Recommendation System Using User-Based Collaborative Filtering and Spectral Clustering

Malim Muhammad
{malim.muhammad@gmail.com}

Department of Mathematics Education, Faculty of Teacher Training and Education, Universitas Muhammadiyah Purwokerto

Abstract. Recommendation systems are tools that can solve this problem by classifying people using User-Based Collaborative Filtering and Spectral Clustering approaches, resulting in more accurate recommendations. Preprocessing the data is the first step in the recommendation process, after which the data is grouped using the Spectral Clustering method. In the process of creating rating predictions and film recommendations based on similarity derived using the Pearson Correlation and Cosine Similarity algorithms, the clustering results are used to determine which users will be neighbors. Based on system experiments that have been conducted using variations in the number of clusters 3, 5, and 7, variations in the number of neighbors 1, 2, 3, 5, and 10, and comparing the results of MAE calculations of rating prediction results using a combination of spectral clustering methods with Pearson correlation, spectral Clustering with cosine similarity, with Pearson correlation and with cosine similarity, get results where the combination of methods with cosine similarity using 3 clusters and two neighbors becomes the method that has the best accuracy in making movie recommendations, namely with an MAE value of 0.3114. This is because the combination of methods has the smallest MAE calculation value. In other words, it has a minimal recommendation error rate. Meanwhile, the recommendation system with spectral Clustering only gets an MAE value of 0.3611, namely by a combination of spectral clustering methods with cosine similarity using 5 clusters and two neighbors. The result of the accuracy of the combination of spectral clustering method with Pearson correlation gets the lowest MAE value of 0.4109 with an average MAE value of 1,159, the combination of spectral clustering method with cosine similarity gets the lowest MAE value of 0.3611 with an average MAE value of 1,190, the combination of a method with Pearson correlation gets the lowest MAE value of 0.4711 with an average MAE value of 0.911 and the combination of methods with cosine similarity get the lowest MAE value of 0.3114 with an average MAE value of 0.807.

Keywords: Recommended System, User-Based Collaborative Filtering, Spectral Clustering, Pearson Correlation, Cosine Similarity

1 Introduction

A recommendation system is a system used to predict items to be communicated to the user based on the user's relationship to other items or to other users. Recommendations relate to various decision-making processes, such as items to buy, music to be listened to, or what information to read [1]. Searches for items that will be recommended to the user are done by looking at the similarity, either the similarity of an item with other items based on content or

ISTED 2021, July 17-18, Purwokerto, Indonesia

DOI 10.4108/eai.17-7-2021.2312409

similarity of taste between one user and another based on the rating given to the item. As the times progress, much research has been done on recommendation systems to find new approaches to addressing more and more complex problems. Recommendation system approaches commonly used in recommendation systems are content-based filtering and collaborative filtering approaches.

Research conducted by Hadi et al. (2020) on film recommendation systems using User-Based Collaborative Filtering and K-Modes methods for user grouping [2]. Calculation of accuracy in this study using the Mean Reciprocal Rank (MMR) method and get a result of 0.17 with a data train and test ratio of 80%; 20% and 0.15 with a data train and test ratio of 60%; 40%. The accuracy of this study is quite low because, based on the MRR rule, the recommendation is said to be less precise if the value is close to 0 with a limit of 0.5 and is said to be appropriate if the value is above 0.5 and close to 1.

Research conducted by Ahuja et al. (2019) about the film recommendation system using K-Means Clustering and K-Nearest Neighbor [3]. In this study, the system accuracy calculation using Root Mean Squared Error (RMSE) and obtained the result that the number of clusters used affects the results of the RMSE calculation, where the smaller the number of clusters used, the smaller the RMSE value produced, with the smallest RMSE value produced is 1.08 using 2 clusters.

Research conducted by Halim et al. (2017) on Film Recommendation System using K-Means Bisecting and Collaborative Filtering [4]. This study resulted in a Mean Absolute Error (MAE) value which is a combination of Bisecting K-Means and User-Based Collaborative Filtering of 1.63, lower than the MAE value, which is a combination of K-Means Noisy and Item-Based Collaborative Filtering. In addition to the recommended method, the distribution of rating values in the dataset also greatly affects the MAE value, where if the distribution of rating values is uneven, it will result in a higher error value in the recommendation system.

