Analysis of User-based and Item-based Prediction Algorithms for Recommender System

Chit Hnin Wai, Khin Thanda Soe
University of Computer Studies, Yangon.
Libra.chw@gmail.com, ktdsoe@gmail.com

Abstract

Recommender Systems typically use techniques from collaborative filtering which recommend items that users with similar preferences have liked in the past and, also predict new rating by averaging ratings between pairs of similar users or items. Predictions come from three sources: predictions based on ratings of the same item by other users, predictions based on different item ratings made by the same user, and ratings predicted based on data from other but similar users rating other but similar items. In this system, we use prediction algorithms to provide users with items that match their interests based on collaborative filtering (CF) approach. Our system use similarity measures between users, and also between items from a single rating criteria .We provide analysis of user-based and item-based prediction algorithms. The accuracy of the algorithms is compared by Mean Absolute Error.

1. Introduction

Recommender systems have evolved in the extremely interactive environment of the web. Recommender systems help online customers avoid information overload by making suggestions regarding which information is most relevant to them [4].

Recommender systems form a specific type of information filtering (IF) techniques that attempts to present information items that are likely of interest to the user [9]. They apply data analysis techniques to the problem of helping customers find which products they would like to purchase at E-Commerce sites by producing a list of top-N recommended products for a given customer [2].

Section 2 describes a recommender system based collaborative filtering. Section 3 describes a set of similarity measures. Section 4 states prediction algorithms and Statistical Accuracy Metric. Section 5 presents the overview of the system and analysis of the prediction algorithms. Section 6 concludes the paper.

2. Recommender System based on Collaborative Filtering

The two basic entities in any Recommender System are the user and the item. The input to a Recommender System depends on the type of the employed filtering algorithm. But, the input belongs to one of the following categories: Ratings, Demographic data and Content data. The output of a recommender system can be either a prediction or a recommendation [3].

In our system, we use input as rating to give by user. Ratings, express the opinion of users on items. Ratings are normally provided by the user and follow a specified numerical scale. A Prediction is expressed as a numerical value, which represents the anticipated opinion of an active user for item. This predicted value should necessarily be within the same numerical scale provided initially by active user. A Recommendation is expressed as a list of N items, which the active user is expected to like the most. This list to include only items that the active user has not already purchased viewed or rated [3].

The basic idea of CF-based algorithms is to provide recommendations or predictions based on the opinions of other user. The opinions of users can be obtained explicitly from the users or by using some implicit measures.

Collaborative filtering systems recommend products to a target customer based on the opinions of other customers. These systems use statistical techniques to find a set of customers known as similar that have a history of agreeing with the target user. Once a similar user is formed, these systems used several algorithms to produce recommendations [1].

The Collaborative filtering algorithms that can be divided into two main categories-

(i) Memory based (User-based)
(ii) Model-based (Item-based)

(i) Memory-based Collaborative filtering Algorithms

Memory–based algorithms provide the entries user-item database to generate a prediction. These system use statistical techniques to find a set of users, know as neighbors, which have a history of agreeing with the target user.

(ii) Model-based Collaborative filtering Algorithms

Model-based algorithms provide item recommendation by first developing a model of user ratings as computing the expected value of a user prediction, given his/her ratings on other items[2].

3. Similarity measures

We compute a set of similarity measures to compare the relevance between users or items. And a set of similarity measures are presented based on the Pearson correlation coefficient, a metric of relevance between two vectors. When the values of these vectors are associated with a user's model then the similarity is called **user-based similarity**. Also, the values of those vectors are associated with an item's model is called **item-based similarity**. The similarity measures can be effectively used to balance the ratings in a prediction algorithm and to improve accuracy.

