BUILDING A RECOMMENDATION SYSTEM FOR ONLINE SHOPPING BASED ON ITEM-BASED COLLABORATIVE FILTERING

Tessy Badriyah, Ira Prasetyaningrum, Basik Adhi P., Iwan Syarif

Department of Information Technology,
Electronic Engineering Polytechnic Institute of Surabaya (EEPIS)
http://www.pens.ac.id

Correspondence email: tessy@pens.ac.id

Abstract

This research applied an innovation in developing online shopping, using recommendation system. Recommendation System applies finding knowledge technique which is called item-based Collaborative Filtering. This works with by building information about items that are preferred by the customers. Collaborative Filtering filters data based on similarities or certain characteristics, so that the system is able to provide information based on patterns from a certain group of data that are almost the same.

With recommendation system, customers could benefit from the recommended items which they may favour, generated automatically by the system. It is hoped that it could improve the convenience to shop and reduce the time needed by customers to search for items. Therefore it could increase the competitiveness of online shops that use a recommendation system.

Keywords: ***online shopping, recommendation system, item-based collaborative filtering***

Presenting First Author's biography

Tessy Badriyah graduated from Portsmouth University, United Kingdom with a Doctor of Philosophy in Computer Science with a specialization in Health Informatics in September, 2013.

Her current research focuses on knowledge discovery for different purpose.

1. Introduction

With the growth of online shops that offer convenience for customers to shop, the greater is the competition across online shops to provide service that give ease for customers to find their desired items

It is not easy for customers to be able to find their desired items, since various items are offered with different specifications. This calls online shops to rightly notice, which products may be favoured by customers who visit their websites. This cannot be done by a marketing staff in real life, because a 24 hours system is needed to stand by, ready to provide recommendations for products that may be favoured by a customer.

This research builds a recommendation system for online shops using the Collaborative Filtering method. Collaborative Filtering filters data based on similar characteristics from customers. Therefore this enables it to provide new information based on the patterns of a group of customers with similarities. Collaborative Filtering technique that is implemented in this research uses an open dataset movielens from http://files.grouplens.org/datasets/movielens/ml-100k/ and uses slope one method that was proposed by Daniel Lemire and Anna Maclachlan (2005). This method is included in item-based collaborative filtering category, which filters and evaluates items that are offered in online shops using preferences from other customers.

This paper composes of discussions as follows: in Chapter 2, several relevant researches would be reviewed. Then Chapter 3 describes methods that will be used. Chapter 4 will discuss the experiment result and analysis. Lastly, Chapter 5 discusses the conclusion from the research that has been done.

2. Related Works

Implementation from Collaborative Filtering was first done by (Goldberg, Nichols, Oki, & Terry, 1992). The system built was still revolved around small communities that have limited scope. However, for a wider community, we cannot rely on each individual in those communities to have already known each other. Subsequently from that advancement, a rating-based recommendation system was developed akin to those by GroupLens research system (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994), (Konstan et al., 1997)], which provides collaborative filtering solution for news websites and films.

Subsequently are the technology that are applied on recommendation system, including Bayesian networks and clustering. Bayesian networks create model-based data training with decision tree for ever node and branches that interprets user information. The modelling generated are very compact, fast and accurate. (Breese, Heckerman, & Kadie, 1998). Meanwhile clustering technique on recommendation system identifies user groups that have similar preferences. After the clusters are created, predictions from an individual could be made by searching average opinions from others. The result of the experiment stated that the implemented clustering technique gives recommendation that is less personal and in some cases, nearest neighbour algorithm gives a better accuracy.

From several alternative methods that can be used for recommendation system, this research raises weighted slope one algorithm proposed by Daniel Lemire and Anna Maclachlan (2005). Weighted slope one method effectively manages sparse data caused by too few users who give ratings than the existing items.

3. The Methods Used

The following will explain how weighted slope one algorithm could solve recommendation system issues and the use of mean absolute error (MAE) to measure performance from the results collected.

A. Weighted Slope One

Weighted slope one calculates item similarity value, added by item rating times by the number of users who chose item. The said total sum is divided by the total user who give rating for item, therefore:

$$P^{wS1}(u)_j = \frac{\sum_{i \in S(u)-\{j\}} (dev_{j,i} + u_i) c_{j,i}}{\sum_{i \in S(u)-\{j\}} c_{j,i}} \quad (1)$$

Where :

$P^{wS1}(u)_j$ is a predicted weighted slope one for user u

$\boldsymbol{dev_{j,i}}$ is deviated value of item j to item i.

