Journal of the Korean Data & Information Science Society
2010, **21**(4), 803–811

한국데이터정보과학회지

The relationship between prediction accuracy and pre-information in collaborative filtering system

Sun Ok Kim[1]

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

Received 2 June 2010, revised 14 July 2010, accepted 19 July 2010

Abstract

This study analyzes the characteristics of preference ratings by dividing estimated values into four groups according to rank correlation coefficient after obtaining preference estimated value to user's ratings by using collaborative filtering algorithm. It is known that the value of standard error of skewness and standard error of kurtosis lower in the group of higher rank correlation coefficient. This explains that the preference of higher rank correlation coefficient has lower extreme values and the differences of preference rating values. In addition, top n recommendation lists are made after obtaining rank fitting by using the result ranks of prediction value and the ranks of real rated values, and this top n is applied to the four groups. The value of top n recommendation is calculated higher in the group of higher rank correlation coefficient, and the recommendation accuracy in the group of higher rank correlation coefficient is higher than that in the group of lower rank correlation coefficient. Thus, when using standard error of skewness and standard error of kurtosis in recommender system, rank correlation coefficient can be higher, and so the accuracy of recommendation prediction can be increased.

Keywords: Collaborative filtering, rank correlation coefficient, recommender system, top n.

1. Introduction

Although e-commerce has increased rapidly in on-line environment, the service has not yet satisfied customers. Because of this situation, recommender system to satisfy customers has been introduced, and utilizing the recommender system in the speedy internet environment is becoming more important. Preference estimated value in the recommender system is calculated according to the information of customers, and the study about recommender algorithm to improve this has been conducted steadily. The study of recommending ranks by using calculated preference estimated value has also been conducted. When recommending items in e-commerce, the algorithm which estimated value is accurate in recommending ranks is needed. It is more important that the method to increase the accuracy of recommending ranks by using preference estimated value needs to be studied.

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

2. Related works

Collaborative filtering system predicts the customer preference of items depending on customer's past attitudes to items (i.e. the preference evaluation of items). Collaborative filtering algorithm can be divided into two categories; memory-based and model-based recommender system. memory-based recommender system consist of user-based rating and model-based recommender system is made up of item-based rating. The prediction of recommending accuracy was researched by using user-based recommender system in memory-based recommender system for this study. Collaborative filtering approach is the one that links customers with the relating items. For example, movie profile depicts the genre of the movie, characters in the movie, and the success rank of the box office, and customer profile can be obtained through the Internet and demography statistic information.

Lee (2006a) presented correspondence mean algorithm (CMA) that was improved from neighborhood based collaborative filtering algorithm, to improve the precise preference prediction of recommender system.

Lee (2006b) researched ranks fitting to recommend Top-N in the recommender system. Also, Lee (2007) studied the ranks of Top-N using improved algorithm in collaborative filtering system. Lee *et al.* (2007a) provides an improvement of prediction accuracy of recommender system. Lee *et al.* (2007b) studied Run of abnormal user's ratings on prediction accuracy of recommender system. Yang *et al.* (2008) researched the characteristics of preference prediction in memory-based collaborative filtering system. Lee (2007c) conducted a research on the correlation of customer's standard deviation and MAE (Mean Absolute Error). Kim *et al.* (2008) studied improvement of prediction accuracy according to critical value. Lee (2008) suggested the way to increase Top-N recommending accuracy by using data supplementation method.