Pearson correlation is selected to compute item-based and user-based similarities taking advantage of both explicit and implicit rating. An explicit rating is defined as the preference of a user to a specific item. Rating ranges from 1 to 5 with 1 expressing greatest aversion to the item and 5 expressing greatest liking to the item. An implicit rating is defined as the preference of a user to specific categories [8]. Before describing the algorithms the following definitions are introduced to facilitate the explanation process:

- A set of m users U={u_x: x=1, 2,…,m};
- A set of n items I={i_x: x=1, 2,…,n};
- A set of p categories C={c_x: x=1, 2,…,p};
- A set of q explicit ratings R= {r_x: x: 1, 2,..., q∧q≤ m*n};
- A set of t implicit ratings R= { r'_x : x =1, 2,...,t∧t≤ m*p} ;
- The explicit rating of a user u_x with reference to an item i_h as r_{u_x, i_h}.
- The average explicit rating of a user u_x as $\bar{r}_{u_x}$.

3.1 User-based similarity

3.1.1 Based on explicit rating

If the set of items that user u_x and u_y have co-rated is defined as $I' = \{i_x : x = 1,2,...,n' \wedge n' \leq n\}$, where n is the total number of items in the database, then the similarity between two users is defined as the Pearson correlation coefficient of their associated rows in the user-item matrix and is given by Eq.(1)

$$\kappa_{x,y} = sim(u_x, u_y) = \frac{\sum_{h=1}^{n'} (r_{u_x,i_h} - \bar{r}_{u_x})(r_{u_y,i_h} - \bar{r}_{u_y})}{\sqrt{\sum_{h=1}^{n'} (r_{u_x,i_h} - \bar{r}_{u_x})^2} \sqrt{\sum_{h=1}^{n'} (r_{u_y,i_h} - \bar{r}_{u_y})^2}} \quad (1)$$

3.1.2 Based on implicit rating

The preference of a user $u_x \in U$ to the category $c_x \in C$ is inferred by the user-category matrix. An implicit rating $r'_{u_x,c_x} \in R'$ to that category is computed as $r'_{u_x,c_x} = (c_{Xpos} / c_{Xpos} + c_{Xneg})*5$ where c_{Xpos}, c_{Xneg} are, respectively, the number of positive and negative ratings that user u_x has implicitly given to category x. The similarity between the two users is defined as the Pearson correlation coefficient of their implicit ratings to all categories $c \in C$, given by Eq.(2) where p is the number of available categories.

$$\lambda_{x,y} = sim(u_x, u_y) = \frac{\sum_{h=1}^{p} (r_{u_h,i_x} - \bar{r}_{i_x})(r_{u_h,i_y} - \bar{r}_{i_y})}{\sqrt{\sum_{h=1}^{p} (r_{u_h,i_x} - \bar{r}_{i_x})^2} \sqrt{\sum_{h=1}^{p} (r_{u_h,i_y} - \bar{r}_{i_y})^2}} \quad (2)$$

3.2 Item-based similarity

3.2.1 Based on explicit ratings

If the set of users that have rated both items and is defined as $U' = \{u_x : x = 1,2,...,m' \wedge m' \leq m\}$,where m is the total number of users in database, then the similarity between two items is defined as the Pearson correlation coefficient of their associated columns in the user-item matrix and is given by Eq.(3)

$$\mu_{x,y} = sim(i_x, i_y) = \frac{\sum_{h=1}^{m'} (r_{u_h,i_x} - \bar{r}_{i_x})(r_{u_h,i_y} - \bar{r}_{i_y})}{\sqrt{\sum_{h=1}^{m'} (r_{u_h,i_x} - \bar{r}_{i_x})^2} \sqrt{\sum_{h=1}^{m'} (r_{u_h,i_y} - \bar{r}_{i_y})^2}} \quad (3)$$

3.2.2 Based on item-category bitmap

The similarity between two items is defined as the Pearson correlation coefficient of their associated rows in the item-category bitmap matrix and is given by Eq.(4),where p is the number of categories and v_{C_h,i_h} is a Boolean value that equals to 1 if the item x belong h or equals to 0 otherwise

$$v_{x,y} = sim(i_x, i_y) = \frac{\sum_{h=1}^{p} (v_{C_h,i_x} - \bar{v}_{i_x})(r_{C_h,i_y} - \bar{v}_{i_y})}{\sqrt{\sum_{h=1}^{p} (v_{u_h,i_x} - \bar{v}_{i_x})^2} \sqrt{\sum_{h=1}^{p} (v_{u_h,i_y} - \bar{v}_{i_y})^2}} \quad (4)$$

4. Prediction Algorithms

Prediction algorithms try to guess the rating that a user is going to provide for an item. This user will be referred as active user u_a and this item as active item i_a.

