Advances in Engineering Research, volume 208
Proceedings of the International Conference on Innovation in Science and Technology (ICIST 2020)

Improve Quality of Recommendation System Using Hybrid Filtering Approach

Annas Al Amin*, Andi Sunyoto, Hanif Al Fatta
Magister of Informatics Engineering
Universitas Amikom Yogyakarta
Sleman, Indonesia
*annas.am@students.amikom.ac.id

***Abstract*—Recommendation systems are widely used on website platforms such as e-commerce, marketplaces, streaming movies to produce appropriate item recommendations for each user. The traditional memory-based collaborative filtering approach is currently used in recommender systems. This approach relies on users' item rating as a basic approach for calculates the similarity of users' responses about products to predict item recommendations, but the weakness is high prediction errors. This study aimed to reduce prediction errors from a memory-based collaborative filtering approach using hybrid filtering, so the recommendation system's quality can be improved. The hybrid filtering approach combined a collaborative filtering approach based on a matrix factorization model and content-based filtering, which can reduce prediction errors to produce accurate item recommendations. The proposed method has been evaluated ten times using the root mean squared error to measure the prediction error. As a result, the hybrid filtering approach produced the smallest prediction error of 0.68, while memory-based collaborative filtering was 2.98. Based on the results, the hybrid filtering approach is better than memory-based collaborative filtering.**

Keywords—recommender system, hybrid filtering, matrix factorization, collaborative filtering, content based filtering

I. INTRODUCTION

The recommendation system is one of the support systems used on a website-based platform such as e-commerce, marketplace, streaming movies to provide item recommendations to users. Currently, the approach that widely used in recommendation systems is memory-based Collaborative Filtering (CF) which is the traditional model for recommendation systems. The memory-based CF approach is one of the common approaches that is often used to recommend an item to a user using the rating value given by the user. The main idea is to predict which items a user might like based on the preferences of other users in giving an item rating value [1]. This approach is also one of the most successful and is often implemented for recommendation systems on a variety of platforms due to the resulting model of providing personalized item recommendations to users [2-4]. There is another approach that is also often used in recommendation systems, namely Content-Based Filtering (CBF) which utilizes the similarity of the content of an item such as categories, descriptions, or other attributes as a basis for providing item recommendations to users. This approach can provide item recommendations to users without using the information on the rating value of the item.

There are weaknesses in the memory-based CF and CBF approaches, the first weakness is in the traditional CF approach, if there are items that do not have a rating, then these items cannot be recommended by the system (Cold Start Problem) [5], the second weakness, this approach ignores content information of the items that can actually be used as a basis for item recommendations, and the third weakness, the memory-based CF approach requires high computation because every time there is new rating data, the model must be retrained, so it produces high prediction errors. In the CBF approach, the first weakness is that the item recommendation results are not personalized, meaning that the item recommendations to users are all the same, and the second weakness is that the system cannot provide recommendations for items that have different content [6]. The purpose of this study is to reduce the prediction error of the traditional memory-based CF approach from previous studies using a combination of a new CF approach based on matrix factorization and CBF models into an approach called Hybrid Filtering (HF) to improve the quality of the recommendation system. The HF approach has several advantages. First, the system can still provide item recommendations to new users who have not given an item rating at all [7], second. The model can provide personalized item recommendations so each user can get various item recommendations [8], third, the HF approach can reduce the computational load because the training process is only done once, so it can reduce the prediction error to be smaller than the memory-based CF approach. The model generated from this HF approach can recommend items based on content similarity because of the capabilities of the CBF approach. After that, the results of item recommendations based on similar content will be processed again using the CF approach based on the matrix factorization model to produce personalized item recommendations. Thus, the HF approach can produce personalized new item

recommendations based on the similarity of item content, and the model can reduce prediction errors smaller than the traditional memory-based CF approach, so the HF approach can improve the quality of the recommendation system.

II. Related Literature

Many other researchers have tried using new techniques and approaches to reduce prediction errors in recommendation systems in order to improve the quality of the recommendation system.