$\boldsymbol{u_i}$ is user rating u for item i

$\boldsymbol{c_{j,i}}$ is the number of users who rated item j

After receiving every prediction ratings for users, the next step is showing recommended users by using top-N item, which is item with highest predicted value, used as a recommendation for user.

The precision of the prediction value for user could be obtained using Mean Absolute Error (MAE).

B. Measuring performance using Mean Absolute Error

Mean Absolute Error calculates the average sum of error from absolute difference prediction value (p) towards the true value (a). Prediction value is obtained from one value towards the removed database. The erased value is kept as an actual value. The smaller MAE value that is generated, the better is the value precision [10].

who chose item. The said total sum is divided by the total user who give rating for item, therefore:

$$MAE = \frac{\sum_{i=1}^{N} |p_i - a_i|}{N} \quad (2)$$

Where :

MAE is the average value of miscalculation.

N is the sum of calculated item

p; is the predicted value to item i

a; is the true rate value to item i

4. Experiment Results and Analysis

This section will discuss the result of conducting experiment and the analysis.

A. Experiment and Analysis towards dummy data

The following experiment illustrate the using of dummy data about the working of weighted slope one algorithm. Dummy data used consist of 4 (four) user ((Reza, Angga, Claudia, Desi)

Tab. 1 Rating Data

	Ayat-ayat Cinta	The Raid	Comic 8
Reza	4	3	4
Angga	0	2	0
Claudia	0	3.5	4

Desi	5	0	3

As explained before, the similarity of item calculated by the rating value given by user on both items. For example, for two items, raid and comic 8, user who effect the similarity between item is Reza and Claudia. Rating value of raid and comic 8 from Reza and Claudia are 3,4,3.5 and 4, respectively.

$$dev_{i,j} = \sum_{u \in S_{i,j}(X)} \frac{u_i - u_j}{card(S_{i,j}(X))} \quad (3)$$

$$dev_{comic8,theraid} = \frac{\text{selisih rating reza}}{2} + \frac{\text{selisih rating claudia}}{2}$$

$$dev_{comic8,theraid} = \frac{4-3}{2} + \frac{4-3.5}{2}$$

$$dev_{comic8,theraid} = 0.5 + 0.25 = 0.75$$

The similarity value from Comic 8 and the Raid is 0.75. Then the similarity values among other items are calculated, therefore we obtain the similarity item as shown in table 2.

Tab. 2The similarity between item

	Ayat-ayat cinta	The Raid	Comic 8
Ayat-ayat cinta	0	1	1
The Raid	-1	0	-0.75
Comic 8	-1	0.75	0

After the similarity values between items are obtained, the predictions from the item that have not been rated by users are calculated using weighted slope one scheme. For example, the prediction for the user Angga who have not given a rating for Ayat-ayat cinta. From the dataset, it is known that Angga has only rated the Raid with 2 points.

$$P^{wS1}(u)_j = \frac{\sum_{i \in S(u)-\{j\}} (dev_{j,i} + u_i) c_{j,i}}{\sum_{i \in S(u)-\{j\}} c_{j,i}} \quad (4)$$

$$P^{wS1}angga_{aac} = \frac{\left((dev_{acc.acc} + u_{acc})c_{acc.acc}\right) + \left((dev_{acc.tr} + u_{tr})c_{acc.tr}\right) + ((dev_{cdk.c8} + u_{c8})c_{cdk.c8})}{c_{acc.acc} + c_{acc.tr} + c_{acc.c8}}$$

$$P^{wS1}angga_{acc} == \frac{\left((0+0)*0\right) + \left((1+2)*1\right) + \left((1+0)*2\right)}{3}$$

$$P^{wS1}(u)_j = \frac{0+3+2}{3} = 1.667$$

From the calculation, we can predict that Angga's rating for Ayat-ayat cinta is 1.667. After we obtained the predicted value from Angga for Comic 8, as shown in table 3, the recommended item could be identified based on top N item. Where the higher the value of prediction, the more recommended an item will be.