4.1 User-based prediction algorithms

User-based prediction algorithms are based on user's average rating and an adjustment to it, as given by

Prediction=user-average + adjustment.

The adjustment is most often a weighted sum that integrates user-based or item-based similarity measures. So, prediction arises as the sum of the two.

4.1.1 User-based with explicit ratings (CF_{UB-ER})

The prediction algorithm is given by Eq. (5), where m' is the number of users that have rated the item i_a and $\bar{r}_{u_a}$ is the user's average rating over the set of items that the active user has rated

$$CF_{UB-ER} = p_{u_a, i_a} = \bar{r}_{u_a} + \frac{\sum_{h=1}^{m'} k_{a,h}\left(r_{u_h, i_a} - \bar{r}_{u_h}\right)}{\sum_{h=1}^{m'} \left|k_{a,h}\right|} \tag{5}$$

4.1.2 User-based with explicit ratings and category boosted ($CF_{UB-ER-CB}$)

The prediction algorithm is given by Eq. (6), where m' is the number of users that have rated the item i_a and $\bar{r}_{u_a}$ is the user's average rating over the set of items that the active user has rated

$$CF_{UB-ER-CB} = p_{u_a, i_a} = \bar{r}_{u_a} + \frac{\sum_{h=1}^{m'} k_{a,h}\left(r_{u_h, i_a} - \bar{r}_{u_h}\right)}{\sum_{h=1}^{m'} \left|k_{a,h}\right|} \tag{6}$$

4.1.3 User-based with implicit rating (CF_{UB-IR})

The prediction algorithm is given by Eq. (7), where m' is the number of users that have rated the item i_a and $\bar{r}_{u_a}$ is the user's average rating over the set of items that the active user has rated

$$CF_{UB-IR} = p_{u_a, i_a} = \bar{r}_{u_a} + \frac{\sum_{h=1}^{m'} \lambda_{a,h}\left(r_{u_h, i_a} - \bar{r}_{u_h}\right)}{\sum_{h=1}^{m'} \left|\lambda_{a,h}\right|} \tag{7}$$

4.2 Item-based prediction algorithms

Item-based prediction algorithms refer to algorithms that are based on item's average rating and an adjustment to it, as given by

Prediction=item-average + adjustment

4.2.1 Item-based with explicit rating (CF_{IB-ER})

The prediction algorithm is given by Eq.(8), where n' is the number of items that the active user u_a has rated and $\bar{r}_{i_a}$ is the item's average rating based on all the rating based that have been submitted for it

$$CF_{IB-ER} = p_{u_a, i_a} = \bar{r}_{i_a} + \frac{\sum_{h=1}^{n'} \mu_{a,h}\left(r_{u_a, i_h} - \bar{r}_{u_a}\right)}{\sum_{h=1}^{n'} \left|\mu_{a,h}\right|} \tag{8}$$

4.2.2 Item-based with implicit rating (CF_{IB-IR})

The prediction algorithm is given by Eq.(9), where n' is the number of items that the active user u_a has rated and $\bar{r}_{i_a}$ is the item's average rating based on all the ratings that have been submitted for it

$$CF_{IB-IR} = p_{u_a, i_a} = \bar{r}_{i_a} + \frac{\sum_{h=1}^{n'} v_{a,h}\left(r_{u_a, i_h} - \bar{r}_{u_a}\right)}{\sum_{h=1}^{n'} \left|\mu_{a,h}\right|} \tag{9}$$

