JITEL (Jurnal Ilmiah Telekomunikasi, Elektronika, dan Listrik Tenaga) p-ISSN: 2774-7972
Vol. 2, No. 2, September 2022, pp. 103-110 e-ISSN: 2775-6696
DOI: https://doi.org/10.35313/jitel.v2.i2.2022.103-110

Singular value decomposition model application for e-commerce recommendation system

Wervyan Shalannanda[1*], Rafi Falih Mulia[2], Arief Insanu Muttaqien[3], Naufal Rafi Hibatullah[4], Annisabelia Firdaus[5]
[1,2,3,4,5]School of Electrical Engineering and Informatics, Institut Teknologi Bandung
Jl. Ganeca 10 Bandung, West Java 40132, Indonesia
[1*]wervyan@itb.ac.id, [2]18118010@telecom.stei.itb.ac.id, [3]18118014@telecom.stei.itb.ac.id,
[4]18118019@telecom.stei.itb.ac.id, [5]18118040@telecom.stei.itb.ac.id

ABSTRACT

A recommendation system is one of the most important things in today's technology. It can suggest products that match the user's preferences. Many fields utilize this system, including e-commerce, using various algorithms. This paper used the matrix factorization-based algorithm, singular value decomposition (SVD), to make a recommendation system based on users' similarities. Afterward, we implement the model against the ModCloth Amazon dataset. The results imply that the SVD algorithm yields the best accuracy compared to other matrix factorization-based algorithms with root mean square error (RMSE) of 1.055586. Then, we optimized the SVD algorithm by changing the hyperparameters of the algorithm to generate better accuracy and yield a model with an RMSE value of 1.041784.

Keywords: e-commerce, recommendation system, matrix factorization, SVD, RMSE

ABSTRAK

Sistem rekomendasi adalah salah satu bagian yang penting bagi teknologi saat ini. Sistem ini dapat menawarkan berbagai produk yang sesuai dengan preferensi pengguna dan digunakan luas di berbagai bidang, termasuk e-commerce. Ada beberapa algoritma yang dapat digunakan untuk membuat suatu sistem rekomendasi. Paper ini menawarkan penggunaan algoritma berbasis faktorisasi matriks, yaitu SVD (singular value decomposition) untuk membat sistem rekomendasi berdasarkan kemiripan perilaku antar user. Model yang dibangun kemudian digunakan pada dataset ModCloth Amazon. Hasil dari penelitian ini menunjukkan bahwa algoritma SVD menghasilkan akurasi yang lebih baik dibandingkan dengan algoritma berbasis faktorisasi matriks lain yang ditunjukkan dengan nilai RMSE (root mean square error) terendah, yaitu 1,055586. Kemudian, peneliti melakukan optimasi dengan mengubah hyperparameters algoritma dengan akurasi lebih baik, yaitu model dengan nilai RMSE 1,041784.

Kata kunci: *e-commerce, sistem rekomendasi, faktorisasi matriks, SVD, RMSE*

1. INTRODUCTION

Electronic commerce (e-commerce) is the activity of buying and selling products or services through the internet. With the rapid development of technology, e-commerce has become an inseparable part of people's daily life [1], [2]. E-commerce provides a wide variety of products and services that allow users to choose the thing that matches their preferences. On the other hand, there have been so-called "information overload." It is difficult for users to quickly and accurately find the right item when faced with a huge amount of product information [3], [4].

From the explanation above, we realized that providing products that suit the needs and desires of the user is very important. Therefore, a recommendation system is needed to provide users with product choices from the e-commerce platform. In addition, the recommendation system must match user preferences with personalized product recommendations to various users. That way, this e-commerce product recommendation system can make it easier and reduce user time to choose products that suit user needs and desires. The e-commerce platform can also make users more consumptive by providing attractive product recommendations to users so that the revenue of the e-commerce platform will increase.

Naskah diterima tanggal 29 Juli 2022, disetujui tanggal 15 September 2022
*E-mail korespondensi

The recommendation system is part of the information filtering system that predicts the user's rating or preference by analyzing multiple users' data points and displays items that users might like. There are three types of recommendation systems, namely: collaborative filtering, content-based filtering, and hybrid recommendation systems. We chose collaborative filtering as the method used in this paper.

