Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 453−459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1214

Improved Collaborative Filtering Recommender System Based on Missing Values Imputation on E-Commerce

Kadek Abi Satria A V P, Z K A Baizal*

Informatics, School of Computing, Telkom University, Bandung, Indonesia
Email: [1]abisatria@student.telkomuniversity.ac.id, [2,*]baizal@telkomuniversity.ac.id
Correspondence Author Email: baizal@telkomuniversity.ac.id
Submitted: **22/01/2022**; Accepted: **03/02/2022**; Published**: 31/03/2022**

Abstract−One of the important aspects in e-commerce is how to recommend a product to users accurately. To achieve this goal, many e-commerce starts to build and research about recommender system. Many methods can be used to build a recommender system, one of them is using the collaborative filtering technique. This technique often experiences data sparsity problem that can impact to the recommender system prediction accuracy. To solve this problem, we apply improved collaborative filtering. This method predicts the missing values in the user item rating matrix. First, we do an initial selection to determine potential users who have the same characteristics with the active user. After that, we calculate the average distance between the active user and the other selected user. Next, we calculate missing values prediction. Missing values predictions is only done for items that have never been rated by other's selected user but has been rated by the active user. We used Amazon electronic product with high sparsity level in this research to simulate the actual condition of e-commerce. We used MAE and RMSE to measure prediction accuracy. The methods we apply succeeds to improve the prediction accuracy compare to the conventional collaborative filtering method. The average MAE for method that we apply is 0.78 and RMSE 1.07.

Keywords: Recommender System; Collaborative Filtering; E-commerce

1. INTRODUCTION

Recommender system is a system that recommends the most relevant items to several users [1]. Recommender system technology is widely used in e-commerce sites [2], [3]. Analyzing and understanding the needs of customers is one of the important things to do in e-commerce platform. In this way, the e-commerce company knows the interests of their customer. By recommender system, e-commerce can provide a good service for the customer and almost similar services are provided by professional salesperson in offline stores [4].

Recommender system requires some data from the users such as history, personal preferences, ratings and comments [4]. Recommender system also can use several attributes of products in e-commerce such as descriptions, product names and categories. This data will be used as input for the recommender system to produce personalized recommendations for each user [5]. Personalized recommendations are an important part of the recommender system. By personalized recommendations, customers can enjoy several products that match with their characteristics and interest. It also helps user to find the product they like faster [6].

There are several methods that can be used for building recommender system, such as collaborative filtering (CF), content-based filtering (CB) and hybrid recommender system [7]. Collaborative filtering is one of the best and most used methods in building a recommender system [8]. This method find other users who give the same rating to a product. Other's user data will be used to determine the right product recommendations for the active users. There are two techniques that can be used in collaborative filtering, such as memory based and model based [9]. However, this technique facing a lot of problem [10], one of them is data sparsity problem. This problem is very common because the number of products and users in e-commerce is very large compared to the existing rating. This problem will affect the accuracy of the recommender system [11].

To overcome this problem, many studies have been conducted. In a study conducted by [12] using the Item K-nn method. The Item K-nn method makes predictions based on the similarity of each item available in e-commerce. Research [13] uses the improved item based collaborative filtering method. This method uses a characteristics combination of similar items to improve the accuracy of the recommendations. Another research is using missing values predictions in the user item rating matrix [14] or known as imputation. In real cases in building a recommender system with collaborative filtering, the user only rates a small part of all available items. On average, users only contribute to providing rating data as much as 1% of all data in the user item rating matrix. If the rating given by each user is very small, recommender system will have difficulty in calculating the similarity of each user. We can modify the user item rating matrix by adding some data to reduce the sparsity of the dataset. Modification of the user item rating matrix has not been applied in several other studies. By applying this step, is one of the high potential solutions to overcome the data sparsity problem.

To modified the user item rating matrix, we can use Imputation method. This method will predict missing values based on several condition and inputing to the user item rating matrix. Missing values predictions can be done based on data relations or using constant values [15]. Predicting with constant values is fast and easy to implement. However, filling with a constant value will cause a change in the originality characteristics of each user, so there will be no personalization for each user [14]. Using data relations in predicting missing values have some advantages compared to constant values. One of the advantages is the personalization of the predictions that will be generated for each user.

