ISBN: 978-1-4799-6526-7

Film Recommendation Systems using Matrix Factorization and Collaborative Filtering

Mirza Ilhami
Informatics Technology Department
Faculty of Informatics Technology
STMIK Mikroskil
Medan, Indonesia
mr.ilhami@gmail.com

Suharjito
Magister in Information Technology
Binus Graduate Program
Bina Nusantara University
Jakarta, Indonesia
harjito@yahoo.com

***Abstract* — Collaborative filtering method was widely used in the recommendation system. This method was able to provide recommendations to the user through the similarity values between users. However, the central issues in this method were new user issue and sparsity. This paper discusses about how to use matrix factorization and nearest-neighbour in film recommendation systems. Both of methods will be used in order to make more accurate recommendations. Based on the experiments results, the combination of matrix factorization and classical collaborative filtering (nearest neighbor) could improve the prediction accuracy. It can be concluded that the combination of matrix factorization and nearest-neighbor produced a better prediction accuracy**

Keywords— Recommendation Systems; Matrix Factorization; Collaborative Filtering.

I. INTRODUCTION (*HEADING 1*)

Recommendation systems has been rapidly growing due to huge data informations available. We need a personalized systems that can be match to us, based on what we read recently, our last activities, and how its relevance to us. Recommendation systems become more popular today and has been used in various fields.

Collaborative filtering has been the mostly used in recommendation systems nowaday. This method was able to give recommendation to user based on their similarity to others. On the other hand, matrix factorization has also been familiar since Netflix Grand Prize [*1*].

Generally, recommendation system can be divide into two parts, the Content-based Filtering and Collaborative Filtering. Content-based filtering based on the similarity of the items to the object that the user liked in the past [*2*]. Meanwhile, Collaborative filtering is the process of evaluation using the information of user behavior or the behavior of other users [*3*].Collaborative filtering grouped by memory-based and model-based.

Some of the central issues in the collaborative filtering is sparsity, content analysis, overspecialization and new user issue (cold start problem) [*4*]. Hybrid method was used in order to solve this problems. Hybrid is a combination between several methods. This method was also a key that brought Bellkor team as the winner in Netflix Prize in 2009 [*5*].

Based on [*6*], matrix factorization can be used to increase the prediction accuracy in scalable dataset. Other than, matrix factorization also can be used to remove the dimension of the item space and retrieve latent relations between items of the dataset.

For experiments purpose, Movielens Hetrec 2011 rating data will be used in this paper. The rating data will be used for computation in recommendation and prediction algorithm.The combination of matrix factorization and classic collaborative filtering (nearest-neighbour) technique will be used to improve the prediction accuracy.

Preliminary section 3 discusses how the dataset will be used. In section 4, we describe the mathematical approach of our recommender. Prediction accuracy was evaluated in section 5. Finally, the conclusions in section 6.

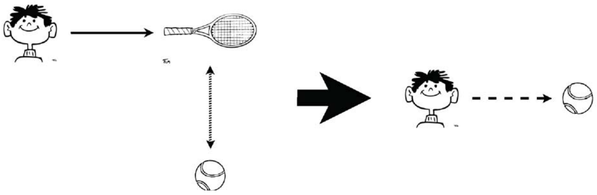
II. RELATED WORK

In this section we briefly review the main works in the context. The list of references is not exhaustive due to the page limit. Our main focus was in collaborative filtering which has been successfully applied to several real world problems, such as Netflix's movie recommendation.

A. Content-based Filtering

Content-based method is one of the oldest methods and most popular in the recommendation. The principle of this method is recommend object that has similarities to some other object, the user preferred in the past [*2*]. The similarity between objects was determined from the values of the characteristics of the object.

Fig. 1 shows an example of a content-based filtering. Well-known relationship marked by full arrows, calculations or similarity of objects marked with an arrow point and predictive relationship marked by dotted arrows.



