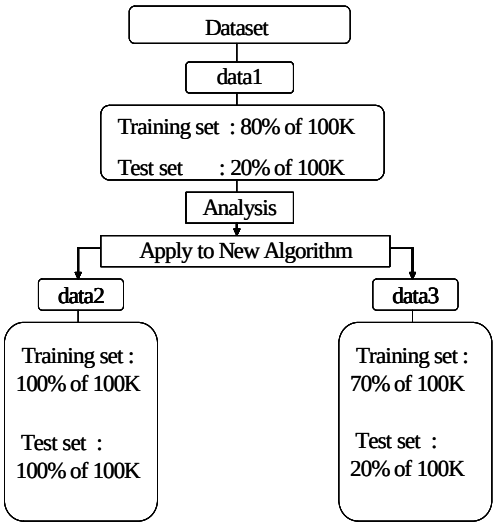

Figure 3.1 Composition of experiment dataset

used to calculate prediction value of new algorithm. The test set is used to get prediction accuracy.

This paper analyze two variable at data1 by current algorithm and define new algorithm. Also, in order to know improvement of prediction accuracy, this paper apply the suggested algorithm to data2 and data3 which is sparsity data. The sparsity data decrease prediction accuracy in general (Kim *et al.*, 2007; Lee *et al.*, 2006; Lee *et al.*, 2007).

But, this paper describe prediction accuracy is little increased at sparsity data.

The data3 was composed with data which has a scarcity compared in data1 in order to evaluate the performance of new suggest algorithm.

This paper suggest new algorithm after analyzing of variable at data1 and then evaluate the performance of the suggest algorithm by applying a same method to data2 that is full dataset and data3 that includes scarcity.

3.2. Analysis of the suggested variable

The recommender systems that uses NBCF algorithm is used to recommend neighbor information. To calculate prediction value of target user, they are important variable that are average target user, correlation of neighbor and average user-ratings of neighbor. This paper analyze of these variable and the EV and study to improve prediction accuracy.

At first, we study value of the EV at data1, defined below.

Table 3.1 The analyzed value of the EV

	Average	Min. Value	Max. Value
EV	0.0265	-1.2802	1.9330

The minimum the EV of data1 is -1.2802 and maximum is 1.9330. The EV distribution

defined below.

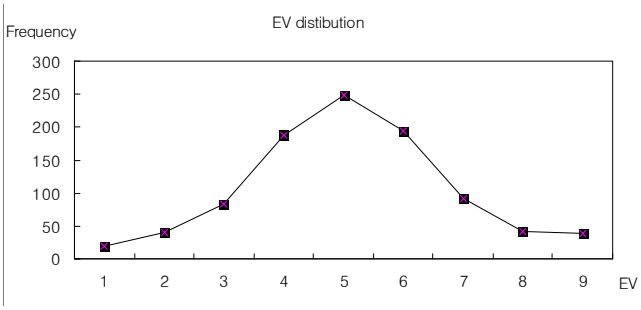

Figure 3.2 The EV distribution at data1

EV data that is smaller than -0.82 are 12, and the data between -0.1 and 0.1 are concentrated.

Table 3.2 The analyze of the EV scope

	1	2	3	4	5	6	7	8	9
scope	-1.3 -0.7	-0.7 -0.5	-0.5 -0.3	-0.3 -0.1	-0.1 0.1	0.1 0.3	0.3 0.5	0.5 0.7	0.7 2.0
N	19	40	83	187	248	194	92	41	39

This paper analyzes two variable of NBCF algorithm to reduce prediction error. The analyzed two variable are defined below.

Table 3.3 Analyze of two variable, $\bar{r_u}$and $\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/|c_{u,j}|$

Variable	Average	Min	Max
$\bar{r_u}$	3.5908	1.4713	4.8889
$\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/\|c_{u,j}\|$	0.0143	-0.7082	0.4359

The average of $\bar{r_u}$ is 3.5908 and average of $\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/|c_{u,j}|$ is 0.0143.

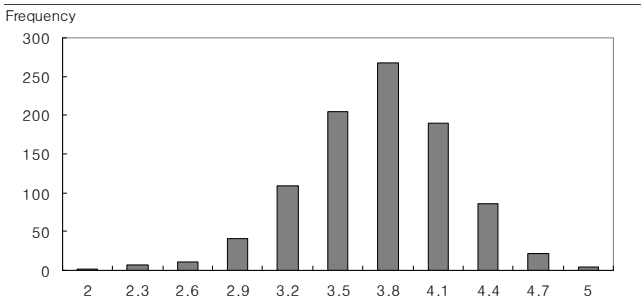

Figure 3.3 $\bar{\mathbf{r_u}}$ distribution at data1

$\bar{r_u}$ is distributed like as Fig. 3.3, and the shape is a similar with the EV distribution chart.