Research conducted by Yusuf et al. (2012) on the development of student value predictor software using spectral Clustering and bagging linear regression methods [5]. The study used spectral Clustering and K-means clustering algorithms to group data as comparisons and regressions with Bootstrap Aggregating Linear Regression using student value data. The calculation of the accuracy of predictions in this study is calculated by the RMSE method. The results of this study showed that software developed with a spectral clustering algorithm that supports bootstrap aggregating linear regression algorithm proved capable of predicting student grades with RMSE error values of about 0.05-0.08 compared to using K-Means Clustering, which obtained RMSE error results of about 0.1.

Based on the description above, researchers created a movie recommendation system using User-Based Collaborative Filtering and using the Spectral Clustering method in user grouping. This is so that the system can recommend movies based on users in the same group so that the results of the recommendations given become better. The accuracy of the recommendation system made will be calculated by the Mean Absolute Error (MAE) method.

Collaborative filtering (CF) is an approach to the recommendation system. This approach provides recommendations on an item by looking for similarities between users to the item. The main idea in this approach is to find out information about past user behavior and opinions of a group of users who are then used to predict which items will be liked or attractive to the user [6]. The term user in Collaborative Filtering refers to the person who gives an assessment of the items in the system, who will later receive recommendations from the system.

Ratings can be collected by explicit, implicit, or by both. An explicit rating is when a user is directly asked to provide an assessment of an item. The implicit rating means the system automatically gains user preferences passively by looking at user behavior. The assessment given is only based on user behavior; for example, when a member in a library decides to borrow a book item, then the member is considered interested or likes the item, and vice versa is considered disinterested or disliked if he does not borrow a book. In this way, the user profile is formed without involving the additional role of the user. The disadvantage of this way is, of course, that the alleged assessment given may be inappropriate [6].

User-Based Collaborative Filtering is a method of recommendation system that provides item recommendations based on similarities between users with each other. Recommended items are items favored by other users who bear similarities to the main user [2]. To find an item that is favored by one user, it must look for other users who have similar tastes or favorites. Here is an illustration of the User-Based Collaborative Filtering method in providing recommendations in the following figure 1.

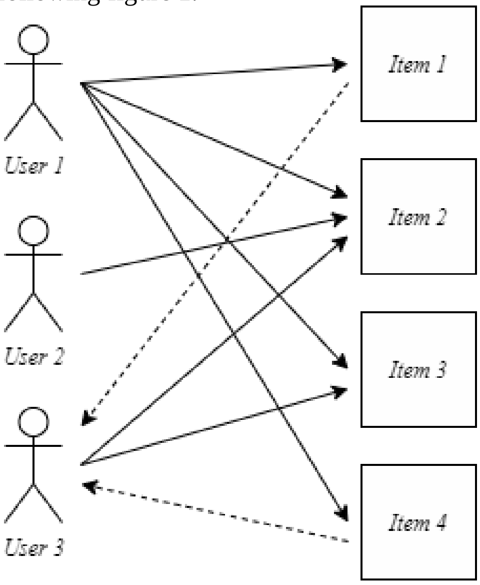


Figure 1. Illustration of User-Based Collaboration Filtering

The most commonly used method is the nearest neighbor method. This method is based on items that have been selected by nearby neighbors or other users that bear similarities to the main user, thus producing predictions of items that are likely to be selected by the main user in the future [7].

The similarity is a method in machine learning that serves to calculate similarities between 2 or more data. Similarity calculations use algorithmic methods such as Pearson correlation, cosine similarity, and many more. Pearson Correlation is one of the algorithms commonly used in calculating the similarity between users and other users. Correlation is a measurement technique that determines the closeness of relationships between two sets of different numbers. Correlation calculations have a condition where the set of numbers calculated must have a fixed order and pair with each other between the two sets. The results of measurement can be either positive relationships or negative relationships. Positive relationships indicate that both sets have a tendency to increase or increase equal values. In contrast, the negative relationship shows that both sets have a tendency to decrease or decrease in equal value [8]. Here is the Pearson Correlation equation for calculating the similarity between users:

$$PC(u,v)=\frac{\sum_{i=1}^{n}(r_{ui}-\overline{r}_{u})(r_{vi}-\overline{r}_{v})}{\sqrt{\sum_{i=1}^{n}(r_{ui}-\overline{r}_{u})^{2}(r_{vi}-\overline{r}_{v})^{2}}} \qquad 1)$$