Tab. 3Rating prediction for user Angga

Item	Rating prediction
Ayat-ayat cinta	1.667
Comic 8	0.875

After all of the ratings of the items are fulfilled as shown in table 4, the accuracy from the prediction is calculated using MAE. The value from the rating data is replaced one by one to calculate the prediction value again. The values that have been replaced will be used as a guide to calculate the margin of error from the prediction value.

Tab. 4Full Rating Data

	Cinta dalan Kardus	The Raid	Comic 8
Reza	4	3	4
Angga	1.667	2	0.875
Claudia	4.833	3.5	4
Desi	5	2.833	3

Before calculating the prediction value for Ayat-ayat cinta rating from user Reza that is being removed, we also calculate the similarity values between the items. Table 5 shows the deviation value from this situation.

Tab. 5Deviation values

	Ayat-ayat cinta	The Raid	Comic 8
Ayat-ayat cinta	0	1.05567	1.20833
The Raid	-1.0557	0	-0.1355
Comic 8	-1.2083	0.1355	0

The calculation of Ayat-ayat cinta predicted values for Reza is as follows:

$$\mathrm{P}^{\mathrm{wS1}}(\mathrm{u})_{\mathrm{j}} = \frac{\sum_{\mathrm{i}\in \mathrm{S(u)}-\{\mathrm{j}\}}(\mathrm{dev}_{\mathrm{j.i}}+\mathrm{u}_{\mathrm{i}})\mathrm{c}_{\mathrm{j.i}}}{\sum_{\mathrm{i}\in \mathrm{S(u)}-\{\mathrm{j}\}}\mathrm{c}_{\mathrm{j,i}}} \tag{5}$$

$$\mathrm{P}^{\mathrm{wS1}}\mathrm{reza}_{\mathrm{aac}} = \frac{\left((\mathrm{dev}_{\mathrm{aac.aac}} + \mathrm{u}_{\mathrm{aac}})\mathrm{c}_{\mathrm{aac.aac}}\right) + ((\mathrm{dev}_{\mathrm{aac.tr}} + \mathrm{u}_{\mathrm{tr}})\mathrm{c}_{\mathrm{aac.tr}}) + ((\mathrm{dev}_{\mathrm{aac.c8}} + \mathrm{u}_{\mathrm{c8}})\mathrm{c}_{\mathrm{aac.c8}})}{\mathrm{c}_{\mathrm{aac.aac}} + \mathrm{c}_{\mathrm{aac.tr}} + \mathrm{c}_{\mathrm{aac.c8}}}$$

$$P^{wS1}angga = \frac{\left((0+0)*0\right) + \left((1.05567+3)*3\right) + \left((1.20883+4)*3\right)}{3}$$

$$P^{wS1}(u)_j = \frac{0 + 12.167 + 15.625}{3} = 4.632$$

Ayat-ayat cinta predicted value for user Reza is 4.632 while the actual value is 4 and therefore the absolute value is 0.632. This is done for the entire user rating so that we know the actual and predicted values from the dataset as shown in table 6.

Tab. 6Deviation values

Username	**Actual**	**Prediksi**	**Absolut Eror**	***Item***
Reza	4	4.635	0.635	Ayat-ayat Cinta
Reza	3	3.545	0.545	The Raid
Reza	4	2.820	1.180	Comic 8
Angga	1.667	2.658	0.991	Ayat-ayat Cinta
Angga	2	0.241	1.759	The Raid
Angga	0.875	1.643	0.768	Comic 8
Claudia	4.833	4.685	0.148	Ayat-ayat Cinta
Claudia	3.5	3.942	0.442	The Raid
Claudia	4	3.707	0.293	Comic 8
Desi	5	3.522	1.478	Ayat-ayat Cinta
Desi	2.833	3.60	0.767	The Raid
Desi	3	3.712	0.712	Comic 8

Table 6 shows the absolute error value from each items, so that afterwards we can obtain the accuracy from the rating predictions.

$$MAE = \frac{\sum_{i=1}^{N} |p_i - a_i|}{N}$$

$$MAE = \frac{9.718}{12} = 0.809$$

The calculations show the accuracy of the predicted rating value for the overall data in table 4 is 0.809.

B. Testing and Analysis towards Grouplens dataset[8]

The scenario testing from Grouplens dataset [8] was taken from the following link:

http://files.grouplens.org/datasets/movielens/ml-100k/

The scenario testing includes: testing MAE values to the amount of data, the comparison of manual calculation and testing of the similarity of users.