Collaborative filtering is collecting and analyzing large amounts of information about a user's behavior, activity, or preference and predicting what users will like based on what they have in common with other users. The main advantage of a collaborative filtering approach is that it does not rely on machine-analyzable content and can accurately recommend complex items without requiring an "understanding" of the item itself. Collaborative filtering consists of memory-based and model-based methods. Many algorithms are used to measure the similarity of users or items in the recommendation system. A collaborative filtering model is chosen using the singular value decomposition (SVD) algorithm in this paper.

SVD is a matrix factorization technique whose main objective is to reduce the number of features of a data set by reducing the spatial dimensions from N to K, where K < N. Matrix factorization can be considered as a method that finds two matrices and the result of multiplication between those two matrices is its original matrix [5]. For the recommendation system, the matrix factorization keeps the same dimensions. Matrix factorization is carried out on the user-item rating matrix.

2. RESEARCH METHOD

2.1 Algorithm

Recommendation systems can be classified into three categories: content-based, collaborative, and hybrid filtering [6]. Content-based filtering uses item features to recommend other items similar to the user's likes, based on their previous actions or explicit feedback [7]., Collaborative filtering uses similarities between users and items simultaneously to provide recommendations to address some of the limitations of content-based filtering. This mechanism allows for serendipitous recommendations; that is, collaborative filtering models can recommend an item to user A based on the interests of a similar user B. Furthermore, the embeddings can be learned automatically without relying on the hand-engineering of features [8]. Meanwhile, hybrid filtering combines content-based recommendation systems and collaborative filtering.

In this paper, collaborative filtering will be used to make product recommendations. Collaborative filtering can be classified into two categories, memory-based and model-based. The memory-based method performs the recommendation process by accessing the database directly. In contrast, the model-based method uses transaction data to create a model that can generate recommendations [8]. Model-based collaborative filtering will be used in this paper. There are several kinds of algorithms in model-based collaborative filterings, e.g., K-Nearest Neighbour (KNN), PMF (Probabilistic Matrix Factorization), NMF (Non-Negative Matrix Factorization), SVD, and Multi-layer Neural Networks. KNN is one of the collaborative filtering methods that use a clustering-based algorithm as its generic modeling approach. In contrast, PMF, NMF, and SVD use a matrix factorization-based algorithm. Lastly, Multi-layer Neural Networks use deep learning as their generic modeling approach.

This paper used the SVD algorithm, a technique that has been generally used to reduce the dimensions of a matrix. It is a matrix factorization technique that takes an m × n matrix A into a multiplication of three matrices as follows:

$$A = U \times S \times V^T \tag{1}$$

where A is a utility matrix, U and V are orthogonal matrices with dimensions m × r and r × n, respectively, describing the relationship between users and items with latent factors (such as a user or item characteristics). V^T denotes the transpose of matrix V, while S is a singular or diagonal matrix with dimensions r×r that represents the strength of latent factors. The diagonal matrix contains all singular values of matrix A whose diagonal entries are positive real numbers in decreasing order [10] [11].

2.2 Library and Tools

This paper uses several python libraries in designing the recommendation system. First, the NumPy library is used. NumPy is used to perform numerical computations on the data, such as creating arrays and doing mathematical calculations. The next library that is used is Pandas. Pandas are used to perform data manipulation and data analysis. For visualization purposes, the Seaborn library is used. This library provides a high-level interface for drawing interactive and informative statistical charts [12].

In creating and evaluating models, the Surprise library is used. Surprise is a Python SciKit for building and analyzing recommendation systems that deal with explicit rating data. In the surprise library, various kinds of algorithms can be used to make recommendation systems, such as baseline algorithm, matrix-factorization-based algorithm (SVD, PMF, SVD++, NMF), and similarity measures (cosine, Pearson, SMD/difference-based similarity measure), and many others [13]. The Surprise library also provides for evaluating and analyzing the model that has been created. We used a cross-validation procedure to measure the model's root mean square error (RMSE). After that, the algorithm hyperparameter is tuned to improve the model accuracy. GridSearchCV is used to look for parameter combinations that yield the best results.