4.3 Statistical Accuracy Metric

We use accuracy metric to evaluate the accuracy of a prediction algorithm. As statistical accuracy measure, mean absolute error (MAE) is used. The MAE is defined as the average absolute different between the n pairs $\langle p_h, r_h \rangle$ of predicted ratings and the actual rating and is given by Eq. (10)

$$\text{MAE} = \frac{\sum_{h=1}^{n} \left|p_h - r_h\right|}{n} \tag{10}$$

5. Overview of the Proposed System

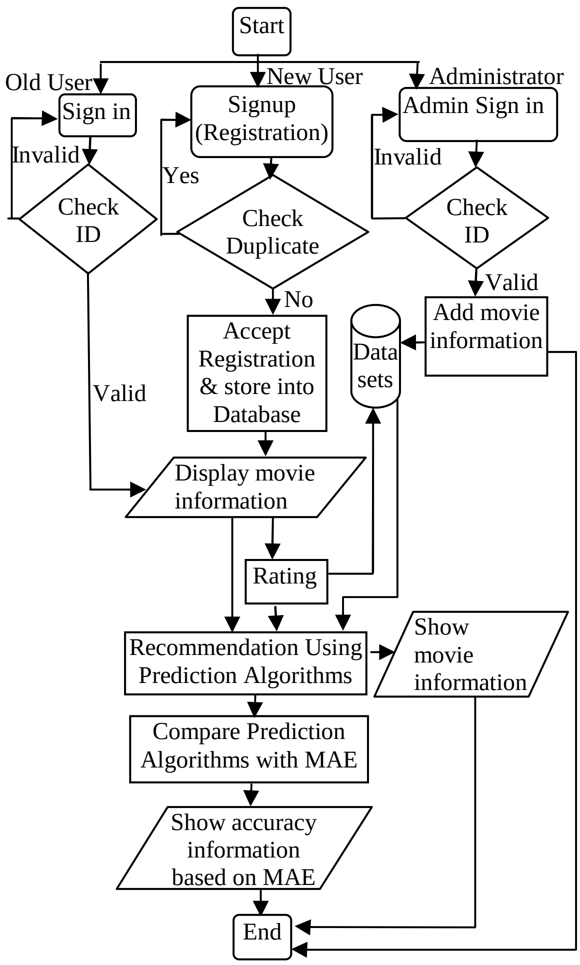

Figure 1.Overview of the Proposed System

In our system, there are three types of user: administrator, new user and old user. The administrator can add the new movie information. If the user is new, he must be registered in the system and the system will give the login UserID. If the user is old, he can sign in with UserID and if the UserID is valid, the user can access the system services. The system provides movie information for the user to give rating or not on each item and presents recommendation movies for user using different prediction algorithms and then shows the accuracy of different prediction algorithms by using Mean Absolute Error (MAE) as shown in Figure 1.

5.1 Analysis of the Prediction Algorithms

In this section, we will provide the analysis of user-based and item-based prediction algorithms using both sample data and Movie Lens Dataset. Firstly, we analyze the algorithms using sample data that contains 4 users (U1 to U4), 5movies (I1 to I5) and rating of each user on the items are shown in Table1.

Table 1 . User ratings to each movie item

Users	I1	I2	I3	I4	I5
U1	1	4	2	5	4
U2	4	2	5	2	?
U3	4	3	?	3	3
U4	2	4	3	?	2

Assume that the user U4 want to take recommendation from the system. There is an unknown rating on item I4. So, we are predicting the ratings of U4 on I4 by using prediction algorithms. First, we calculate similarity between active user and other users. So, Sim (U4, U1) is 0.38, Sim (U4, U2) is -0.5 and Sim (U4, U3) is -0.5 and calculate predicted ratings for user 4. The predicted results computed with different algorithms are shown in Table 2.

Table 2. Predicted result computed by different Algorithms

Prediction Algorithms	Collaborative Filtering				
	UB-ER	UB-ER-CB	UB-IR	IB-ER	IB-IR
P(U4,I4)	3.78	3.69	2.77	3.56	3.18

Next, the recommended movies are given to U4 based on the ratings on items as shown in Table 3.

