Movie Recommender System Comparison of User-based and Item-based Collaborative Filtering Systems

1st Imam Dwicahya[1], 2nd P. H. Prima Rosa[2], 3rd Robertus Adi Nugroho[3]
{imamdwc@gmail.com[1], rosa@usd.ac.id[2], robertus.adi@usd.ac.id[3]}

Faculty of Science and Technology, Sanata Dharma University, Indonesia[123]

Abstract. Collaborative Filtering (CF) is a method that is widely used in recommendation systems. There are two approaches that are often used in CF, namely User-based CF and Item-based CF. The User-based CF approach requires several similar users to predict the rating of a new item. Meanwhile, Item-based CF requires several similar items to predict the rating of a new item. The number of similar users or similar items involved in predicting ratings will affect the computing load. This research aims to see the effect of the number of neighbors (similar user or similar items) used on the level of accuracy of rating predictions for an item. By using different numbers of neighbors for both User-based CF and Item-based CF, the results of the experiment show that the number of neighbors affects the level of accuracy although not too significant.

Keywords: Recommender system, user-based collaborative filtering, item-based collaborative filtering.

1 Introduction

In this information age, people no longer have difficulties in sharing information. Internet technology makes it easy for someone to access this information [1]. But, this convenience causes a huge amount of information on the internet. The challenge for humans in this information age is to find information that is in accordance with the needs of this vast information pool. Search Engine is present to answer that problem. However, it is not able to filter information personally[2], [3]. Therefore, a recommendation system appears to answer these problems. A recommendation system is a system that is able to provide predictions on whether an item will be liked by its users [4].

Two approaches commonly used in recommendation systems are Content-based Filtering and Collaborative Filtering. The Content-based Filtering approach predicts whether an item is preferred by the user or not based on similarity of the item with items that have been rated well by the user[5]–[7]. This approach is not able to bring up new items that user likes. The Collaborative Filtering approach is present to overcome the problem. The Collaborative Filtering approach recommends items using the principle of users who have similarities will like similar items and similar items favored by similar users[8]–[10].

Collaborative Filtering (CF) has two approaches, User-based CF and Item-based CF. User-based CF predicts the rating given by a user to an item based on the rating given by other similar users to the same item. Whereas, Item-based CF predicts the rating given by a user to an item based on the rating given by other users on similar items. Other similar users or items are called

ICSTI 2018, October 19-20, Yogyakarta, Indonesia

DOI 10.4108/eai.19-10-2018.2282541

neighbors. User-based CF and Item-based CF rely on nearest neighbors to predict ratings. The difference between User-based CF and Item-based CF is the type of neighbor.

There is no provision for the number of nearest neighbors that must be used in predicting ratings. However, the number of nearest neighbors used will affect the computational load of the prediction process and the accuracy of the predicted rating. This study tries to compare the accuracy of the prediction results in the User-based CF and Item-based CF approaches in each number of nearest neighbors used. Some experiments were carried out by changing the maximum number of nearest neighbors involved in the prediction process. The accuracy of prediction results will be measured using Mean Absolute Error (MAE)[11]–[13].

The dataset used in this research is a movie rating taken from MovieLens.org. It is an open dataset managed by GroupLens, a laboratory research at the University of Minnesota (https://movielens.org/). The calculation of similarity between users or items using Pearson-Correlation.

2 Method

This research uses a movie rating dataset from MovieLens.org which contains 100,000 ratings from 943 users and 1682 movies. The range of rating values is 1 to 5. The dataset is divided into training data and testing data. Training data of 95,299 ratings are used to determine the similarity between users (User-based CF) or between items (Item-based CF). While the testing data is 4,701 ratings, used to calculate the accuracy of the prediction results.

Testing data is selected by taking the first five ratings given by each user. The similarity between users (User-based CF) or items (Item-based CF) is calculated using the Pearson Correlation (PC). Using Pearson Correlation, positive relations and negative relations between two users or two items can be known [2]. The Pearson Correlation formula used for the User-based CF approach can be seen in equation (1).

$$PC(u,v) = \frac{\sum_{i \epsilon \mathcal{I}_{uv}} (r_{ui} - \overline{r_u})(r_{vi} - \overline{r_v})}{\sqrt{\sum_{i \epsilon \mathcal{I}_{uv}} (r_{ui} - \overline{r_u})^2 \sum_{i \epsilon \mathcal{I}_{uv}} (r_{vi} - \bar{r}_v)^2}} \qquad (1)$$

Where,

r_{ui} = the rating given by user u for item i
r_{vi} = the rating given by user v for item i
$\overline{r_u}$ = the average rating given by user u
$\overline{r_v}$ = the average rating given by user v

The Pearson Correlation Formula used for the User-based CF approach can be seen in equation (2).