Starting from the first researcher [5] using a Hybrid approach with a combination of Knowledge-Based Recommender and CF approaches, researchers not only use item rating data but also utilize interaction data from user profiles on items as one of the inputs for the recommendation system. The prediction error results from this hybrid model are measured using Root Mean Square Deviation (RMSE) and produce a prediction error of 0.347. On the other hand, there are researchers [8] who take advantage of implicit feedback from users in the form of similar ratings between users, and positive user behaviour obtained from the conversion of a given rating. If a user gives a rating of not less than 3, then the behaviour is considered positive, but if it is less than 3, then the behaviour is harmful. In this way, the researcher can solve the data sparsity problem so that it can reduce the prediction error of the recommendation system. The third researcher [9] also proposed a new method called K-RecSys to reduce prediction errors in the traditional CF approach used for fashion retail e-commerce. This method uses click data on items that users have seen when visiting an online store, and item sales data obtained from offline stores to provide item recommendations based on the same category and different categories. The results of the method proposed by the researcher were measured by dividing the two groups of item recommendations, the Control Group from the old recommendation system, and the Experimental Group from the new recommendation system proposed by the researcher. After testing, the results of the recommendations from the Experimental Group get the most interactions compared to the Control Group. Thus the model proposed by the researcher has a better performance. Researchers [10] also overcome the problem of data sparsity to reduce prediction errors by proposing a Fusion Matrix Factorization model to measure the multi-factor similarity between users. Researchers use the extreme behaviour similarity measure and linear similarity algorithm, after obtaining a multi-factor similarity or neighbour matrix, then combined with the original matrix called the Fusing process to obtain a prediction matrix based on the representation of latent factors from users and items. The results of the model from this researcher get the smallest RMSE and MAE values compared to the traditional CF approach.

The use of memory-based CF can produce high prediction errors, but the researcher [11] can overcome this by using clustering techniques in a hybrid approach. Researchers use the K-Means clustering algorithm to classify items based on similarity in item ratings, then use the Weighted Average of Deviation algorithm to predict the rating in the cluster. This clustering technique can reduce the prediction error of memory-based CF, but it takes much time for the computation process. Reducing the value of prediction errors and overcoming the sparsity of data in the recommendation system can be overcome by combining the Matrix Factorization and Deep Learning approaches as was done by researchers [12] to improve the CF approach. The researcher uses the NMF algorithm to reduce the error rate, then measures the impact of the feature representation calculated with the Quadratic Polynomial Regression formula to obtain latent features by increasing Item-average Clustering. The classification results of the CF algorithm based on in-depth learning analysis using the Non-Linearity calculation formula of the user-item produces a recommendation system with a hybrid approach. The results of the new model proposed by this researcher can reduce prediction errors in MSE estimation and overcome data sparsity problems and can improve the quality of the recommendation system.

From the research described above, most researchers use an approach that only utilizes the item rating data in the model. The difference in this study is that the data used is not only item rating but also item content information to produce personalized recommendations for items that have similar content, and can reduce prediction errors from the traditional memory-based CF approach to improve the quality of the recommendation system using the HF approach.

III. Proposed Methods

A. Dataset

Dataset taken from MovieLens [13], MovieLens is a recommendation system developed by GroupLens. This dataset is open source and has been widely used for research related to recommendation systems. This dataset contains a collection of movie rating data from various genres with a range of movie rating 1 to 5. Details of the dataset can be seen in table 1.

TABLE I. Dataset Detail

Attribute Name	Number
Metadata (Genres, Cast, Keywords, Director)	3706
Movie Title	3706
UserID	6040
Rating Movie	1.000.209

B. Data and Text Pre-Processing

The UserID attribute is filtered to find users who have given ratings ranging from 150 to 200. After being filtered, 518 users are obtained with a total rating of 89290. This filtering process is carried out so that the number of rating data for each user is balanced. Then the metadata attribute which contains text data, if in the attribute there are words that are not informative such as conjunctions, pronouns, for example "and", "the", and "him/her", the word will be deleted so the metadata

attribute can represent the informative content. In the testing and evaluation stage of the model later, the attribute rating of the movie with the number 89290 ratings will be divided into 80,361 training data and 8929 testing data to be evaluated ten times using RMSE, then the results will be compared with the memory-based CF approach in previous studies.

C. Hybrid Filtering

This Hybrid Filtering is a combination of CBF and CF approaches based on the Matrix Factorization (MF) model. First, the CBF approach is used to produce item recommendations based on similar content. Then the rating items will be used as input for the MF model-based CF approach to producing a recommendation system hybrid model that can reduce prediction errors and produce personalized item recommendations. Details of the flow of the HF approach implementation can be seen in figure 1.

