A Neighbor Selection Method Based on Priority of the Number of Common Items in Collaborative Filtering

Hyeong-Joon Kwon and Kwang-Seok Hong

School of Information and Communication Engineering, Sungkyunkwan University,
300, Chunchun-dong, Jangan-gu, Suwon, Kyungki-do, 440-746, South Korea
katsyuki@skku.edu and kshong@skku.ac.kr

Abstract. Memory-based collaborative filtering (MBCF) selects the top-k neighbors in order to predict a rating for an item the target user has not yet experienced. We propose a method to minimize the similarity errors with the existing neighbor selection method by considering the number of common items between two objects. To verify the proposed method, we compare experimental results on the existing method and the proposed method. As a result, we were able to confirm that the proposed method can improve the prediction accuracy by proposed neighbor selection.

Keywords: Collaborative Filtering, Neighborhood Selection, Personalization

1 Introduction

The most important performance evaluation criterion of MBCF, which is to predict a rating of an items for target user, is the prediction accuracy, which is estimated from the error between the CF and the actual rating. One of reasons for the decline in CF prediction accuracy is the data sparsity problem, which is due to an insufficient number of user ratings from the user-item rating dataset [1-2]. Existing studies have proposed various methods to solve data sparsity problems [3-4]. One of them is to develop the novel neighbor selection method [3].

In this paper, we deal with neighbor selection in order to reduce the data sparsity problem. The proposed method considers the number of co-items as a priority when similarity is evaluated between two objects. The proposed neighbor selection method minimizes the decrease in performance that results from a lack of the number of co-items and increases the prediction accuracy in many co-items. We have confirmed the effectiveness of the proposed method with a full-rating experiment using the Movielens 100k dataset.

2 Proposed Neighbor Selection Method

The MBCF evaluates the similarity between a target object and every other object, and then selects a given number (constant k) of nearest neighbors. At this point, a
priority of the similarity is generally considered. Existing studies asserted that PCC and COS give the wrong similarity for a few co-items and this condition is observed sufficiently often. We drew frequency distributions from the Movielens 100k: UA dataset. We made a histogram of similarities with the total number of pairs of users in the training set and observed the frequency of the number of co-items.

![Histogram of Movielens 100K: UA](image)

Fig. 1. Histogram of Movielens 100K: UA

The histograms are shown in Fig. 1. The x-axis of the histograms indicates the number of co-item ratings while the y-axis indicates the frequency. The Movielens 100k dataset consists of 943 users and 1,682 items. Therefore, the maximum value of the x-axis is 1,682 and the maximum value of the y-axis is 444,153. The PCC and COS show wrong similarity value for a few co-items. The threshold is approximately 5 co-items and under. Fig. 1 includes 411,096 items with 5 or fewer co-items, which corresponds to just 92.55% of training results. Movielens 100k: UA is separated by a dataset distributor. The number of items in the test data set is 9,430.

The PCC gives a similarity of 0 or 1 when the number of co-items is 1 and 2. In addition, when the number of co-items is 3 and 4 has 0 or 1 as the similarity with high probability. In other words, PCC and COS do not classify similar points between objects with few co-rating items. We see that the existing neighbor selection method cannot select optimal neighbors because it considers similarity independently of the number of co-items. We have shown a critical problem of a priority of similarity via Fig. 1. To reduce this problem, we propose a novel neighbor selection method. The proposed method first considers the number of co-items. After similarity evaluation, the traditional neighbor selection method chooses neighbors according to similarity whereas the proposed method chooses neighbors according to the number of co-items.

3 Experiment and Result

We use Movielens 100k: UA, which were explained in Section 3, to verify the effectiveness of the proposed method. A full-rating experiment of Movielens 100k: UA using the traditional and proposed neighbor selection method is shown in Fig. 2. By comparison, PCC is known as robust when there are many co-items. PCC is accurate in objects with many co-items based on existing studies. Therefore, the
The proposed method should decrease the MAE of PCC because the proposed method chooses neighbors with many co-items before those with high similarity. As a result, the proposed method can improve the prediction accuracy of the CF system.

![Experimental results of Movielens 100k: UA](image)

**Fig. 2.** Experimental results of Movielens 100k: UA

### 4 Conclusion

We have suggested a novel neighbor selection method for MBCF. Our approach supplements the weakness of the existing method with a priority of similarity, which is to show wrong similarity by PCC and COS. In future work, we will experiment on effectiveness of proposed method with a various dataset.

**Acknowledgments.** This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0021411), and by MKE, Korea under ITRC NIPA-2012-(H0301-12-3001).

### References