2.1. Preference prediction algorithm

1) Neighborhood based collaborative filtering algorithm (NBCFA)

First automatic collaborative filtering algorithm was proposed by GroupLens for recommending articles or news in UseNet News discussion forum based on internet, named as Neighborhood-Based Collaborative Filtering Algorithm (NBCFA) by Resnick *et al.* (1994). The first stage of GroupLens system used similarity weight for calculating each user's preference analogy as Pearson's correlation coefficient and used all relationships among users' preference to writings. To predict the degree of preference of target user who wanted to recommend specific articles, the following NBCFA was calculated to predict rating value for a target user.

$$\widehat{U}_x = \overline{U} + \frac{\sum_{j \in raters}(J_x - \overline{J})r_{uj}}{\sum_{j \in raters}|r_{uj}|}, \text{ where } \overline{J} = \frac{\sum_i^n J_i}{n}, i \neq x \tag{2.1}$$

In equation (2.1), the $\widehat{U}_x$ denotes the prediction value of the preference of the target user u over the target item x, the $\overline{U}$ is the mean of the all preference ratings that have been rated by the user u, the j_x is the preference rating of the neighbor user j over the target item x, and the $\overline{J}$ is the mean of the all preference ratings of the neighbor user j except the rating of target item x. Raters are users who rate the preference of the item in the data set. The r_{uj} is the similarity weight of both the user u and the neighbor user j.

2) Correspondence mean algorithm (CMA)

Correspondence mean algorithm (CMA) for improving prediction accuracy of NBCFA was proposed by Lee *et al.* (2007). The concept of NBCFA was to predict target user's preference ratings by selecting neighbor users who had rated on the same target item which will be calculated for the target user and unseen as yet, and their preference relationships between the target user and his or her neighbor users are defined as similarity weights. However, NBCFA had some defects as unnecessary or a bit much target user or neighbor users' information was reflected on calculating or prediction process. So, CMA was proposed for amending those defects. Next equation is CMA.

$$\widehat{U}_x = \overline{U_{match}} + \frac{\sum_{j \in raters}(J_x - \overline{J_{match}})r_{uj}}{\sum_{j \in raters}|r_{uj}|} \tag{2.2}$$

In the NBCFA, $\overline{U}$ is the mean of the preferences of the user u. In this case of using the mean of the entire rating of the user u. the preference of the user u is overestimated, which leads to a possibility that the preference of the user j might be sufficiently reflected. This is why in the CMA $\overline{U}_{match}$ is used, which is the mean of means of the preferences that are marked by both the user u and the neighbor user j. The mean $\overline{J}_{match}$ the mean of the preferences marked by both the u and the j, and it is calculated by the same way of calculating the Pearson's correlation coefficient.

Several techniques have been used to evaluate recommender system. Those techniques have been divided by Herlocker *et al.* (2004) into three categories, predictive accuracy metrics, classification accuracy metrics and rank accuracy metrics. The predictive accuracy metrics measure how close the predicted ratings by algorithm are to the true ratings in the test dataset. In this study, Mean absolute error (MAE), one of the predictive accuracy metrics, is used to evaluate the performance of each algorithm.

$$MAE = \frac{1}{N}\sum_{j}^{N}|R_{uj} - \widehat{R}_{uj}| \tag{2.3}$$

In this equation, R_{uj} is the true rating of user given to the item j and $\widehat{R}_{uj}$ is the prediction value of user u to the item j.

2.2. Prediction ranks and top n recommendation

Recommender system recommends items that users may be fond of in e-commerce by providing upper N number of item lists. At this, the recommendation of item lists of upper N numbers is defined as top n recommendation. Recommender system based on collaborative filtering approach recommends top n by using the information of neighboring customer as the information of items. Sarwar *et al.* (2000) presented top n recommendation as follows.

1) Recommendation of the highest frequency items: Preference frequency of neighboring customer's items is obtained by using the information of neighboring customer who marks the preference of items. According to this frequency, arranging the very items, N number of items which are higher frequency are recommended.

2) Association Rule Based Recommendation: By using association rule, the relation of former-purchased rule and the next rule is arranged according to the confidence, the higher confidence items are recommended.

Data used in this research is rated 1-5 in preference of each individual's response. So, recommendation of top n cannot be recommended according to the frequency of purchased items. In this study, estimated rating value of each individual is obtained and then according to the ranks of estimated rating value the top n can be recommended.