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 453−459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1214

In this research, we focus to solve the data sparsity problem in recommender system using improved collaborative filtering, which apply imputation method in user item rating matrix. This method has similar step from research that has been conducted by [14]. But in our research, we do data preprocessing at first step to reduce the computational time of the recommender system. First step, data preprocessing is done by selecting users who have number of items rated above the specified parameters. After that, a user selection is made based on the preferences of the active user. This step will select users who have the same number of items rated as active users. After the selection, the calculation is carried out to find the average distance between active users and the other's selected users. Next step is predicting missing values. This step is done by calculating the predicted value for items that have not been rated by other users but have been rated by active users. The value to be imputed is obtained from the calculation of the rating value of the active user and the average distance between the active user and other's selected users.

The purpose of this research is to build a recommender system that has smaller error value than the conventional collaborative filtering and matrix factorization techniques. This comparison aims to prove that the method applied in this research is better in overcoming data sparsity problem.

2. RESEARCH METHODOLOGY

2.1 System Design

In this research, we use electronic products dataset from Amazon [16]. The dataset contains a total of 1,700,000 rating data for 200,000 products. This dataset has a sparsity level up to 99.80%. It shows that the sparsity in the dataset is very high. We only used 200.000 ratings data in the dataset to reduce the computation of the recommender system. We use the 5-fold cross validation technique in each experiment to obtain more valid research results. The stages in this research are generally shown in Figure 1. Based on Figure 1, this research has 8 important step to build a recommender system.

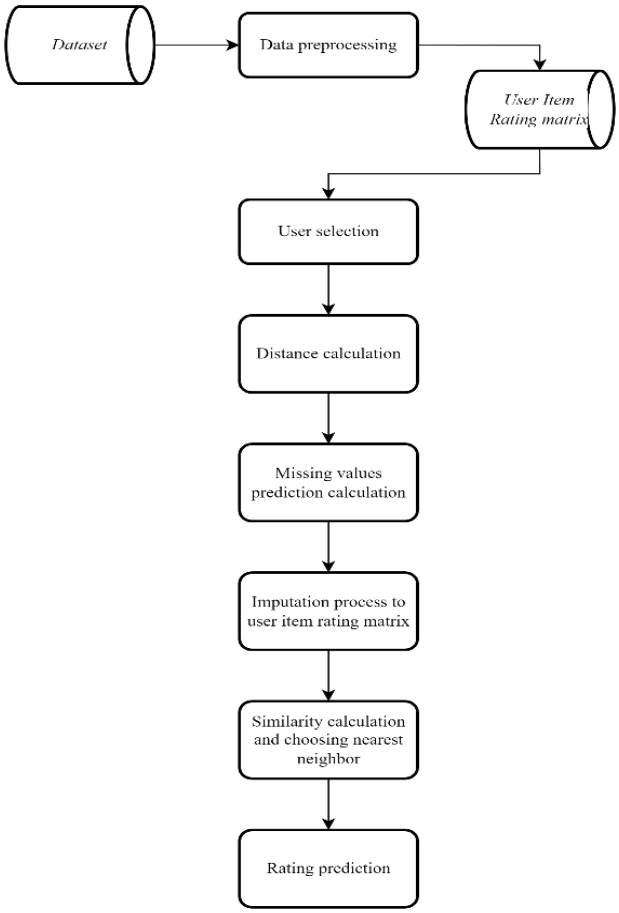

Figure 1. The Design

In the dataset, data preprocessing is carried out first. Dataset is filtered by selecting users who have number of items rated greather than n. The determination of n will be adjusted to the conditions of the dataset. After that, we select users who have the same number of items rated as active users. We provide an example of a user item rating matrix in Table 1 to understand the solution applied in this research.

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 453−459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1214

Table 1. User item rating matrix example

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

In Table 1, U1 is assumed to be an active user. User selection is done by looking at the number of ratings on the same item as the active user. If the number of rating items of the sam item given is greater than the specified parameter θ, then the user is selected as a potential user for calculations in the next step. In table 1, we assume that the parameter θ is 2, so U3 will be removed from the user item rating matrix and only U2 and U4 will be left.