MAE testing to the amount of the data aims to determine the relationship between the accuracy of prediction value through MAE value to see the influence of the number of rating data involved. This test used the program as shown in the picture 4.32. In the program that we developed, the ongoing process can be identified and also the users and item that are involved. At the end of the program the MAE values are shown.

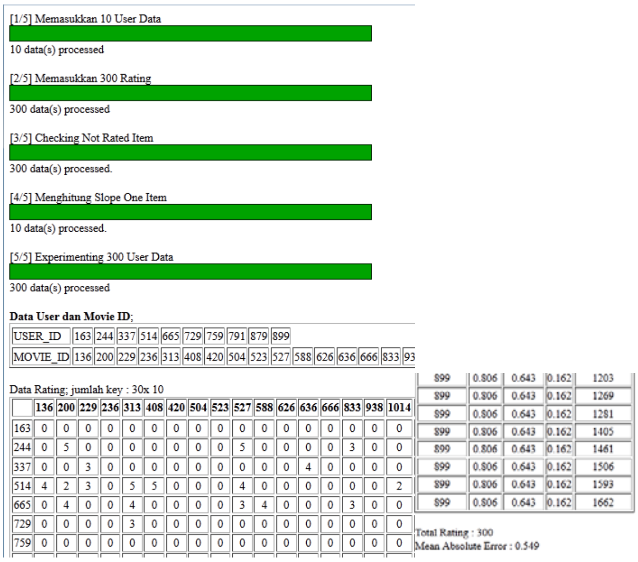

Fig. 1 Piece of program of MAE testing

The experiment involved users and item randomly with the various number of data that has been determined before. The experiment is done ten times for every variety of the amount of data.

Tab. 7MAE testing on various number of data

		Number of Data				
		150	**300**	**600**	**750**	**900**
N Experiments	**1**	0.506	0.699	0.687	0.624	0.599
	2	0.544	0.683	0.701	0.595	0.732
	3	0.601	0.799	0.593	0.523	0.716
	4	0.384	0.761	0.663	0.6	0.67
	5	0.912	0.71	0.668	0.632	0.651
	6	0.87	0.572	0.572	0.662	0.651
	7	0.802	0.491	0.618	0.537	0.594
	8	0.726	0.837	0.736	0.703	0.594
	9	0.87	0.694	0.691	0.692	0.598
	10	0.766	0.727	0.708	0.669	0.664
	Average	**0.6981**	**0.6973**	**0.6637**	**0.6237**	**0.6469**

Based on table 7, the MAE experiment of the use of various data. The amount of data 750 has the smallest average MAE value compared to other sum of data. The average trend of MAE tends to decrease. This shows that the more rating values are included, the better the predicted accuracy will be.

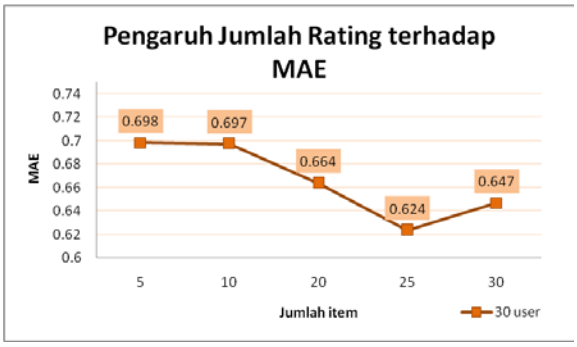

Fig. 2 The average of number of rating towards MAE

Testing manual calculation uses four processes, initialisation and calculation of the similarity values between items, calculating a particular predicted user rating and sorting the top N items from recommendation result.

The following explain the detail of four processes:

1. Initialisation. By using a pieve of four users' ratings for three items (Big Hero 6, American Sniper and Unbroken) to table 8. Which recommendation will be given to user Daniel?

Tab. 8Rating table on the manual experiment

	Big Hero 6	**American Sniper**	**Unbroken**
Alex	4	3	0
Christi	0	5	4
Eka	4	0	3
Daniel	0	0	3

2. By calculating the similarity between items using formula (1) we can take the similarity value between items on the rating table.