2.3. Rank fitting

Rank fitting presents the numbers of the ranks that estimated rating value and real preferred rating value agree with each other, and according to the rank ratio of estimated rating value, rank fitting is presented. At this, the recommendation lists of top n are decided by the ranks of estimated rating value.

1) Adjusted rank fittingRank fitting is defined as follows.

$$\frac{N(Top\_N(R_u) \cap Top\_N(\widehat{R}_u))}{N(Top\_N(\widehat{R}_u))} \cdot 100 \tag{2.4}$$

Where, $N(Top\_N(\widehat{R}_u))$ is the numbers of the upper estimated rating value in $\widehat{R}_u$, preference estimated value of target customer u, and $N(Top\_N(R_u) \cap Top\_N(\widehat{R}_u))N(Top\_N(\widehat{R}_u))$ is the numbers of the upper N lists of estimated preference value and the numbers of real preferred upper N items corresponding with this.

2) Spearman correlation coefficientRank fitting is represented by the agreement ratio of the ranks of preference prediction value and rating value. Spearman correlation coefficient is used to find out fitting to estimated preference ranks and the real preference ranks. Equation of spearman correlation coefficient is as follows;

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(R_i - \widehat{R}_i)}{n(n^2-1)},\ -1 \le r_s \le 1 \tag{2.5}$$

Where, R_i is the ranks of real ratings which each individual rated in dataset, $\widehat{R}_i$ is the ranks of estimated preference ratings corresponding to the real ratings.

3. Method

In recommender system, it is preferred that the difference between the real preferred rating value and estimated rating value is small. However, in the real top n recommendation, top n recommendation is conducted according to the ranks of estimated rating value. So, even though the difference of estimated rating value is small, it is important for the ranks of estimated rating value and the ranks of real preferred rating value to agree with. Though the difference of estimated rating value is big, when the ranks of estimated rating value agree with the real preferred rating value, better recommendation can be provided to customers. Table 3.1 shows the example that the ranks of top n agree more in the case of small difference and relatively big difference of estimated value.

In the above Table 3.1, MAE in Estimated Value (1) is smaller than that of in Estimated Value (2), but Top-N rank accuracy is higher in Estimated Value (2). Customer is interested not in the prediction value in Estimated Value (1) but in the rank accuracy. So, the ranks of

Table 3.1 Recommendation accuracy and Top-N ranks (Lee, 2009a)

Rating Value	Estimated Value (1)	Estimated Value (2)	Top-N Ranks in Estimated Value (1)	Top-N Ranks in Estimated Value (2)
3	3.1	2.3	3	4
4	4.5	2.8	1	2
5	4.3	2.9	2	1
3	2.9	2.6	4	3
2	2.5	2.2	5	5
MAE	0.38	0.92	MAE is smaller in Estimated Value (1)	
Spearman's Rank Correlation	0.872	0.975	Rank accuracy is higher in Estimated Value (2)	

the recommending lists play a important role in the customer's satisfaction and the reliance (Lee, 2009a).

100 MovieLens dataset in GroupLens is used to analyze the experimental data. The dataset is divided into two groups; 80% of training data and 20% of test data. Estimated value is calculated by using NBCFA, and the procedures of this experiment are as follows;

1) Estimated rating value of the entire data is calculated by NBCFA.

2) Spearman correlation coefficient between estimated value and the real preferred rating value is calculated.

3) Four groups of respondents are categorized according to the differences of calculated spearman correlation coefficient.

4) Whether which pre-information of respondents in four groups has significance is tested.

5) Whether top n (N=2, 3) of respondents in four groups is tested.

4. Experiment and analysis

4.1. Experimental dataset

The experimental data is 100k MovieLens dataset in MovieLens dataset of GroupLens. 943 users in 100k MovieLens dataset rate 1682 movies according to the preference rates with a score 1-5, and the total number 100,000 rates are divided into 80% of test data and 20% of training data. Preference rating value of training data is calculated by applying collaborative filtering to test data. The rank relation of calculated estimated value and preference rating value is used in the experiment by using spearman rank correlation coefficient. The following Figure 4.1 shows the frequency of preference of the total data in training dataset. Table 4.1 shows the distribution according to the quartile of standard error of kurtosis and standard error of skewness in test data.