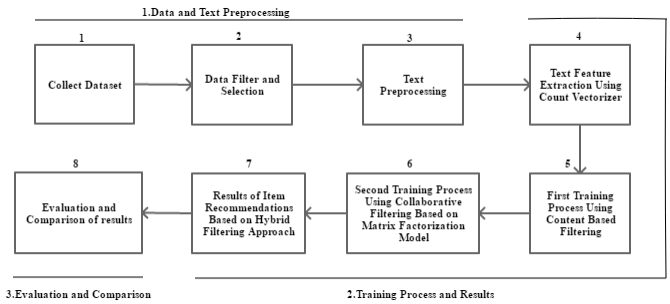

Fig. 1. Implementation of hybrid filtering approach.

1) Content based filtering: Basically, this approach is used to generate item recommendations based on similar content such as descriptions, titles, names of items that have been seen by the user. This approach uses Natural Language Processing (NLP) for computing text data. In the metadata attribute containing the text documents, it will be converted into numeric using one of the bag of words models, namely Count Vectorizer [14]. After the data is encoded, then perform calculations with the cosine similarity and cosine distance formulas to find the similarity in the content of items that have been seen by the user. The similarity of items is obtained from the calculation of the closest value distance to items that have been seen by the user. The cosine similarity formula can be seen in equation 1 and the cosine distance in equation 2.

$$similarity = cos(\theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

Variable A_i and B_i are components of vectors A and B that represent similarities between items.

$$distance = Dc(A,B) = 1 - Sc(A, B) \quad (2)$$

Variable Dc(A, B) is the similarity item based on the distance between values vector items A dan B, then Sc(A, B) is the result of calculating similarity value vector items A dan B.

2) Collaborative filtering based on matrix factorization model: This approach is used to produce personalized item recommendations by predicting the rating of the items based on the results of the first training process using the CBF approach. After that, the second training process is carried out to predict the item rating using one of the Matrix Factorization methods, namely Singular Value Decomposition (SVD). The result of the second training process from the CF approach based on the Matrix Factorization model is called the Hybrid Filtering approach. This item rating prediction process is to produce personalized item recommendations based on the order of the highest to lowest rating prediction values. The result of this hybrid approach is personalized item recommendations based on the similarity of item content. The SVD formula can be seen in equation 3.

$$R = M(\sum)U^T \quad (3)$$

Variable R is a matrix that represents user rating for each item. Variable M is the eigenvector of the matrix RR^T, variable U is eigenvector of the matrix R^TR, and ($\sum$) is the association of eigenvalues into a diagonal matrix.

D. Root Mean Squared Error (RMSE)

After the item rating prediction stage is carried out, the final step is to evaluate the performance of the HF approach by estimating the prediction error using the RMSE formula, then comparing the results with the memory-based CF approach from previous studies using the same dataset. This technique is applied by adding the actual rating value minus the predicted rating value and then squaring it, then dividing it by the amount of data, then rooting it. RMSE is a way to evaluate the model by measuring the level of prediction error generated by the model [8]. The RMSE formula can be seen in equation 4.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(A_i - P_i)^2}{n}} \quad (4)$$

Variable A_i is the actual item rating value, variable P_i is the value of item rating prediction, and variable n is the actual amount of rating data.

IV. Results and Evaluation

At this stage, to prove that the HF approach can improve the quality of the recommendation system and can reduce the prediction error to be smaller than the memory-based CF approach from previous studies, an evaluation process will be carried out to compare the performance of the HF approach model with memory-based CF based on the results of the RMSE value. If the prediction error is getting smaller, the accuracy of the model is getting better, but if the prediction error is getting bigger, the accuracy of the model is getting worse. The evaluation process will be carried out ten times, using 80361 training data and 8929 testing data. Each evaluation will calculate the RMSE value and compare the difference of prediction errors from the two approaches. Then

the final stage will average the RMSE value from 10 evaluations to decide the best approach model based on the smallest RMSE value. The results of the evaluation and comparison of the prediction errors of the two approaches can be seen in figure 2.

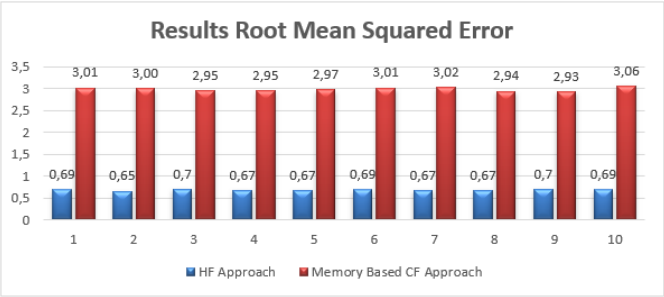

Fig. 2. Evaluation and comparison results.