978-1-4799-6527-4/14/$31.00 ©2014 IEEE

Fig. 1: Content-based filtering [*4*]

This method was widely used in text-based applications, to recommend a document or web site, according to [*4*]. One implementation of a content-based is the Music Gnome Project and WebMate.

B. Collaborative Filtering

Collaborative filtering was one of the most widely used techniques for recommendation systems. Companies like Google [*7*], Amazon [*8*] using the recommendation, at least in part based on collaborative filtering. CFwas a popular recommendation algorithm that bases its predictions and recommendations on ratings or opinions of other users in the system [*3*]. Typically, predictions about user interests were obtained by collecting taste information from many other similar users. The fundamental assumption of this method was that the tastes of other users can be selected and assembled in such a way to give a reasonable prediction of the active user's preference. CF requires a large amount of information to be processed, including large-scale data sets such as the e-commerce and web applications. Over the past few decades CF has been continuously improved and has become one of the most prominent personalization techniques in the field of recommendation systems.

Fig. 2. show an example of a collaborative filtering. Full-line arrows represent the relationships were already known. Both subjects had ties with the goods and therefore they were considered to have similarities. Recommendations marked with dotted arrows. An object can be recommended for selected of objects related to the same subject, so the subject was not related to the object.

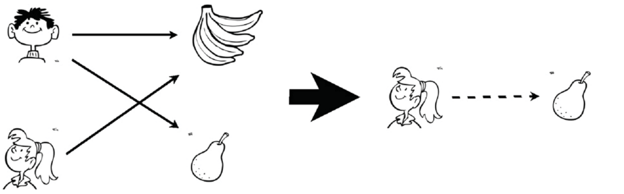


Fig. 2:Collaborative filtering [*4*]

Based on the approach of unknown relationships, collaborative filtering can be divided into two groups: a memory-based and model-based.

C. Hybrid Recommendation Method

Hybrid method combining two or more methods of recommendation with the aim to improve recommendation results. There were a number of approaches to incorporate practical advice among collaborative filtering, content-based filtering, knowledge-based demographics and recommendations. The most popular was a combination of content-based filtering and collaborative filtering [*4*].

The combination of memory-based collaborative filtering and collaborative filtering-based models were often used in contemporary commercial recommendations.Fig. 3. show an example of hybrid method.

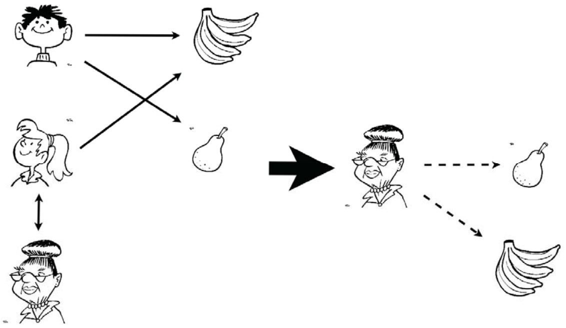


Fig. 3:Hybrid Method [*4*]

Hybrid method can help overcome some of the drawbacks of the method recommendations, eg. new item that has not been rated by any user. Several methods can be used additional recommendations [*9*].

D. Matrix Factorization

Collaborative filtering recommendation method has been widely used in the nearest neighbour (NNH) algorithm combined with the pearson correlation or cosine similarity to calculate the predicted values. In terms of sparse rating, NNH approach usually experience difficulties in finding the right match and consequently produce a weak recommendation with accuracy. Furthermore NNH calculation algorithm complexity tends to increase with number of users and the number of items, which means the recommendation system will have serious problems in scalability [*6*].

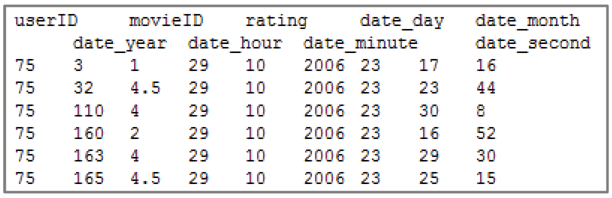
To overcome this problem, a matrix factorization approach has been proven to be an efficient method for rating-based recommendation [*10*].