$$PC(i,j) = \frac{\sum_{i \epsilon \mathcal{U}_{ij}} (r_{ui} - \overline{r_i})(r_{uj} - \overline{r_j})}{\sqrt{\sum_{i \epsilon \mathcal{U}_{ij}} (r_{ui} - \overline{r_i})^2 \sum_{i \epsilon \mathcal{U}_{ij}} (r_{uj} - \bar{r}_j)^2}} \qquad (2)$$

Where,

r_{ui} = the rating given by user u for item i
r_{uj} = the rating given by user u for item j
$\overline{r_i}$ = the average rating for item i

$\overline{r_j}$ = the average rating for item j

Sort the similarities from the largest to the smallest. Then, take a number of nearest neighbors to predict rating. This research evaluated several numbers of neighbors including 10, 30, 70, 100, and all neighbors. After the nearest neighbors are obtained, for User-based CF, the rating prediction given by the user for an item could be done using the following formula (3):

$$\hat{r}_{ui} = \frac{\sum_{v \epsilon \mathcal{N}_i(u)} sim(u,v)\, r_{vi}}{\sum_{v \epsilon \mathcal{N}_i(u)} |sim(u,v)|} \quad (3)$$

Where,

$\hat{r}_{ui}$ = the predicted rating given by user u for item i
$\mathcal{N}_i(u)$ = set of simthe ilar user who have rated item i
$sim(u, v)$ = similarity between user u and user v
r_{vi} = the real rating given by user v for item i

For Item-based CF, the rating prediction could be done using the following formula (4):

$$\hat{r}_{ui} = \frac{\sum_{j \epsilon \mathcal{N}_u(i)} sim(i,j)\, r_{uj}}{\sum_{j \epsilon \mathcal{N}_u(i)} |sim(i,j)|} \quad (4)$$

Where,

$\hat{r}_{ui}$ = the predicted rating given by user u for item i
$\mathcal{N}_u(i)$ = set of items rated by user u most similar to item i
$sim(i, j)$ = similarity between item i and item j
r_{uj} = the real rating given by user u for item j

To calculate the magnitude of the error between the predicted rating and the actual rating, Mean Absolute Error (MAE) could be used [4] [9]. Measurements using MAE follow the following formula:

$$MAE = \frac{\sum_{r_{ui} \epsilon \mathcal{R}_{test}} |\hat{r}_{ui} - r_{ui}|}{|\mathcal{R}_{test}|} \quad (5)$$

Where,

$\mathcal{R}_{test}$= training set
$|\mathcal{R}_{test}|$ = number of training set
$\hat{r}_{ui}$ = the predicted rating given by user u for item i
r_{ui} = the real rating given by user u for item i

The methodology used in this research could be illustrated in Figure 1.

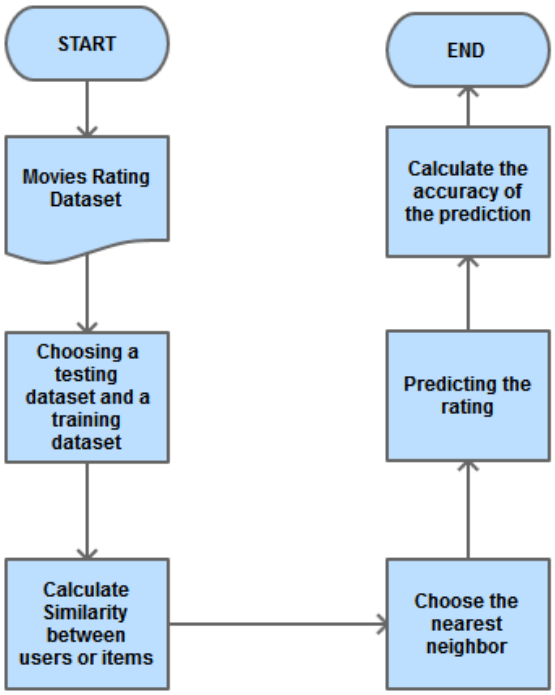

Fig. 1. Research Methodology

3 Discussion

After running several experiments on the number of neighbors used, the MAE measurement results are obtained as shown in Table 1. From the table, it can be seen that the MAE value will tend to decrease when the number of neighbors used is increasing. Figure 2 shows that trend. The accuracy obtained from the User-based CF approach is lower than that of Item-based CF. This shows that Item-based CF has better accuracy than User-based CF. The difference in accuracy is not too significant. The average difference is only about 0.006526856.

Table 1. User-based CF and Item-based CF MAE Comparison

Number of Neighbors	MAE User-based CF	MAE Item-based CF
10	0.826020531	0.833880016
30	0.804927121	0.800977548
50	0.802093284	0.796545415
70	0.800633522	0.794311284
100	0.801171485	0.793895061
All Neighbors	0.802709142	0.794503596

If you look at Figure 2, it can be seen that when the number of neighbors is 10, User-based CF is superior to Item-based CF. Whereas when the number of neighbors is more than or equal to 30, Item-based CF is superior to User-based CF. From this experiment, it can be said that User-based CF is better than Item-based CF when the number of neighbors is used a little. If the number of neighbors used is increasing, Item-based CF provides better accuracy than User-based CF.