Table 4.1 Distribution of standard error of kurtosis and standard error of skewness

Classification	Min	1 quartile	2 quartile	3 quartile	Max	Mean
Standard error of kurtosis	0.4	0.83	1.19	1.59	2.62	1.29
Standard error of skewness	0.2	0.43	0.62	0.79	1.01	0.62

Table 4.2 shows the distribution according to the quartile of top 2 and top 3 in test data.

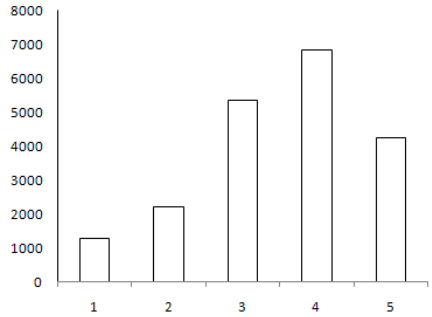

Figure 4.1 Frequency of preference in training data

Table 4.2 Distribution of top 2, top 3

Classification	Min	1 quartile	2 quartile	3 quartile	Max	Mean
top 2	0	50	50	100	100	59.11
top 3	0	33.33	66.66	100	100	66.06

The respondents are divided into four groups according to the rank correlation coefficient. Table 4.3 shows the distribution of respondents' rating value in test data.

Table 4.3 Frequency of preference according to groups

rating	Frequency				
	Group1	Group2	Group3	Group4	Total
1	5 (0.4%)	70 (5.5%)	618 (48.9%)	572 (45.2%)	1265 (100%)
2	14 (0.6%)	80 (3.6%)	1100 (49.9%)	1011 (45.9%)	2205 (100%)
3	37 (0.7%)	216 (4.0%)	2780 (51.8%)	2329 (43.4%)	5362 (100%)
4	56 (0.8%)	400 (5.9%)	3428 (50.1%)	2952 (43.2%)	6836 (100%)
5	37 (0.9%)	271 (6.4%)	2047 (48.1%)	1903 (44.7%)	4258 (100%)
Total	149 (0.7%)	1037 (5.2%)	9973 (50.1%)	8767 (44.0%)	19926 (100%)

χ^2 =56.765 df=12 significance probability=0.00**

Cross table analysis of the four groups of respondents divided by rank correlation coefficient and preference rating value shows the relationship each other in statistics. In relative comparison, the higher correlation coefficient group shows lower response in preference rating value, and the lower correlation coefficient group shows higher response in preference rating value.

4.2. Experiment data analysis and results

The respondents are divided into four groups according to the rank correlation coefficient. To find out the relationship between these groups and estimated rank, distribution analysis by standard error of skewness and standard error of kurtosis is as the following Table 4.4 and Table 4.5.

Standard error of skewness and standard error of kurtosis show significant differences according to the groups. The group of higher rank correlation coefficient shows smaller

Table 4.4 Oneway ANOVA of standard error of kurtosis according to groups

Group	Mean	Sum of squares Between Groups	Sum of squares Within Groups	Mean squares Between Groups	Mean squares Within Groups	F	Sig.	Duncan's
1	1.92							
2	1.65	29.96	275.14	9.98	0.29	33.72	0.00**	3,421
3	1.18							
4	1.27							

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

Table 4.5 Oneway ANOVA of standard error of skewness according to groups

Group	Mean	Sum of squares Between Groups	Sum of squares Within Groups	Mean squares Between Groups	Mean squares Within Groups	F	Sig.	Duncan's
1	0.86							
2	0.77	4.95	40.82	1.65	0.04	37.54	0.00**	3,421
3	0.57							
4	0.61							

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

standard error of kurtosis and standard error of skewness. This can be explained that when recommending top n, by standard error of kurtosis and standard error of skewness, the pre-information of respondents, the agreement of estimated ranks can be detected in advance. The respondents are divided into four groups according to the rank correlation coefficient. To find out the relationship between these groups and top 2 and top 3, distribution analysis is as following Table 4.6 and Table 4.7.