The next step is to calculate the average distance between active users and selected users. The purpose of this step is to find the closeness of the characteristics between active users and other's selected users. After that, we calculate the predicted missing values which will be imputed in the user item rating matrix. The calculations are based on item ratings that have not been given by other users but have been rated by active users. In this example, we will predict missing values on the U2 rating for I5 and U4 for I1. The calculated value will use the average distance between active users and other users that has been calculated in the previous step. Finally, we calculated similarity using the Pearson correlation coefficient (PCC) and make recommendation according to user-based collaborative filtering.

2.2 Data Preprocessing

Dataset Preprocessing is done by selecting users who have number of items rated greater than n. The value of n will be determined by the developer who wants to implement this research based on conditions of the dataset.

2.3 User Selection

This step is carried out to help the recommender system more easily to choose other users who may have similar preferences to active users. One of the parameters used in user selection is θ. Selection is made on the user item rating matrix, if the same number of items rated by other users is less than θ, it will be removed from the user item rating matrix.

$$UIR\ Matrix = \begin{cases} remove, & if\ r_{a,i} \cap r_{b,i} = k \\ & 0 \le k < \theta \\ keep, & if\ r_{a,i} \cap r_{b,i} = k \\ & k \ge \theta \end{cases} \quad (1)$$

Where $r_{a,i}$ is the rating given by active users to item i, $r_{b,i}$ is the rating given by other users to item i and k is total same items that has been rated by active and other users.

2.4 Calculate Distance Between Users

The similarity between active users and other users can be calculated by the difference from the rating given to items. The calculation is done by calculating all the differences in the rating given by active users with other users.

$$diff(U_a, U_b) = \sum_{i=1}^{m} |r_{a,i} - r_{b,i}| \quad (2)$$

Where U_a is an active user and U_b is other user that has been selected and m is total same item that is rated by the active users and other users. Next, the average of each rating difference will be calculated using formula 3.

$$AVG(U_a, U_b) = \frac{U\ diff(U_a, U_b)}{m} \quad (3)$$

Where m is total same item that is rated by the active users and other users.

2.5 Missing Values Predictions

The missing value prediction calculation is done by filling all ratings that not have been rated by other users but has been rated by the active users. The calculation is show in formula 4.

$$fill(U_b, i \to 0) = |r_{a,i} - AVG(U_a, U_b)| \quad (4)$$

Where $fill\ (U_b, i \to 0)$is empty rating data from other users and has been rated by active users.

2.6 Similarity Calculation

Similarity calculation in this research using pearson corellation coefficient [17]. This similarity has a value from -1 to +1. If the relationship between users is very strong, it will be marked close to +1, if the relationship between users is negative, it will be marked close to -1 and if the relationship between users is not related at all, it will be marked with a value of 0.

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 453−459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1214

$$Sim_{a,b} = \frac{\sum_{i=1}^{n}(r_{a,i} - \overline{r_a})(r_{b,i} - \overline{r_b})}{\sqrt{\sum_{i=1}^{n}(r_{a,i} - \overline{r_a})^2 (r_{b,i} - \overline{r_b})^2}} \quad (5)$$

Where r_a is the average rating of user a and r_b is the average rating of user b.

2.7 Rating Prediction

After calculating the similarity of each user, then predictions are made on the items that will be liked by active users. The calculation will use the formula 6.

$$P_{a,i} = \overline{r_a} + \frac{\sum_{b=1}^{n}(r_{b,i} - \overline{r_b}) Sim_{a,b}}{\sum_{b=1}^{n} Sim_{a,b}} \quad (6)$$

Where $P_{a,i}$ is the prediction of item i to the active user, $\text{Sim}_{a,b}$ is the similarity value of user a and user b and n is the total number of active user nearest neighbor.

3. RESULT AND DISCUSSION

In this research we use electronic products dataset from Amazon. We used 200,000 rating data from the dataset. In this research we used 5-fold cross validation method to obtain more accurate results. Datasets was divided into 4 parts, dataset containing 50,000 rating data, 100,000 rating data, 150,000 rating data and 200,000 rating data. This division aims to find out how the performance and prediction results generated by the recommender system for different dataset conditions. The distribution of training and testing data is 70:30. We used collaborative filtering (CF) and matrix factorization (MF) techniques to compare with improved collaborative filtering (ICF) method. To measure the error, we used mean absolute error (MAE) and root mean square error (RMSE).