Table 3. Recommended movie for U4

Prediction Algorithms	Collaborative Filtering				
	UB-ER	UB-ER-CB	UB-IR	IB-ER	IB-IR
Movie Title	I2	I2	I2	I2	I2
	I4	I4	I3	I4	I4
	I3	I3	I4	I3	I3

The system compare the accuracy of prediction algorithms by Mean Absolute Error (MAE).Table 4 shows the accuracy of prediction algorithms.

Table 4. Statistical Accuracy of Prediction Algorithms

Algorithms	Mean Absolute Error
CF_{UB-ER}	0.734
$CF_{UB-ER-CB}$	1.20
CF_{UB-IR}	1.30
CF_{IB-ER}	0.454
CF_{IB-IR}	1.002

We use Mean Absolute Error (MAE) in which the lower the MAE value, the more accurate the predictions. Therefore, the lowest MAE value predicted with item based explicit rating (CF_{IB-ER}) is 0.454. Second, lower MAE value is 0.734 in user-based explicit rating (CF_{UB-ER}).Third; lower MAE value is 1.002 in item-based implicit rating (CF_{IB-IR}) and so on. So, using CF_{IB-ER} provides the most accurate recommendation than other algorithms.

Now, we analyze user-based and item-based prediction algorithms using Movie Lens Recommender system datasets with different users and different ratings that contain the information about the user in the u.user dataset, information about the movie in the u.item dataset, information of the user and rating in the u.data datasets. There are 50 users, 60 users, 70 users, 80 users and 100 users and 500 ratings, 600 ratings, 700 ratings, 800 ratings and 1000 ratings and1682 movies. The comparison of prediction algorithms in term of Mean Absolute Error are shown in Table 5 and U means users and r means ratings.

Table 5. Statistical Accuracy of the different prediction Algorithms

Prediction Algorithm	U-50 r-500	U-60 r-600	U-70 r-700	U-80 r-800	U-100 r-1000
CF_{UB-ER}	0.647	0.633	0.615	0.613	0.615
$CF_{UB-ER-CB}$	0.920	0.934	0.908	0.908	0.889
CF_{UB-IR}	0.903	0.896	0.887	0.892	0.882
CF_{IB-ER}	0.216	0.246	0.263	0.270	0.299
CF_{IB-IR}	0.512	0.5347	0.5595	0.5650	0.5967

The results show that the item-based prediction algorithms are more accurate than user-based prediction algorithms although the number of users and ratings is different.

6. Conclusion

The recommendation systems can predict user behavior patterns without any knowledge of the user in advance. User-based and Item-based prediction algorithms are used for the recommender system. In this paper, we analyzed these algorithms and the analysis results express that CF_{IB-ER} is the most accurate algorithm compared with others.

7. References

[1] B. Sarwar, G. Karypis, J. Konstan, and J. Ried, "Analysis of Recommendation Algorithms for E-Commence", In Proceedings of the ACM EC'00 Conference, 2000, Minneapolia, Minnesto, pp158-167.

[2] B. Sarwar, G. Karypis, J. Konstan, and J. Ried, "Item-based Collaborative filtering Recommendation Algorithms", In Proceedings of the WWW10 Conference, 2001, HongKong.

[3] E. Vozalis, K. G.Margatistis, "Analysis of Recommender System's Algorithms", Recommender System Conference.

[4] G. Adomavicius and YoungOk Kwon, University of Minnesota. "New Recommendation Techniques for Multicriteria Rating Systems", Published by the IEEE Computer Soecity, IEEE Intelligent System, 2007.

[5] http://www.GroupLens.org.

[6] J. Wang, A. P.de Vries, M. J.T.Reinders, "Unifying User-based and Item-based Collaborative filtering Approaches by Similarity Fusion", In Proceedings of the ACM Confrecnce, 2006, Seattle, Washington, USA.

[7] M. Papagelis, D. Plexousakis, "Qualitative analysis of user-based and item-based prediction algorithms for recommendation agents", Engineering Applications of Artificial Intelligent, 2005, pp 781-789.

[8] N. Leavitt, "Recommendation Technology: Will it Boost E-commerce?", Published by the IEEE Computer Society, May 2006.

[9]Recommender System, http://www.wikipedia.org.