Where:

- $PC(u,v)$ is the similarity value between user u and user v
- r_{ui} and r_{vi} is the user rating ui and vi to item-i
- $\overline{r}_{u}$ and $\overline{r}_{v}$ is the average rating u and v to item-i
- n is the number of items

Cosine similarity is a method used to calculate the similarity of each user with each other. Different rating scales between different users will result in different similarity values [8]. Here is the Cosine Similarity equation for calculating similarity between users:

$$Sim(u_i,u_k)=\frac{r_i.r_k}{|r_i||r_k|}=\frac{\sum_{j=1}^{m}r_{ij}r_{ik}}{\sqrt{\sum_{j=1}^{m}r_{ij}^{2}\sum_{j=1}^{m}r_{kj}^{2}}} \qquad 2)$$

Where:

- $Sim(u_i,u_k)$ is the similarity value between the ith user and the kth user
- u_i and u_k is the ith user and the kth user
- r_i and r_k is the ith user rating and the kth user
- m is the number of items

The calculation of rating predictions on the recommendation system is used to find predictions of rating values given by users against specific items. This predictive calculation is implemented as the final step of the Collaborative Filtering approach in providing recommendations. After the similarity value between users or items has been obtained, the next step is to determine the number of neighbors to determine the predicted value of the rating. One method of calculating rating prediction is the weighted average method [7]. Here is the weighted average formula for calculating predictions:

$$P_{(a,i)}=\overline{r}_{a}+\frac{\sum_{u=1}^{n}(r_{u,i}-\overline{r}_{u})\times sim(a,u)}{\sum_{i=1}^{n}sim(a,u)} \qquad 3)$$

Where:

- $P_{(a,i)}$ is the user's rating prediction of item-i
- n is the number of neighbors.

- $\overline{r_a}$ is the average user rating.
- $r_{u,i}$ is the user rating u to item-i
- $\overline{r_u}$ is the average user rating u
- $sim(a,u)$ is the similarity value between user a and user u

Mean Absolute Error is one method for calculating the accuracy rate of system recommendations based on the magnitude of errors from the results of the system's rating predictions against the actual rating that the user gives to an item [1]. MAE evaluation uses a simple calculation technique, which is to calculate the difference of all items that have been rated by the user and have a rating prediction value. The difference will be absolutely (become a positive value) then averaged. From the results of MAE calculations, it is clear how far the difference in the value of the rating prediction is given by the system with the actual rating value. The greater the value produced by MAE, it can be interpreted that the value of the rating prediction by the system is increasingly inaccurate; conversely, if the resulting MAE value is close to 0, then the prediction by the system is more accurate [8]. Here is the formula for calculating MAE.

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \quad 4)$$

Where:

- MAE is the magnitude of the error of the prediction result
- p_i is a prediction rating.
- q_i is the real rating.
- N is many original ratings and predictions.

Clustering is one of the methods of data exploration used in finding patterns in a dataset. In general, the pattern can be seen from the similarity of properties, characteristics, or characteristics of the records in the dataset [5]. The clustering process will group data items into a small number of groups in such a way that each group has an essential similarity that will later facilitate the search for data based on existing similarities [10].