III. METHODOLOGY

There were several things that need to be done before the recommendation can be made. Firstly, it needsto extract the dataset, in order to process recommendations. This section will explain how the datawill be used, a hybrid recommendation techniques and evaluation methods.

A. Dataset Analysis

The Movielens dataset was used in this research, contains 855,598 ratings for 10197 movies and 2113 users. This dataset was an extension of Movielens10M dataset, published by Groupleans research group. It links the movies of Movielens dataset with their corresponding web pages at Internet Movie Database (IMDb)and Rotten Tomatoesmovie review systems. The dataset was released in HetRec 2011. The ratings data represent as a list of triples *userID*, *itemID* and *rating*. Split the dataset into two parts, 80% for training and 20% for testing the algorithm. The dataset contains actors, countries, directors and genres. For experiment purpose, it only use the data rating from users to make recommendations. Fig. 4. show how the data format look like.



```
userID      movieID     rating      date_day    date_month
      date_year   date_hour   date_minute       date_second
75    3     1     29    10    2006  23    17    16
75    32    4.5   29    10    2006  23    23    44
75    110   4     29    10    2006  23    30    8
75    160   2     29    10    2006  23    16    52
75    163   4     29    10    2006  23    29    30
75    165   4.5   29    10    2006  23    25    15
```

Fig. 4: Rating data format

Our main focus was on the *userID*, *movieID* and *rating*. Those field will convert into matrix two-dimensional space, first dimension was the number of users and second dimension is the number of films *($M_{2113\ x\ 10197}$)*. The valueswere represented with numeric. However, the result of matrix *M* will be sparse, mostly the user-item value was empty field. Thus, only the available rating data will used in training algorithm.

B. *Matrix Notation*

Our experiment willfocus on the film and the user entity. Movie data, and user ratings were available in different data formats. Each file consists of fields that were interrelated and can be extracted to form a matrix notation that will be used to make recommendation system. Table1. shows the matrix notation that will be used to recommendation system.

TABLE 1: Matrix notation

user/item	film1	film2	film3	n-films
user1	1	-	2	-
user2	2	3	-	-
user3	-	2	1	3
user4	5	4	-	1
user5	3.5	1.5	5	2
user6	2.5	2	3	1
n-user	1	3	-	1

From the table above, rows represented the number of users and columns represented the number of movies. Whereas, the cell represented the rating value that user given. In table above show the construction of the matrix. Then, that matrix will be used into recommendation system. The matrix dimension is number of users multiple number of movies. Using matrix factorization can find the missing values from matrix notation.

C. *Hybridization*

Various techniques have been proposed as a basis for recommendation systems: collaborative filtering, content-based filtering, knowledge-based screening and demographic techniques. Each of these techniques has its drawbacks, such as the cold-start problems in collaborative filtering and content-based filtering [*11*].

There a lot of technique that can be used in recommendation system, the mostly used was nearest-neighbour, item-based filtering [*8*] and matrix factorization [*12*]. Matrix factorization will be used in this paper. Then, several experiments to compare the prediction accuracy between matrix factorization and the the combination of matrix factorization and nearest-neighbour (NNH). Combining two or more technique was called hybrid [*11*]. The objective of hybrid was to increase the recommendation accuracy and reduce the error. Fig. 2. shows the architecture of hybrid.

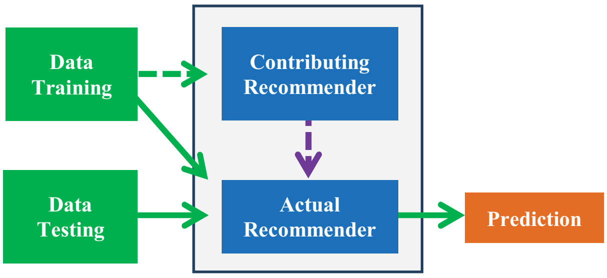


Fig. 5: Hybrid Recommendation System Architecture [*11*]

Fig. 5. shows the big picture of hybrid recommendation system works. The underlying from its architecture was data training and data testing. The ratio of data training and data testing were usually 80:20, however, some papers also use 70:30. Both of the data will be used to produce the prediction value.