Based on the results of the evaluation and comparison of the prediction errors of the two approaches in Figure 2, it can be seen that after ten evaluations, the HF approach can produce the smallest prediction error in the second iteration with an RMSE value of 0.65. In contrast, the memory-based CF approach produces the highest RMSE value in the tenth iteration of 3.06.

V. Conclusion

Based on the results of the evaluation and comparison of prediction errors that have been carried out in this study, it can be concluded that the HF approach can reduce prediction errors smaller than the traditional memory-based CF approach, from the results of 10 evaluations that have been carried out, the HF approach can produce the smallest average prediction error of 0.68. In contrast, the memory-based CF approach produces the largest average prediction error of 2.98. Based on the evaluation results, the Hybrid Filtering approach is better than the memory-based collaborative filtering from previous studies. So, our proposed Hybrid Filtering approach could improve the quality of the recommendation system.

References

[1] N. Nassar, A. Jafar, and Y. Rahhal, “A novel deep multi-criteria collaborative filtering model for recommendation system,” Knowledge-Based Syst., vol. 187, p. 104811, 2020.

[2] S. Jiang, S. C. Fang, Q. An, and J. E. Lavery, “A sub-one quasi-norm-based similarity measure for collaborative filtering in recommender systems,” Inf. Sci. (Ny)., vol. 487, pp. 142–155, 2019.

[3] Y. Lv, Y. Zheng, F. Wei, C. Wang, and C. Wang, “AICF: Attention-based item collaborative filtering,” Adv. Eng. Informatics, vol. 44, no. February, p. 101090, 2020.

[4] K. Li, X. Zhou, F. Lin, W. Zeng, B. Wang, and G. Alterovitz, “Sparse online collaborative filtering with dynamic regularization,” Inf. Sci. (Ny)., vol. 505, pp. 535–548, 2019.

[5] S. Rahmawati, D. Nurjanah, and R. Rismala, “Analisis dan Implementasi pendekatan Hybrid untuk Sistem Rekomendasi Pekerjaan dengan Metode Knowledge Based dan Collaborative Filtering,” Indones. J. Comput., vol. 3, no. 2, p. 11, 2018.

[6] A. S. Tewari, “Generating Items Recommendations by Fusing Content and User-Item based Collaborative Filtering,” Procedia Comput. Sci., vol. 167, no. 2019, pp. 1934–1940, 2020.

[7] S. Natarajan, S. Vairavasundaram, S. Natarajan, and A. H. Gandomi, “Resolving data sparsity and cold start problem in collaborative filtering recommender system using Linked Open Data,” Expert Syst. Appl., vol. 149, p. 113248, 2020.

[8] Y. Hu, F. Xiong, D. Lu, X. Wang, X. Xiong, and H. Chen, “Movie collaborative filtering with multiplex implicit feedbacks,” Neurocomputing, vol. 398, pp. 485–494, 2020.

[9] H. Hwangbo, Y. S. Kim, and K. J. Cha, “Recommendation system development for fashion retail e-commerce,” Electron. Commer. Res. Appl., vol. 28, pp. 94–101, 2018.

[10] C. Feng, J. Liang, P. Song, and Z. Wang, “A fusion collaborative filtering method for sparse data in recommender systems,” Inf. Sci. (Ny)., vol. 521, pp. 365–379, 2020.

[11] G. Geetha, M. Safa, C. Fancy, and D. Saranya, “A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System,” J. Phys. Conf. Ser., vol. 1000, no. 1, 2018.

[12] Hanafi, N. Suryana, and A. S. B. H. Basari, “Hybridization approach to eliminate sparse data based on nonnegative matrix factorization & deep learning,” J. Theor. Appl. Inf. Technol., vol. 96, no. 14, pp. 4502–4512, 2018.

[13] F. M. Harper and J. A. Konstan, “The movielens datasets: History and context,” ACM Trans. Interact. Intell. Syst., vol. 5, no. 4, pp. 1–19, 2015.

[14] R. F. Amanullah, E. Utami, and A. Sunyoto, “Citation detection on scientific journal using support vector machine,” 2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019, pp. 549–553, 2019.