2.3 Dataset

This paper uses the ModCloth Amazon dataset from Kaggle. The dataset consists of some products sold on ModCloth Amazon which are equipped with the username, item, and rating of the product purchased, and there is a timestamp which means the time when the user purchased the product. In this dataset, information is also provided about the products purchased by the user, which are the size, fit, attribute, and brand. These components will be very useful in creating a recommendation system. Figure 1 shows the sample data of the aforementioned dataset.

	item_id	user_id	rating	timestamp	size	fit	user_attr	model_attr	category	brand
0	7443	Alex	4	2010-01-21 08:00:00+00:00	NaN	NaN	Small	Small	Dresses	NaN
1	7443	carolyn.agan	3	2010-01-27 08:00:00+00:00	NaN	NaN	NaN	Small	Dresses	NaN
2	7443	Robyn	4	2010-01-29 08:00:00+00:00	NaN	NaN	Small	Small	Dresses	NaN
3	7443	De	4	2010-02-13 08:00:00+00:00	NaN	NaN	NaN	Small	Dresses	NaN
4	7443	tasha	4	2010-02-18 08:00:00+00:00	NaN	NaN	Small	Small	Dresses	NaN

Figure 1. Dataset overview

The $item\_id$ column displays the code of the item type that the user has purchased; the $user\_id$ column displays the username of the buyer, $rating$ is the rating of a product purchased by $user\_id$, $and\ timestamp$ is the time when a user buys the product. $Size$, fit, $user\_attr$, $model\_attr$, $category$, and brand is information about the products purchased by the user. For $size$, fit, $user\_attr$, and $model\_attr$ are the units for size. $Category$ and $brand$ are the types of the product.

The dataset has a size of 99893 x 10; there are 99893 rows and ten columns. Because there are 99893 rows, the number of unique products is only 1020. The unique value of $item\_id$ is 1020, which means there are 1020 different products in this dataset. The unique value of $user\_id$ is 44783; there are 44783 people involved in this dataset. The number of unique $user\_id$ is 44783, which means that there are several products that users continuously purchase. Some users continue to buy products repeatedly on ModCloth Amazon.

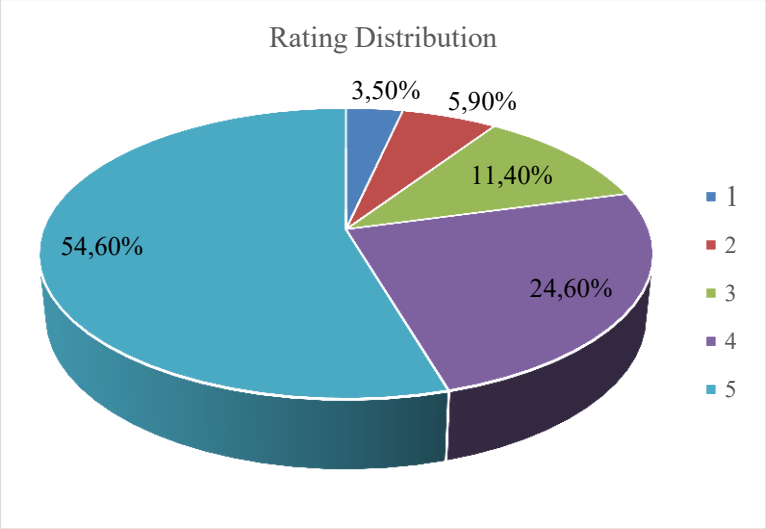

Figure 2. Rating distribution graph

The dataset has a rating distribution, as in Figure 2. The x-axis is the rating, and the y-axis is the number of users who choose the rating. The higher the rating value, the more users will choose the rating. More users think that the products on Amazon's ModCloth are equivalent to 5-star class or very good quality. Meanwhile, some fewer users choose 1-4 ratings.

3. RESULTS AND DISCUSSION

To check the accuracy of the algorithm used, the RMSE (root mean squared error) metric is used as the evaluation metric. RMSE is a widely used evaluation metric that has been proven to be effective for measuring recommendation system performance. RMSE results show the difference between the predicted results and the actual data. Then, RMSE is defined as follows [5]:

$$RMSE = \sqrt{\frac{\sum_{(\{i,j\} \in TestSet)} (r_{ij} - \widehat{r_{ij}})}{T}} \quad (2)$$

where T represents the total number of samples in the dataset, r_{ij} is the actual rating given by user i to item j, and $\widehat{r_{ij}}$ is the result of the predicted rating of user i to item j. For the evaluation method, the k-fold cross-validation method to avoid random choosing and to get better accuracy. Cross-validation is a statistical method to estimate the quality of a machine learning model on new data, while k value denotes the number of partitions of the dataset. There is no formal rule to determine the k value, but we need to consider the bias-variance trade-off as one of the consequences of choosing k value. References [14] and [15] stated that the choice of k is usually 5 or 10 and these values have been shown to yield neither very high bias nor excessive variance in the test error rate. This cross-validation method is used to avoid random choosing and to get better accuracy. This paper used $k = 5$.