D. *Evaluation Method*

The recommendation systems accuracy depend on several factors. Such as, number of data, *k*-values, normalization, and many others. There were a lot of performance metrics that can be used in research. Generally, it will be used to evaluate and analyze the system performance. Some of metric tools to evaluate recommender system according to microsoft research were Root Mean Square Error (RMSE) and Mean Square Error (MAE) [20]. These metric will be used in our research, in order to computes how close the estimates were to the values actually observed. This experiment want to minimize the prediction Root Mean Square Error (RMSE) on test sets:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{r}_i - r_i)^2}$$

In equation above $\bar{r}_i$denotes the predicted rating and $r_i$were the actual rating of the movie. Whereas, the mathematical notation for MAE is:

$$MAE = \frac{1}{|T|}\sum_{(u,i)\in T}|\hat{r}_{ui} - r_{ui}|$$

From the equation above $\hat{r}_{ui}$ is prediction value from user *u* to item *i* dan $r_{ui}$is actual value.

IV. RECOMMENDER SYSTEM

In this section, the step-by-step of proposed method will be more details. It will describe more about the recommendation technique, data normalization, matrix factorization and its combination with classic collaborative filtering.

A. Normalization

The data rating from MovieLens, it was a range between rating. The range was from 1 until 5 (0, if no rating). The normalizationwas used to remove the bias. One common method of normalization involves having values of each feature range from 0 to 1.

In order to increase the prediction accuracy, the "global effect" has to be removed. Generally the combination of user, item and all average rating value ($\bar{r}$, $\bar{r}_u$&$\bar{r}_i$) were substracted from the original data $r_{ui}$to remove the popularity effect [*13*].

$$\tilde{r}_{ui} = r_{ui} - \alpha\bar{r} - \beta\bar{r}_u - \gamma\bar{r}_i$$

After normalize the data, then perform matrix factorization to get predictions and recommendations.

B. Matrix Factorization

In many cases, matrix factorization was used to remove the dimension of the item space and retrieve latent relations between items of the dataset [*12*], [*14*]. Based on the demonstration on Neflix Grand Prize, matrix factorization model was the best nearest-neighbour technique in recommendation system.

Recommendation systems have a large enough data and very sparse, this causes a slow process and computational predictions inaccurate. So therefore, it need special technique in order to minimize the spread of data and speed up the computation process, it can be done by matrix factorization [*12*]. There were many methods used to process factorize the matrix, such as Singular Value Decomposition (SVD), Non-negative Matrix Factorization (NMF), and others. SVD approach will be used in this paper, it factorize matrix rating to be three low dimensional space, left-singular vector (V), singular value (S) and right-singular vector (W).

$$H_k = V_{mxk} \,.\, S_{kxk} \,.\, W_{kxn}^T$$

Multiple k examples were used to make the experiment and investigate the result for next experiment.

C. Predictions

The comparisonof prediction value and original rating were performed in order to get the evaluation about how accurate the algorithm works. To predict rating r_{ui}, SVD class reconstructs the original matrix

$$M' = U\textstyle\sum_k V^T$$

And the rating prediction equals to:

$$rating\ (u,i) \;=\; M'ij$$

A rating prediction approximation will very close to the original rating values, and it also deliver some predictions of the unknown values or missing values. Neighbourhood algorithm uses the ratings of the similar users (or items) to predict the values of the input matrix. The only difference with plain SVD is the way how it computes the predictions. To compute the prediction, it uses this equation:

$$rating\ (u,i) = \frac{\sum_{j\in S^k(i;u)} S_{ij} r_{uj}}{\sum_{j\in S^k(i;u)} S_{ij}}$$

Where S^k *(i; u)* is the set of k that rate by u, which were most similar to i. S_{ij} is the similarity between i and j.