Spectral Clustering is one of the clustering methods that grouping based on the similarity of each data. These similarities are seen from the relationship between data with each other. Spectral Clustering has formed a graph of existing data. Where the vertex of the graph is every record on the data and edge is the relationship between data which is usually in the form of distance from two related records [5]. The steps in doing Spectral Clustering are as follows [11]:

1) Forming a similarity matrix (W)

The similarity matrix is formed based on the relationships between data. If there is a relationship, then there is a value between data with each other, while if there is no relationship will be worth 0. The diagonal value in the similarity matrix will be worth 0 because there is no relation to the data itself. Calculation of similarity values is calculated by exponential distance equation using the formula:

$$W_{ij} = \exp\left(\frac{-\left\| S_i - S_j \right\|^2}{2\sigma^2} \right) \quad 5)$$

Where:

- W is a similarity value.
- i and j is a data number.
- S is data
- σ sigma value as a scale parameter to control similarities

2) Forming a diagonal matrix (D)

The diagonal matrix contains the number of edges connected to each diagonal data. The formula for calculating diagonal values is:

$$D_{ij} = \sum_{j=1}^{n} W_{ij} \quad 6)$$

Where:

- D is a diagonal value.
- n is the number of data
- i is the data line number
- j is the data column number
- W is a similarity value.

3) Forming a Laplacian matrix (L)

The Laplacian matrix is formed using the result of a degree matrix (D) minus the similarity matrix (W). For certain datasets, the Laplacian matrix can also be calculated using the following normalization formula:

$$L_{sym} = D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \quad 7)$$

Where:

- L is the Laplacian value
- D is a diagonal matrix.
- W It's a similarity value.

4) Calculates the value of k eigenvectors of the Laplacian matrix (L), where k is the parameter of the number of clusters. The k-Eigen matrix is the first k eigenvectors of the Laplacian matrix.
5) Normalize data with a k-Eigen matrix so that a new matrix will be formed that represents each value of the normalization result.
6) The results of the normalization data are then clustered with K-Means Clustering. The ith data will be entered in a cluster if and only if the i-year normalization data enters the same cluster.

2 Research Methods

In this study, the item that became the object of the recommendation was a film from a data set of MovieLens.org that had been passed by the user. MovieLens.org is an open data set for development and research in the field of recommendation systems managed and run by GroupLens, a research laboratory at the University of Minnesota (https://movielens.org/). The data source used in this study is in the form of *ratings,* and *movies* data from *movielens.org*. The data used consists of 1,048,575 *rating* data, 7120 *user* data, and 14,026 *movie* data. The data used has a CSV format *(comma separated values)* and is contained in 2 *files,* namely *"ratings.csv"* and *"movies.csv."* For data, *"ratings.csv"* contains *userId, movieId, rating,* and *timestamp*. As for the data, *"movies.csv"* contains *movieId, title,* and *genres.*

System design is created with the aim of determining the system workflow to be created and minimize the occurrence of errors in the flow of program processes. The design of the system in this study can be seen in figure 2.