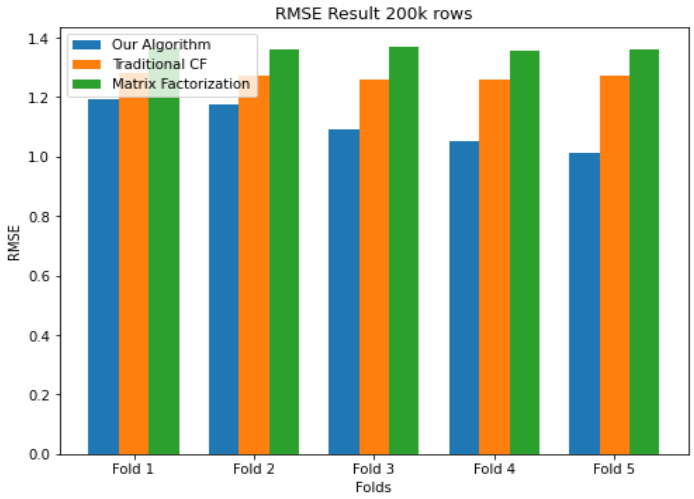

Figure 2. RMSE Result 200k Rows

Figure 2. Shows the RMSE results obtained from research using 200,000 rating data. From these results, the smallest error value obtain by ICF is in the 5th fold with a value of 1.01. The smallest error value obtained by CF is in the 4th fold with a value of 1.25. The smallest error value obtained by MF is in the 4th fold with a value of 1.35.

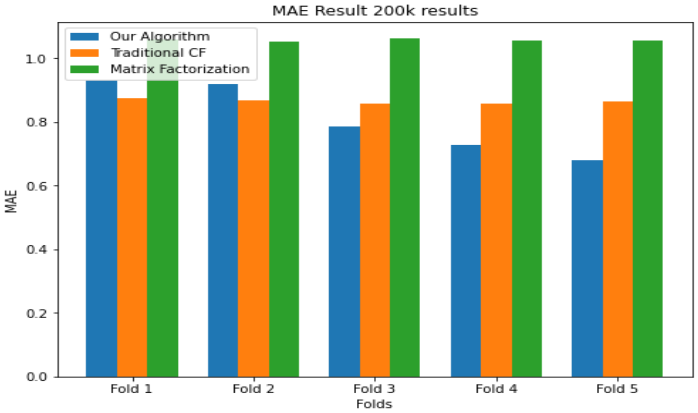

Figure 3. MAE Result 200k Rows

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 453−459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1214

Figure 3 shows the MAE results obtained from research using 200,000 rating data. From these results, the ICF algorithm gets the smallest error value in the 5th fold with a value of 0.67. The CF algorithm gets the smallest error value in the 3rd fold with a value of 0.85. The MF algorithm gets the smallest error value in the 2nd fold with a value of 1.05.

Table 2. Average error value using 200k rows rating

Evaluation Matrix	ICF	CF	MF
MAE	0,80	0,86	1,05
RMSE	1,10	1,26	1,36

We obtained the results of average error value using 5-fold cross validation method in 200k rows dataset as shown on table 2. The average MAE error value obtained by the ICF algorithm is the smallest compared to the CF and MF algorithms. For the average RMSE error value, the ICF algorithm gets the lowest error value compared to CF and MF.