D. Nearest Neighbour

Another method that will be used in this paper was the nearest neigbour approach. Nearest neighbor was a classical collaborative filtering technique that still used in recommendation systems. This technique will predict the value and compare it with matrix factorization. Performed by this technique was to find the similarity to the films or users. From the experiments carried out in section 5.2 proves that the combination of matrix factorization and nearest neighbor can improve the prediction accuracy for better recommendation system.

V. EXPERIMENTS

In this section we carried out tests on datasets that have been mentioned above. Tests performed several times on the value of k to determine the accuracy of the results of the methods that have been mentioned in section IV.

A. MF Experiment

The most difficult thing to get better accuracy was to determine the optimal k values as paramater in computation. Therefore, multiple test by iterating the k valueswere performed.

As ilustrated in Fig. 4, the prediction accuracy was strongly depend on the k values.

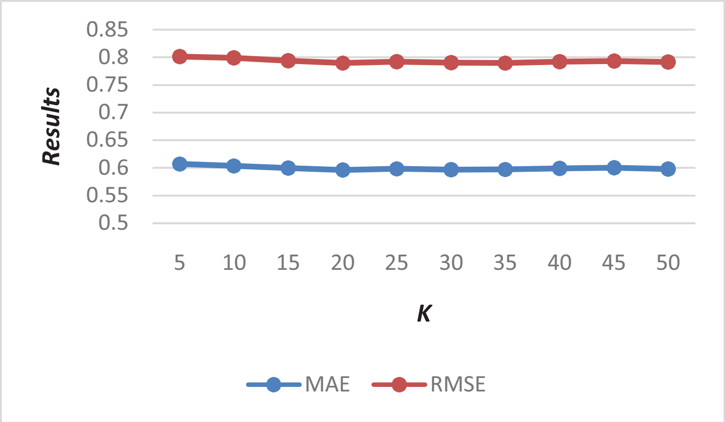


Fig. 6: MF in k-dimensional space

From Fig. 6.it shows that matrix factorization can result in an average prediction accuracy well.From several tested *k*, obtained the best of RMSE and MAE, respectively were 0,78 and 0,59 with *k* values were 20 and 35. Results metrics that close to 0 were the most good.

B. *MF + CF Experiment*

After getting the results of experiments with matrix factorization method, the next experiments will be performed by using the combination of matrix factorization and collaborative filtering.

It needs neighbourhood algorithm, which use the similarity values of user ratings to predict the value of the input matrix. The prediction accuracy resulting from experiment using MF + CF based on the value of *k* illustrated in the Fig. 5.

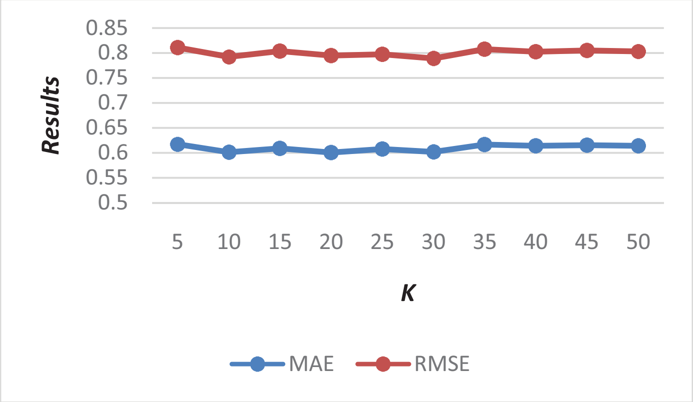


Fig. 7: MF+ CF in k-dimensional space

From the fig. 7. shows that the value of *k* = 30 was the lowest, meaning that *k* = 30 was the most optimum done in this case. When we compared with the matrix factorization (MF) in previous experiment, the combination of MF & CF has a better accuracy rate, were proven by the smaller RMSE value or close to 0. All the experimental results can be seen in the Table 4.