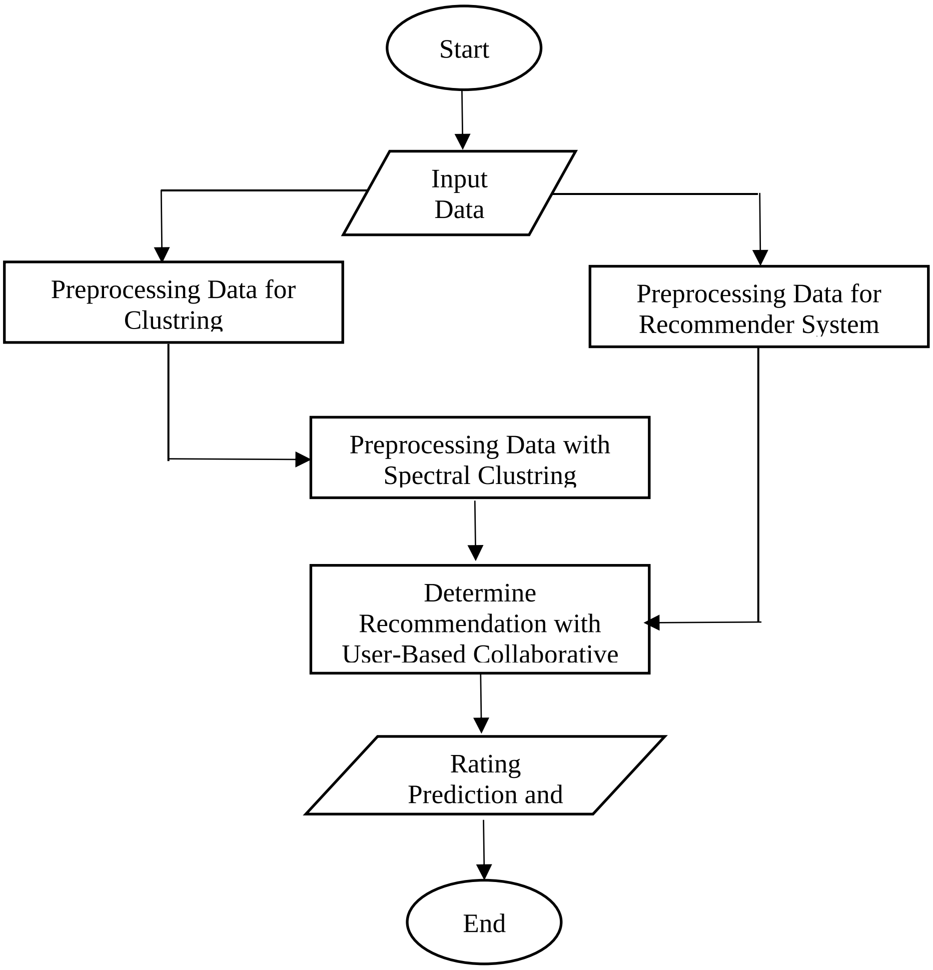

Start
Input Data
Preprocessing Data for Clustring
Preprocessing Data for Recommender System
Preprocessing Data with Spectral Clustring
Determine Recommendation with User-Based Collaborative
Rating Prediction and
End

Figure 2. Design of *User-Based Collaborative Filtering* System with *Spectral Clustering*

In the design of this system, data input comes from "ratings.csv" and "movies.csv" files. After the data is entered, then it will be continued with preprocessing data divided into two, namely preprocessing data for Clustering and preprocessing data for recommendation systems. The next stage is to cluster data with the Spectral Clustering method, where the results of the clustering process will be used to determine movie recommendations with User-Based Collaborative Filtering which will later produce a prediction of the user's rating of a particular film and the accuracy of the film recommendation accuracy will be calculated using the Mean Absolute Error (MAE) method.

3 Results and Discussion

The preprocessing stage is done to prepare the data so that it is ready for use. The preprocessing stage is divided into 2, namely data preprocessing for the clustering process and data preprocessing for the recommendation system process. The initial stage in both preprocessing processes is to form a pivot table that contains the user rating value for each existing film so that a pivot table measuring 7120 rows × 14026 columns will be formed. This aims to find out which movies have been and have not been given a rating by each user.

This data preprocessing process is done using a previously formed pivot table. Then the NaN value on the pivot table is changed to 0. In the new pivot table, the process of standardizing data using the StandardScaler() function, which then results in the standardization of data in normalization using the normalize() function, and finally, the normalized data is reduced in dimension using the Principal Component Analysis (PCA) method by using the PCA function with the number of components formed is 2, resulting in an output of data dimensions of 7120 × 2. Preprocessing data is used in the clustering process using the spectral clustering method.

The process of data clustering is done using preprocessing data. The clustering process is carried out using two methods, namely spectral Clustering with variations in the number of clusters, namely 3, 5, and 7. The results of this clustering method will be compared to the results in calculating user rating predictions against certain films and in making movie recommendations. Here is the code in the clustering process using spectral Clustering. The clustering process is carried out using the python library: cluster class, the Spectral Clustering function for clustering data with the Spectral Clustering method. In the method used, the parameters of the cluster are changed according to the number of clusters that have been determined, namely 3, 5, and 7.