Distribution and category of $\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/|c_{u,j}|$, defined below.

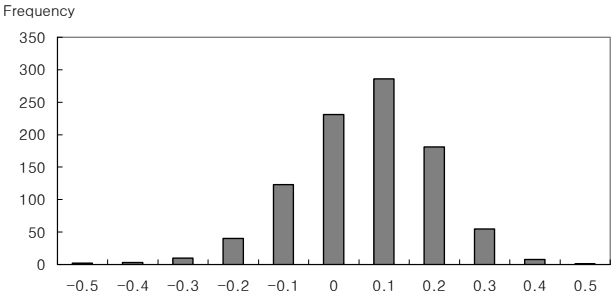

Figure 3.4 $\sum_{\mathbf{j=1}}^{\mathbf{k}}(\mathbf{r_j} - \bar{\mathbf{r_j}})\mathbf{c_{u,j}}/|\mathbf{c_{u,j}}|$ distribution at data1

Variable $\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/|c_{u,j}|$ is bell shape distribution that is similar with $\bar{r_u}$. Their center is 0.0143. It's average value.

Hence, we suggest the method to reduce error by analyzing of the EV and two variable, $\bar{r_u}$ and $\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/|c_{u,j}|$. The relation of two variable and the EV, defined below.

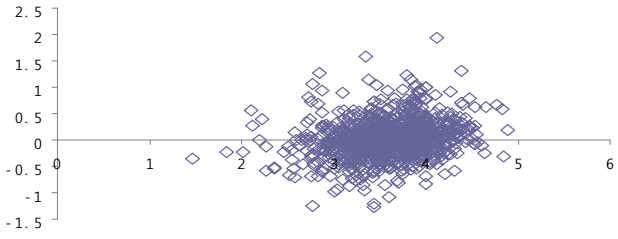

Figure 3.5 The relation plot of the EV and $\bar{r_u}$

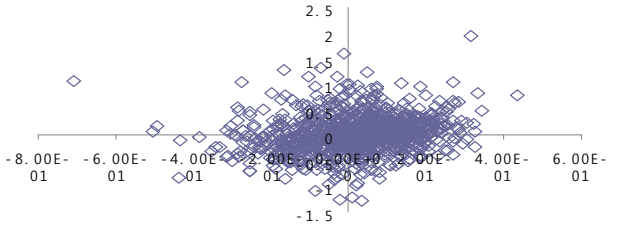

Figure 3.6 The relation plot of the EV and $\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/|c_{u,j}|$

Hence, we used correlation coefficient to know the relation of the EV and two variable, $\bar{r_u}$ and $\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/|c_{u,j}|$.

Table 3.4 A correlation coefficient with personal EV of NBCF and two variable

Variable	Personal EV of NBCF		
	N	Cor.	Significance Probability
$\bar{r_u}$	943	0.2274	0.00*
$\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/\lvert c_{u,j}\rvert$	943	0.1818	0.00*

*: p<0.01

4. New algorithm

This paper demonstrate that two variable influenced to the EV and confirmed statistically. Hence, this paper suggests the method to improve of prediction accuracy of recommender systems by using Exploratory Data Analysis (EDA) about these two variable.

Exploratory Data Analysis (EDA) is analyzing method to get data structure and feature, and Least Squares Method (LSM) in EDA is method to get straight equation that passes by nearest each point on scatter gram. The calculated straight equation should pass the nearest each point on scatter gram. Therefore, this straight equation is perpendicular to each point, and the perpendicular distance should be minimized and it is summation of the square of each point. Hence each point on scatter gram is assigned to $(\bar{r_u},\ \sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/\lvert c_{u,j}\rvert,\ r_{u,x} - p_{u,x})$, the set of points that correspond to $(\bar{r_u},\ \sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/\lvert c_{u,j}\rvert)$ is expressed by straight line that compose with $(\bar{r_u},\ \sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/\lvert c_{u,j}\rvert,\ Y^*)$. The fitted line defines as the two variable, shown in equation (4.1)

$$Y^* = a\bar{r_u} + b\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/\lvert c_{u,j}\rvert + c. \tag{4.1}$$

The fitted line is calculated by a, b and c that is satisfied with the following conditions.

$$S = \min\sum_{x=1}^{n}(r_{u,x} - p_{u,x} - Y^*)^2. \tag{4.2}$$

Hence, the following expression is used to get the fitted line.

$$\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = \frac{\partial S}{\partial c} = 0. \tag{4.3}$$

4.1. Definition of the new algorithm

This paper suggest new algorithm of recommender systems as EDA that use LSM. The suggested new algorithm expressed to fitted function, defined below.

$$F_{data}(a,b,c) = a\bar{r_u} + b\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/\lvert c_{u,j}\rvert + c. \tag{4.4}$$

Constant a, b, c may be changed by data, and they are calculated by expression (4.3).