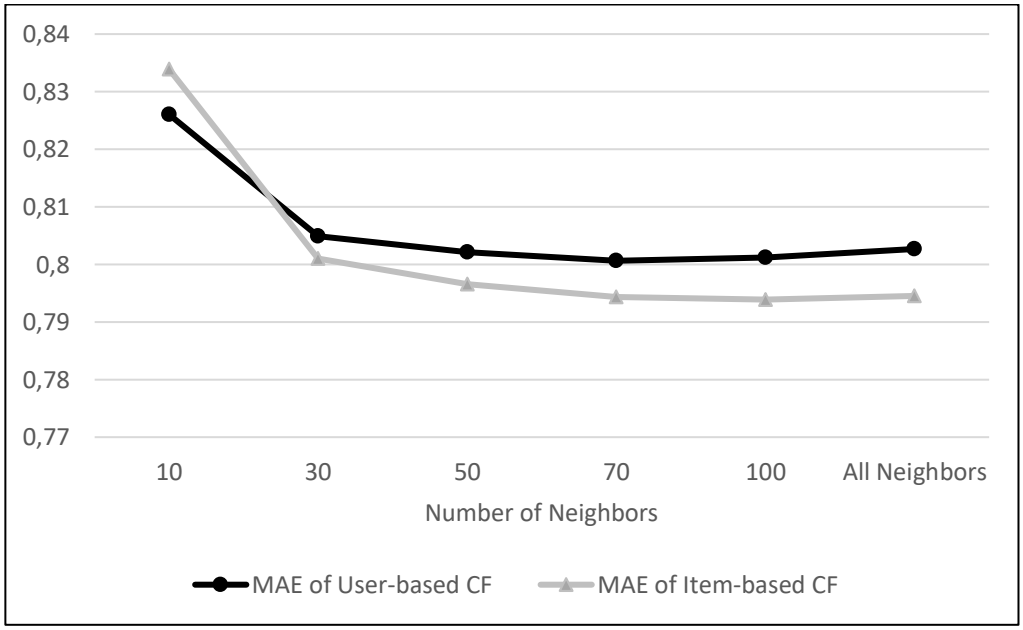

Fig. 2. User-based CF and Item-based CF MAE Comparison

4 Conclusion

By looking at the results of experiments on the number of nearest neighbors used in predicting a rating given by a user for an item, it can be concluded that the number of nearest neighbors used, the smaller the error value of a prediction, in other words, the accuracy of an approach is better. This happens to both User-based CF and Item-based CFs. The difference in accuracy between User-based CF and Item-based CF is not too significant for all the number of nearest neighbors. In the end, it can be concluded that in general the Item-based CF approach has better accuracy than User-based CF.

References

[1] G. Adomavicius and A. Tuzhilin, "Context-aware recommender systems," in *Recommender Systems Handbook, Second Edition*, 2015, pp. 191–226.

[2] R. Hock, "Search Engines," *Encyclopedia of Library and Information Sciences, Third Edition*, no. August 2013. pp. 4630–4637, 2009.

[3] J. Cho and S. Roy, "Impact Of Search Engines On Page Popularity Categories and Subject Descriptors," *Search*, pp. 20–29, 2004.

[4] M. Retrieval, P. Ii, and C. G. M. Snoek, "Video Search Engines," *Advances*, pp. 2010–2010, 2010.

[5] G. Pasi, G. Bordogna, and R. Villa, "A multi-criteria content-based filtering system," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, 2007, p. 775.

[6] T. A. Almeida and A. Yamakami, "Content-based spam filtering," in *Proceedings of the International Joint Conference on Neural Networks*, 2010.

[7] J. M. Gómez Hidalgo, G. C. Bringas, E. P. Sánz, and F. C. García, "Content-based SMS spam filtering," in *Proceedings of the 2006 ACM symposium on Document engineering - DocEng '06*, 2006, p. 107.

[8] Y. Koren and R. Bell, "Advances in collaborative filtering," in *Recommender Systems Handbook, Second Edition*, 2015, pp. 77–118.

[9] J. Canny, "Collaborative filtering with privacy," in *Proceedings - IEEE Symposium on Security and Privacy*, 2002, vol. 2002–January, pp. 45–57.

[10] J. M. Yang and K. F. Li, "An inference-based collaborative filtering approach," in *Proceedings - DASC 2007: Third IEEE International Symposium on Dependable, Autonomic and Secure Computing*, 2007, pp. 84–91.

[11] J. Fürnkranz *et al.*, "Mean Absolute Error," in *Encyclopedia of Machine Learning*, 2011, pp. 652–652.

[12] W. B. Langdon, J. Dolado, F. Sarro, and M. Harman, "Exact Mean Absolute Error of Baseline Predictor, MARP0," *Inf. Softw. Technol.*, vol. 73, pp. 16–18, 2016.

[13] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean Absolute Percentage Error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, 2016.