Tab. 9The similarity table on the manual experiment

	Big Hero 6	**American Sniper**	**Unbroken**
Big Hero 6	**0**	**1**	**1**
American Sniper	**-1**	**0**	**1**
Unbroken	**-1**	**-1**	**0**

3. Calculating prediction value of specific user. One of the users has been chosen, for example Daniel, who has two items that have not been filled. Therefore prediction ratings need to be given to those particular items. The following is the calculation result using weighted slope one scheme for user Daniel.

Tab. 10Prediction rating for user Daniel

	Big Hero 6	**American Sniper**	**Unbroken**
Daniel	2.5	1.5	3

4. Sorting Top N item from the recommendation. After sorting has been done for prediction rating for Daniel, we found that Big Hero 6 is the most recommended item, followed by American Sniper.

Tab. 11The sorting result of prediction value

Title	Rating
Big Hero 6	2.5
American Sniper	1.5

Testing of calculation using computer is made by the same process using manual calculation. After that from these two different tests (manual and computer) a comparison will be made to find out whether there are similarities.

1. Initialisation. The following Fig 3 shows the initialisation multi-dimensional data from one of the users, Daniel. We are calculating the prediction value.

```
$data_misbar_user = array(
        0 =>'Alex',
        1 =>'Christi',
        2 =>'Eka',
        3 =>'Daniel',
);

$data_misbar = array(
    "Big Hero 6" => array(
        0 =>'4',
        1 =>'0',
        2 =>'4',
        3 =>'0',
    ),
    "American Sniper" => array(
        0 =>'3',
        1 =>'5',
        2 =>'0',
        3 =>'0',
    ),"Unbroken" => array(
        0 =>'0',
        1 =>'4',
        2 =>'3',
        3 =>'3',
    ),
);
```

```
//User yang akan dicari
$UserFind="Daniel";
```

Fig. 3The Initialization of Testing Program

2. Program is run to calculate the similarity value. The similarity value between item for the data in table 8 is shown in Fig 4.

Model Deviasi

	Big Hero 6	American Sniper	Unbroken
Big Hero 6	0	1	1
American Sniper	-1	0	1
Unbroken	-1	-1	0

Fig. 4The similarity result from the program

3. After calculating the prediction value for item that have not been rated, we determine the Top N Item. Big Hero 6 becomes the movie that is most recommended, followed by American Sniper.

Rekomendasi Pengguna Daniel untuk Big Hero 6 : 2.5
Rekomendasi Pengguna Daniel untuk American Sniper : 1.5

Fig. 5Top N items as recommendation

The result of manual calculation and capture program show the value and recommendation of Top N that is the same

5. Conclusion

In this research we are implementing knowledge discovery technique known as Collaborative Filtering with Slope One method for recommendation system. The experiment result shows that system could calculate the item similarity value and determines the rating prediction for items that have not yet been rated. From the analysis it was found that the reduction of partial data could increase the value precision from the prediction result.

The overall experiment result and analysis have shown that algorithm for recommendations have successfully been implemented to the system. With the presence of this recommendation system, every customer could quickly obtain the recommended items which they prefer. This could improve the competitiveness from online shops, compared to others who do not use recommendation system. The recommendation system facility could improve the competitiveness of online shops to give service that satisfied their customers.

References

[1] Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. Paper presented at the Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, Madison, Wisconsin.

[2] Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. Commun. ACM, 35(12), 61-70. doi:10.1145/138859.138867

[3] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). GroupLens: applying collaborative filtering to Usenet news. Commun. ACM, 40(3), 77-87. doi:10.1145/245108.245126

[4] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews. Paper presented at the Proceedings of the 1994 ACM conference on Computer supported cooperative work, Chapel Hill, North Carolina, USA.

[5] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. Paper presented at the Proceedings of the 10th international conference on World Wide Web, Hong Kong, Hong Kong.

[6] Schafer, J. B., Konstan, J., & Riedl, J. (1999). Recommender systems in e-commerce. Paper presented at the Proceedings of the 1st ACM conference on Electronic commerce, Denver, Colorado, USA

[7] Lemire, Daniel and Maclachlan, Anna. Slope One Predictors for Online Rating-Based Collaborative Filtering. California : SIAM Data Mining (SDM'05), 2005.

[8] GroupLens. datasets. grouplens. [Online] 2014. http://grouplens.org/datasets/movielens/.

[9] Zacharski, Ron. A Programmer's Guide to Data Mining : The Ancient Art of the Numerati. guidetodatamining.com. [Online] 2012. http://guidetodatamining.com/