4.2. New algorithm that applied to dataset

We must calculate the fitted line to know fitted equation. And, we need to analyze two variable, $\bar{r_u}$ and $\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/|c_{u,j}|$, to apply each experiment data to the fitted line. The result which applied the two variable to expression (4.3) is expressed to the following equation.

$$\sum_{x=1}^{m} r_{u,x}\bar{r_u} - a\sum_{x=1}^{m}\bar{r_u}\bar{r_u} - b\sum_{x=1}^{m}\sum_{j=1}^{k}\frac{(r_j - \bar{r_j})c_{u,j}}{|c_{u,j}|}\bar{r_u} - c\sum_{x=1}^{m}\bar{r_u} = 0.$$

$$\begin{aligned}&\sum_{x=1}^{m} r_{u,x}\sum_{j=1}^{k}\frac{(r_j - \bar{r_j})c_{u,j}}{|c_{u,j}|} - a\sum_{x=1}^{m}\bar{r_u}\sum_{j=1}^{k}\frac{(r_j - \bar{r_j})c_{u,j}}{|c_{u,j}|} \\ &- b\sum_{x=1}^{m}\sum_{j=1}^{k}\frac{(r_j - \bar{r_j})c_{u,j}}{|c_{u,j}|}\sum_{j=1}^{k}\frac{(r_j - \bar{r_j})c_{u,j}}{|c_{u,j}|} - c\sum_{x=1}^{m}\sum_{j=1}^{k}\frac{(r_j - \bar{r_j})c_{u,j}}{|c_{u,j}|} = 0.\end{aligned} \tag{4.5}$$

$$\sum_{x=1}^{m} r_{u,x} - a\sum_{x=1}^{m}\bar{r_u} - b\sum_{x=1}^{m}\sum_{j=1}^{k}\frac{(r_j - \bar{r_j})c_{u,j}}{|c_{u,j}|} - c\sum_{x=1}^{m} 1 = 0.$$

The computed result of the two variable to expression (4.5) is defined below.

Table 4.1 The fitted equation that depend on data

dataset	a	b	c
data1	0.83520	1.17559	0.58005
data2	0.77990	1.38200	0.82864
data3	0.83597	1.14728	0.57942

Hence, The result which applied data to new algorithm is expressed to the following fitted equation.

$$F_{data1}(a,b,c) = 0.83520\bar{r_u} + 1.17559\sum_{j=1}^{k}\frac{(r_j - \bar{r_j})c_{u,j}}{|c_{u,j}|} + 0.58005.$$

$$F_{data2}(a,b,c) = 0.77990\bar{r_u} + 1.38200\sum_{j=1}^{k}\frac{(r_j - \bar{r_j})c_{u,j}}{|c_{u,j}|} + 0.82864. \tag{4.6}$$

$$F_{data3}(a,b,c) = 0.83597\bar{r_u} + 1.14728\sum_{j=1}^{k}\frac{(r_j - \bar{r_j})c_{u,j}}{|c_{u,j}|} + 0.57942.$$

Table 4.2 The result of total the MAE that applied the fitted equation

dataset	MAE		N	t	Significance Probability
	NBCF Algorithm	New Algorithm			
data1	0.75271	0.75093	19973	2.72	0.00*
data2	0.61476	0.5885	99984	44.38	0.00*
data3	0.75867	0.75736	19969	2.13	0.01**

*:p<0.01, **:p<0.05

4.3. Prediction accuracy of new algorithm

We applied new algorithm that use fitted equation to training data and calculated prediction value. Also, we calculated the MAE of new algorithm with test set. Hence, the result that compared the total MAE of NBCFA with new algorithm is defined below.

The MAE that applied data1 to the suggested fitted equation is 0.75093. It is better than 0.75271, the MAE of NBCFA. Also, The MAE of data 2 is 0.58850. It's better than 0.61476, the MAE of NBCFA. In case of data3 that include scarcity, the value that applied to fitted equation is smaller. As a result of experiment, the result that applied to fitted equation is improved.