In the clustering process using the spectral clustering method, the similarity matrix is calculated using the Nearest Neighbors method specified in affinity parameters so that the Spectral Clustering function used has directly formed a similarity matrix, diagonal matrix, Laplacian matrix, eigenvalue and eigenvector, and data clustering. Here are the results of data visualization and the division of the number of members for each cluster:

Table1. Data visualization and division of the number of cluster members (k=3)

Data Visualization
Spectral Clustering (k=3)

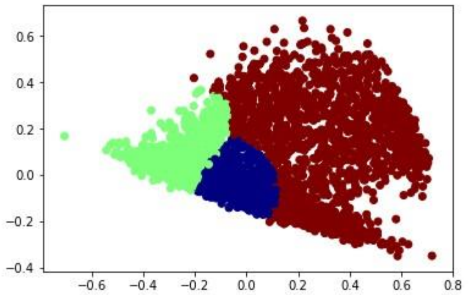
			
Data Sharing			
Spectral Clustering			
Cluster	0	1	2
Sum	2852	1993	2275

Table2. Data visualization and division of the number of cluster members (k=5)

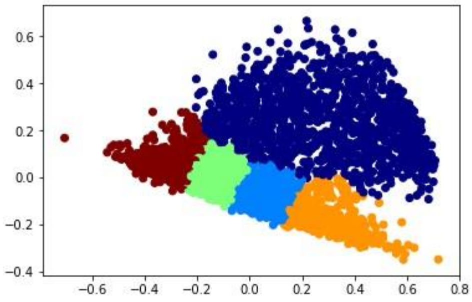
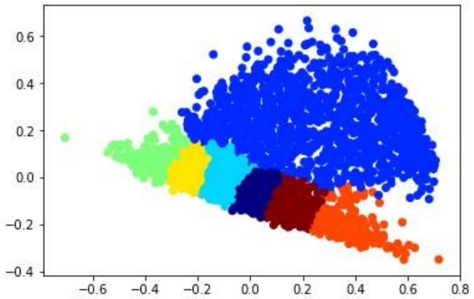
Data Visualization					
Spectral Clustering (k=5)					
					
Data Sharing					
Spectral Clustering					
Cluster	0	1	2	3	4
Sum	1200	1665	2015	937	1303

Table 3. Data visualization and division of the number of *cluster* members (k= 7)

Data Visualization
Spectral Clustering (k=7)

							
Data Sharing							
Spectral Clustering							
Cluster	0	1	2	3	4	5	6
Sum	1180	1254	1425	591	1256	482	932

Based on system experiments that have been conducted using variations in the number of clusters 3, 5, and 7, variations in the number of neighbors 1, 2, 3, 5, and 10, and comparing the results of MAE calculations of rating prediction results using a combination of spectral clustering methods with Pearson correlation, spectral Clustering with cosine similarity, with Pearson correlation and with cosine similarity, get results where the combination of methods with cosine similarity using 3 clusters and two neighbors becomes the method that has the best accuracy in making film recommendations, namely with MAE value 0.3114. This is because the combination of methods has the smallest MAE calculation value. In other words, it has a minimal recommendation error rate. Meanwhile, the recommendation system with spectral Clustering only gets an MAE value of 0.3611, namely by a combination of spectral clustering methods with cosine similarity using 5 clusters and two neighbors.

The result of accuracy combination spectral clustering method with Pearson correlation gets the lowest MAE value of 0.4109 with an average MAE value of 1.159, The combination of spectral clustering method with cosine similarity gets the lowest MAE value of 0.3611 with an average MAE value of 1,190, the combination of the method with Pearson correlation gets the lowest MAE value of 0.4711 with an average MAE value of 0.911 and the combination of methods with cosine similarity gets the lowest MAE value of 0.311 with an average MAE value of 0.807.

In terms of making movie recommendations, based on the system testing that has been done, it can be seen that each combination of methods used will produce different film recommendation results. In addition, the difference in the number of clusters used also affects the results of the resulting film recommendations. This can occur due to differences and member changes in each cluster generated by the spectral clustering method, thus changing the list of users used in performing rating predictions.