Table 4.6 Oneway ANOVA of top 2 according to groups

Group	Mean	Sum of squares Between Groups	Sum of squares Within Groups	Mean squares Between Groups	Mean squares Within Groups	F	Sig.	Duncan's
1	0.15							
2	0.40	192035	992643	64011	1068	59.91	0.00**	1234
3	0.52							
4	0.74							

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

Table 4.7 Oneway ANOVA of top 3 according to groups

Group	Mean	Sum of squares Between Groups	Sum of squares Within Groups	Mean squares Between Groups	Mean squares Within Groups	F	Sig.	Duncan's
1	0.53							
2	0.54	102225	695794	34075	748	45.49	0.00**	1,2,34
3	0.58							
4	0.78							

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

top n (N=2,3) shows significant differences according to the groups. The group of higher

rank correlation coefficient shows higher top 2 and top 3. This means that as estimated value and rating value agree with each other, top n is higher.

5. Conclusion

In this study, after obtaining estimated value of preference on the test data of 100k MovieLens dataset by using collaborative filtering algorithm, and then spearman correlation coefficient between the real rating value and estimated value is calculated. According to the size of spearman correlation coefficient, respondents are divided into four groups. Characteristics of preference ratings in these groups are analyzed. It is found out that as the rank correlation coefficient in groups is higher, the value of standard error of kurtosis becomes smaller. In addition, it takes it for granted that as rank correlation coefficient is higher, top n becomes higher. This results mean that as standard error of kurtosis and standard error of skewness in ratings in test data, pre-information of respondents, can be perceived in advance, N value of top n differentiates according to the respondents. I would like to suggest the followings in further studies.

First, when recommending items to customers in e-commerce, the number of N value of top n should be suggested by using pre-information of customer. Second, considering the relationship between MAE and top n, a new and comprehensible recommending guideline for recommending should be established. Third, what this recommending landmark can be related to customer's loyalty should be studied and what relation customer's loyalty to items and MAE, top n can have needs to be studied.

References

Herlocker, J., Konstan, J., Terveen, L. J. and Riedl, J. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, **22**, 5-53.

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, 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. (2006a). A study on the rank fitting of the recommender system for Top-N recommender. *Journal of Korean Data & Information Science*, **8**, 2597-2607.

Lee, H. C. (2006b). 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. (2007). A study on the Top-N using the improved algorithm in the collaborative filtering. *Journal of The Korean Data Analysis Society*, **9**, 491-500.

Lee, H. C. (2008). A study on the relationship between sparsity and Top-N recommendation in the recommender system. *Journal of The Korean Data Analysis Society*, **11**, 2389-2399.

Lee, H. C. (2009a). The characteristic of response and Top-N accuracy in collaborative filtering of recommender system. *Journal of The Korean Data Analysis Society*, **11**, 2919-2930.

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

Lee, S, J., Kim, S. O. and Lee, H. C. (2007b). The relationship of prediction accuracy and the run of abnormal user's ratings in collaborative filtering. *Journal of The Korean Data Analysis Society*, **9**, 2043-2054.

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

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.

Resnick, P. and Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, **40**, 56-58.

Sarwar, B. M., Karypis, G., Konstan, J. and Riedl, J. (2000). Analysis of recommender algorithms for e-commerce. *In Proceedings of the 2st ACM Conference on Electronic Commerce*, 158-167.

Yang, K. M., Lee, H. C. and Park, Y. S. (2008). The feature of preference prediction by memory-based collaborative algorithm. *Journal of The Korean Data Analysis Society*, **10**, 591-601.