Hence, we apply the MAE to personal experimental data with fitted equation.

data 1 improved 0.78065 of NBCFA to 0.77857 value that applied the fitted equation. The MAE of NBAFA is 0.78065 and the MAE of using fitted equation is 0.77857. In case of data 2, the result that applied fitted equation is 0.57602. It's better than 0.76436, the MAE of NBCFA. Hence, the result that applied fitted equation is improved than the result of NBCFA. In case of data 3, the result that applied fitted equation is 0.78656. It's better than 0.78820, the MAE of NBCFA.

Table 4.3 Personal the MAE that applied the fitted equation

dataset	Algorithm	MAE	N	t	Significance Probability
data1	NBCF Algorithm	0.78065	943	2.08	0.01*
	New Algorithm	0.77857			
data2	NBCF Algorithm	0.76436	943	54.8	0.00*
	New Algorithm	0.57602			
data3	NBCF Algorithm	0.7882	943	1.53	0.06*
	New Algorithm	0.78656			

*:p<0.1

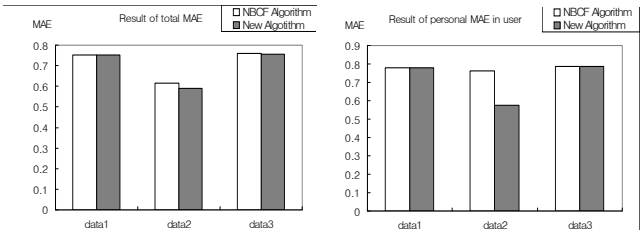

Figure 4.1 Comparison chart that total the MAE and personal MAE that applied the fitted equation

5. Conclusion

In the recommender systems that use NBCF, we compute the prediction accuracy from a relationship with user-ratings, average of target user, the similarity weighting rate with neighbor and average of neighbor. Hence, we suggest new algorithm by analyzing of variable, $\bar{r_u}$ and $\sum_{j=1}^{k}(r_j - \bar{r_j})c_{u,j}/|c_{u,j}|$. The new algorithm use fitted line that applied LSM in EDA. The result which applied the experimental data to fitted line is expressed to the fitted equation. We compared the MAE of fitted equation with the MAE of NBCFA and found that the result which applied the fitted equation to all of data is improved than the MAE of NBCFA.

In personal MAE and total MAE, the result that applied the fitted algorithm is better than the MAE of NBCFA.

Hence, we expect that the prediction accuracy of recommender systems is improved when the fitted equation is applied to each data.

References

Kim, S. O. (2008). Improving the MAE by removing lower rated items in recommender system. *Journal of the Korean Data & Information Science Society*, **19**, 819-830.

Kim, S. O., Lee, K. H., Lee, S. J. and Lee, H. C. (2009). Study of the mean and information of neighbors in NBCFA. *Proceedings of the Spring Korea Society of IT Services*, 345-348.

Kim, S. O. and Lee, S. J. (2007). The effect of data sparsity on prediction accuracy in recommender system. *Journal of the Korean Society for Intenet Information*, **8**, 9-15.

Kim, S. O., Lee, S. J. and Lee, H. C. (2008). A study on improvement of prediction accuracy by critical value. *Journal of the Korean Data Analysis Society*, **10**, 591-601.

Konstan, B., Miller, D., Herlocker, J., Gordon, L. and Riedl, J. (1997). GroupLens : Applying collaborative filtering to usenet news. *Communications of the ACM*, **40**, 77-87.

Lee, H. C. (2006). On the effect of significance of correlation coefficient for recommender system. *Journal of the Korean Data & Information Science Society*, **17**, 1129-1139.

Lee, H. C, Kim, S. O and Lee, S. J. (2007). A study on the interrelationship between the prediction error and the rating's pattern in collaborative. *Journal of Korean Data & Information Science Society*, **18**, 659-668.

Lee, H. C., Lee, S. J. and Jung, Y. J. (2006). The effect of co-rating on the recommender system of user base. *Journal of the Korean Data & Information Science Society*, **17**, 775-784.

Lee, S. J., Kim, S. O. and Lee, H. C. (2007). Pre-Evaluation for detecting abnormal users in recommender system. *Journal of the Korean Data & Information Science Society*, **18**, 619-628.

Melville, P., Mooney, R. and Nagarajan, R. (2002). Content boosted collaborative filtering for improved recommendations. *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, 187-192.

Pazzani, M. J. (1999). Framework for collaborative, content based and demographic filtering. *Artificial Intelligent Review*, 394-408.

Resnick, P. N., Iacovou, M., Bergstrom, P., Bergstrom, J. and Riedl, J. (1994). GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, 175-186.