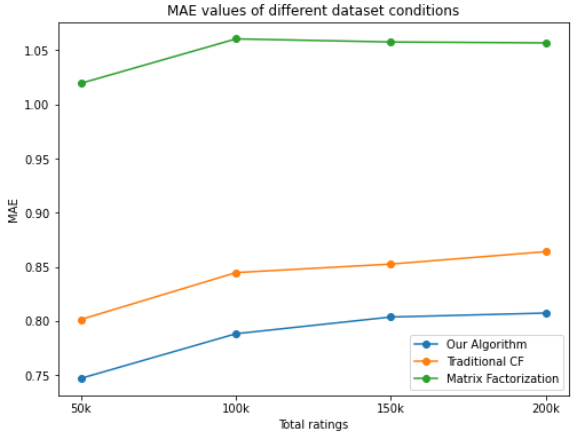

Figure 4. MAE Values

Figure 4 shows the results obtained by the recommender system with a different amount of data used. In experiment using 50k data, we used parameters n = 3 and θ = 3 resulting MAE value of 0.75. In experiment using 100k data, we used the parameters n = 3 and θ = 5 resulting MAE value of 0.79. In the experiment with 150k data, we used the parameters n = 3 and θ = 5 resulting MAE value of 0.80. In the last experiment with 200k data we used the parameters n = 3 and θ = 5 resulting in MAE value of 0.81.

The determination of these parameters is adjusted to the conditions of the dataset to achieve optimal results. Figure 4. shows that the MAE value of the ICF method is smaller than that of the CF and MF methods. The comparison of MAE values obtained by ICF and CF is almost the same with a difference of 0.05. The error value obtained by ICF has a large difference when compared to the MF algorithm.

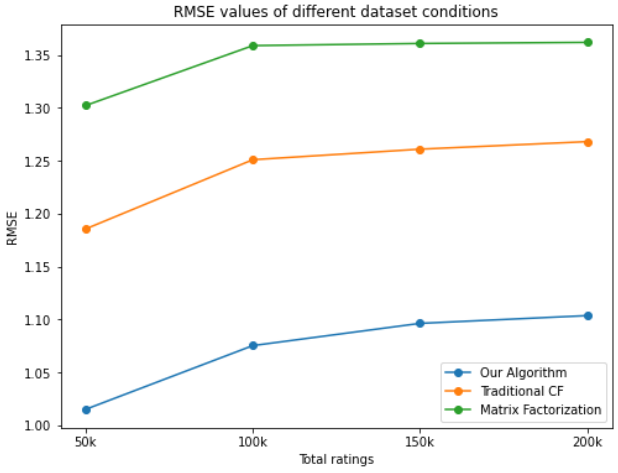

Figure 5. RMSE Values

Figure 5. Shows the RMSE value comparison. The error value obtained by ICF is the lowest compared to CF and MF. In the 50k data experiment, RMSE value for ICF is 1.01. In the 100k data experiment, RMSE value for ICF is 1.07. In the 150k data experiment, RMSE value for ICF is 1.09. In the last experiment using 200k data, RMSE value for ICF is 1.10. The comparison between the RMSE and MAE error results is slightly different. In Figure 5. Shows

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 453−459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1214

that the difference in RMSE obtained by CF and ICF is more high compare to MAE result. This RMSE result shows that the CF algorithm makes predictions that are quite far compared to the original rating given by the user.

4. CONCLUSION

Based on the research that has been done, it can be concluded that the improved collaborative filtering method in the e-commerce domain has improved the prediction accuracy compared to the traditional collaborative filtering and matrix factorization methods. Parameter n and θ affect the prediction accuracy. Each dataset has different optimal parameter values, so it is necessary to do several experiments to find the optimal value. This parameter is used to select the reliable users also to reduce the computation of the recommender system. The research results obtained using 5-fold cross validation and using different amounts of data. The results show that the improved collaborative filtering has the lowest error value in each experiment. The average MAE value for ICF is 0.78 and the average RMSE value for ICF is 1.07. The results shows that the predicted rating generated by ICF is more stable, the predicted rating value is not much different from the original rating given by the user. This statement can be seen from the results of the RMSE value between ICF, conventional collaborative and matrix factorization. In experiments using different amounts of data, it can be concluded that increasing the amount of data is not always increase the recommender system accuracy. Three methods that we tested experienced a decrease in accuracy, because the amount of data we used does not always reduce the sparsity level of the dataset. For further research, the improvement of the algorithm in this research is still very possible. Further research that can be done is to use more different dataset conditions and more comparison methods. And also, changing the parameter for first step in this method to obatin optimal result for the recommender system.