TABLE 2: Comparison of RMSE metrics for both methods

K	**MF**	**MF + CF**
5	0,800966	0.810545
10	0,798626	0.79244
15	0,793362	0.803481
20	0,789719	0.794638
25	0,791834	0.79704
30	0,789986	**0.788649**
35	0,**789649**	0.80757
40	0,79193	0.802361
45	0,792893	0.804966
50	0,791348	0.803322

Table 2. above show the experimental results of the comparison between two proposed methods. For the matrix factorization method, the smallest RMSE value was 0.789649, which *k* = 35. While for MF + CF method, the smallest RMSE value was 0.788649 *k* = 30.

VI. CONCLUSIONS

The combination of MF and CF can indeed be used to make predictions and recommendation systems. But it doest not have a huge if we compare it with the MF only. In experiment we have done, the difference were only 0.001.

Of course, there were many other parameters that can determine the accuracy of predictions, such as the number of datasets, content features, normalization, and others. With the existance of this paper we hope it can enrich the knowledge in the field of recommendation systems and can be used in the future for further experiments to improve the prediction accuracy.

ACKNOWLEDGMENT

For the finishing of this experiment, thank to Mr. Bens Pardamean, M.Sc, P.hD., for any criticsm and suggestions. And also thank to Grouplens for the dataset and for the rec-sys python framework, thanks to Ocelma.

REFERENCES

[1] Yehuda Koren. (2009) The BellKor Solution to the Netflix Grand Prize. [Online]. http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf

[2] Michael J Pazzani and Daniel Billsus, "Content-Based Recommendation Systems," *The Adaptive Web Lecture Notes in Computer Science*, vol. 4321, pp. 325-341, 2007.

[3] Michael D Ekstrand, John T Riedl, and Joseph A Konstan, "Collaborative Filtering Recommender Systems," *Foundations and Trends in Human–Computer Interaction*, vol. 4, no. 2, pp. 81 - 173, 2010.

[4] Gediminas Adomavicius and Alexander Tuzhilin, "Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Transaction On Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734-749, 2005.

[5] Yehuda Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2008, pp. 426-434.

[6] Yehuda Koren, "Factor in the Neighbors: Scalable and Accurate Collaborative Filtering," *ACM Transactions on Knowledge Discovery from Data*, pp. 1-24, 2010.

[7] Abhinandan Das, Mayur Datar, and Ashutosh Garg, "Google News Personalization: Scalable Online Collaborative Filtering.," in *International World Wide Web Conference Committee (IW3C2)*, 2007, pp. 271-280.

[8] Greg Linden, Brent Smith, and Jeremy York, "Amazon.com Recommendation: Item-To-Item Collaborative Filtering," 2003.

[9] Paolo Cremonesi, Roberto Turrin, and Fabio Airoldi, "Hybrid algorithms for recommending new items," in *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, New York, 2011, pp. 33-40.

[10] G Takács, I Pilászy, B Németh, and Domonkos Tikk, "Investigation of Various Matrix Factorization Methods for Large Recommender Systems," in *IEEE International Conference on Data Mining Workshops*, Pisa, 2008., pp. 553-562.

[11] Robin Burke, "Hybrid Web Recommender Systems," in *Vol. 4321 of Lecture Notes in Computer Science*. Berlin Heidelberg: Springer-Verlag, 2007, p. 377.

[12] Yehuda Koren, Robert Bell, and Chris Volinsky, "Matrix Factorization Techniques for Recommender Systems.," in *IEEE Computer Society.*, 2009, pp. 42-49.

[13] Upendra Shardanand and Pattie Maes, "Social information filtering: algorithms for automating "word of mouth"," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, 1995, pp. 210-217.

[14] Xiaoyuan Su and Taghi M Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," *Journal Advances in Artificial Intelligence*, pp. 1-19, 2009.

.