4 Conclusion

Based on research that has been done, it is research on comparing user-based CF methods and CF item-based methods in providing better recommendations on data sets in the form of movie objects. The user-based CF method and the item-based CF method can be used to predict a user's rating of a movie. User-based CF and item-based CF methods have the potential to be applied to a website film recommendation system. Based on system experiments that have been conducted using variations in the number of clusters 3, 5, and 7, variations in the number of neighbors 1, 2, 3, 5, and 10, and comparing the results of MAE calculations of rating prediction results using a combination of spectral clustering methods with Pearson correlation, spectral Clustering with cosine similarity, with Pearson correlation and with cosine similarity, get results where the combination of methods with cosine similarity using 3 clusters and two neighbors becomes the method that has the best accuracy in making film recommendations, namely with a value of MAE 0.3114. This is because the combination of methods has the smallest MAE calculation value. In other words, it has a minimal recommendation error rate. Meanwhile, the recommendation system with spectral Clustering only gets an MAE value of 0.3611, namely by a combination of spectral clustering methods with cosine similarity using 5 clusters and two neighbors. The result of the accuracy of the combination of spectral clustering method with Pearson correlation gets the lowest MAE value of 0.4109 with an average MAE value of 1,159, the combination of spectral clustering method with cosine similarity gets the lowest MAE value of 0.3611 with an average MAE value of 1,190, the combination of the method with Pearson correlation gets the lowest MAE value of 0.4711 with an average MAE value of 0.911 and the combination of methods with cosine similarity gets the lowest MAE value of 0.3114 with an average MAE value of 0.807.

Acknowledgments. The authors would like to express their gratitude to Universitas Muhammadiyah Purwokerto for allowing us to assist and support in completing this research.

References

[1] Ricci, F., Rokach, L., Shapira, B., Kantor, P. B., & Ricci, F. (2011). Recommender Systems Handbook. In Recommender Systems Handbook. https://doi.org/10.1007/978-0-387-85820-3

[2] Hadi, I., Santoso, L. W., Tjondrowiguno, A. N., & Siwalankerto, J. (2020). The Film Recommendation System uses User-based Collaborative Filtering and K-modes Clustering. Infra Journal.

[3] Ahuja, R., Solanki, A., & Nayyar, A. (2019). Movie recommender system using k- means clustering and k-nearest neighbor. Proceedings of the 9th International Conference On Cloud Computing, Data Science and Engineering, Confluence 2019, 263–268. https://doi.org/10.1109/CONFLUENCE.2019.8776969

[4] Halim, A., Gohzali, H., Panjaitan, D.M., & Maulana, I. (2017). The Film Recommendation System uses K-Means Noise and Collaborative Filtering. Citisee, 1(3), 37–41.

[5] Yusuf, A., Ginardi, H., & Arieshanti, I. (2012). Student Value Predictor Software Development Using Spectral Clustering methods and Linear Regression Bagging. Its Technical Journal, 1(2), A246–A250. http://ejurnal.its.ac.id/index.php/teknik/article/view/645

[6] Dzumiroh, L., & Saptono, R. (2016). Application of Collaborative Filtering Method Using Implicit Rating on Film Selection Recommendation System in VCD Rental. Journal of Technology &Information ITSmart, 1(2), 54. https://doi.org/10.20961/its.v1i2.590

[7] Babu, M. S. P., & Kumar, B. R. S. (2011). An Implementation of the User-based Collaborative Filtering Algorithm. (IJCSIT) International Journal of Computer Science and Information Technologies, 2(3), 1283–1286. https://doi.org/10.1016/S0013-4686(01)00598-9

[8] Theodorus, A., & Budiyanto Setyohadi, D. (2016). User-Based Collaborative Filtering By Utilizing Pearson-Correlation To Find Nearby Neighbors In The Recommendation System. Master's Thesis in Information Technology, Atma Jaya University Yogyakarta, 1–6. http://e- journal.uajy.ac.id/8924/

[9] Pamuji, A. (2017). Public Housing Credit Recommendation System Using Collaborative Filtering Method. Exacta Factor, 10(1), 1–9.

[10] Son, I.M. A. W., Indrawan, G., & Aryanto, K. Y. E. (2018). Recommendation System Based on Tabanan Regional Library Transaction Data using K-Means Clustering. Indonesian Journal of

Computer Science (JIKI), 3(1), 18–22. http://119.252.161.254/e-journal/index.php/jik/article/view/2749/1314

[11] Chen, W. Y., Song, Y., Bai, H., Lin, C. J., & Chang, E. Y. (2011). Parallel spectral clustering in distributed systems. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(3), 568–586. https://doi.org/10.1109/TPAMI.2010.88