REFERENCES

[1] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," Knowledge-Based Systems, vol. 46, pp. 109–132, Jul. 2013, doi: 10.1016/j.knosys.2013.03.012.
[2] A. Hidayatullah and M. A. Anugerah, "A Recommender System for E-Commerce Using Multi-objective Ranked Bandits Algorithm," in 2018 International Conference on Computing, Engineering, and Design (ICCED), Sep. 2018, pp. 170–174. doi: 10.1109/ICCED.2018.00041.
[3] D. S. Ken Arnett, Z. K. A. Baizal, and Adiwijaya, "Recommender system based on user functional requirements using Euclidean fuzzy," in 2015 3rd International Conference on Information and Communication Technology (ICoICT), May 2015, pp. 455–460. doi: 10.1109/ICoICT.2015.7231467.
[4] X. Zhao, "A Study on E-commerce Recommender System Based on Big Data," in 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Apr. 2019, pp. 222–226. doi: 10.1109/ICCCBDA.2019.8725694.
[5] F. Abbas and X. Niu, "Computational Serendipitous Recommender System Frameworks: A Literature Survey," in 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Nov. 2019, pp. 1–8. doi: 10.1109/AICCSA47632.2019.9035339.
[6] Z. K. A. Baizal, D. H. Widyantoro, and N. U. Maulidevi, "Computational model for generating interactions in conversational recommender system based on product functional requirements," Data & Knowledge Engineering, vol. 128, p. 101813, Jul. 2020, doi: 10.1016/j.datak.2020.101813.
[7] B. Patel, P. Desai, and U. Panchal, "Methods of recommender system: A review," in 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Mar. 2017, pp. 1–4. doi: 10.1109/ICIIECS.2017.8275856.
[8] D. Kluver, M. D. Ekstrand, and J. A. Konstan, "Rating-based collaborative filtering: Algorithms and evaluation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10100 LNCS, Springer Verlag, 2018, pp. 344–390. doi: 10.1007/978-3-319-90092-6_10.
[9] M. I. Ardimansyah, A. F. Huda, and Z. K. A. Baizal, "Preprocessing matrix factorization for solving data sparsity on memory-based collaborative filtering," in 2017 3rd International Conference on Science in Information Technology (ICSITech), Oct. 2017, pp. 521–525. doi: 10.1109/ICSITech.2017.8257168.
[10] M. R. Zarei and M. R. Moosavi, "A Memory-Based Collaborative Filtering Recommender System Using Social Ties," in 2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA), Mar. 2019, pp. 263–267. doi: 10.1109/PRIA.2019.8786023.
[11] Z. Fayyaz, M. Ebrahimian, D. Nawara, A. Ibrahim, and R. Kashef, "Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities," Applied Sciences, vol. 10, no. 21, p. 7748, Nov. 2020, doi: 10.3390/app10217748.
[12] A. Sang and S. K. Vishwakarma, "A ranking based recommender system for cold start & data sparsity problem," in 2017 Tenth International Conference on Contemporary Computing (IC3), Aug. 2017, pp. 1–3. doi: 10.1109/IC3.2017.8284347.
[13] P. Yu, "Merging Attribute Characteristics in Collaborative Filtering to Alleviate Data Sparsity and Cold Start," in 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Mar. 2019, pp. 569–573. doi: 10.1109/ITNEC.2019.8729461.
[14] M. A. Hassan, M. G. M. Johar, and A. I. Hajamydeen, "A Framework for Recommender Systems Using Improved Collaborative Filtering," in 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), Jun. 2019, pp. 168–173. doi: 10.1109/I2CACIS.2019.8825046.
[15] T. Anwar, T. Siswantining, D. Sarwinda, S. M. Soemartojo, and A. Bustamam, "A study on missing values imputation using K-Harmonic means algorithm: Mixed datasets," in AIP Conference Proceedings, Dec. 2019, vol. 2202. doi: 10.1063/1.5141651.

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 453−459
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1214

[16] R. He and J. McAuley, "Ups and Downs," in Proceedings of the 25th International Conference on World Wide Web, Apr. 2016, pp. 507–517. doi: 10.1145/2872427.2883037.
[17] Y. Liu, Y. Mu, K. Chen, Y. Li, and J. Guo, "Daily Activity Feature Selection in Smart Homes Based on Pearson Correlation Coefficient," Neural Processing Letters, vol. 51, no. 2, pp. 1771–1787, Apr. 2020, doi: 10.1007/s11063-019-10185-8.