Table 1. Model results from different matrix factorization algorithms

Algorithm	test_rmse	fit_time	test_time
SVD	1.055586	5.201821	0.183002
SVDpp	1.060906	25.681538	1.013987
CoClustering	1.37415	6.258202	0.175923
SlopeOne	1.149594	0.606814	0.563476
NMF	1.195050	8.99335	0.152779

This research solely used RMSE to evaluate the accuracy of the models. A large RMSE value indicates that the resulting error will be greater, so to find out which matrix factorization algorithm method has the best quality and must be chosen, it can be seen from the method that has the smallest

RMSE value. To determine the best matrix factorization algorithm method among SVD, SVDpp, CoClustering, SlopeOne, and NMF, we compared the respective RMSE values.

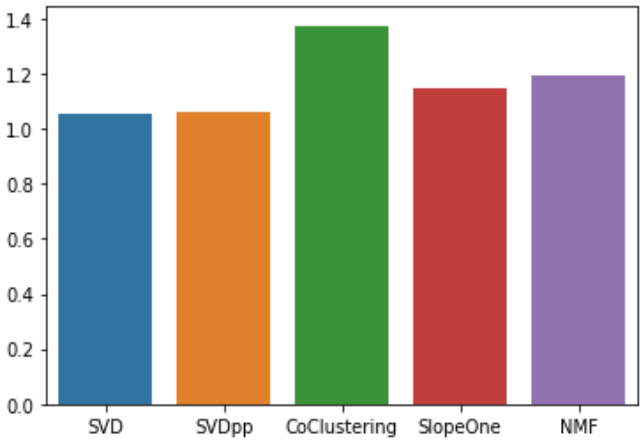

Figure 3. RMSE comparison between different matrix factorization algorithms

Table 1, in the *test_rmse* column, and Figure 3 show that SVD has the smallest RMSE compared to other algorithms, which is 1.055586. It can be concluded that SVD has the best accuracy among matrix factorization algorithm methods in this research. Using SVD as a method to create an e-commerce recommendation system is the right choice.

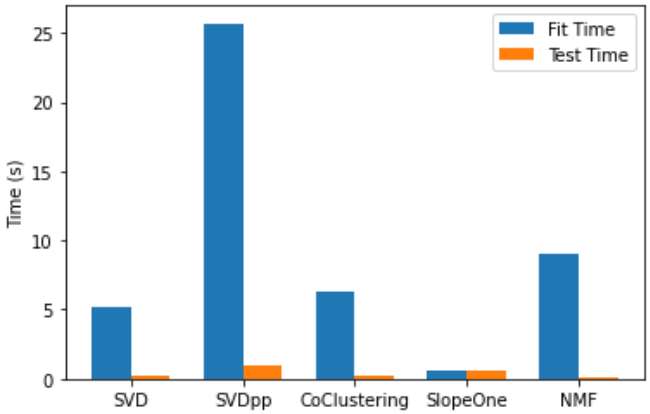

Figure 4. Fit and time comparison between different matrix factorization algorithms

Figure 4 shows that the fit time is always much greater than the test time in each matrix factorization algorithm; this indicates that the k-fold method used here has worked well because the time allocated for fit or training should be longer, which indicates the processed training data is much greater compared to test data. Since SVD and SVDpp are the top two algorithms that are roughly neck and neck in terms of accuracy, fit time and test time can be utilized as tie-breakers. Furthermore, these parameters need to be considered in the deployment phase since these parameters represent the delay that will affect the user experience. Figure 4 also shows that fit time and test time of SVD are significantly shorter than the fit time and test time of SVDpp. In this matter, SVD is only inferior to SlopOne's fit time and NMF's test time. However, since the accuracy of both algorithms is quite inferior to SVD, we can conclude that the best algorithm in this research is SVD.

After knowing that SVD is the best algorithm, the SVD algorithm can be optimized with GridSearchCV. In GridSearchCV, some parameters can be changed by adding options to the common parameter values to get the best RMSE value. Table 2 shows the first optimization with the parameters being compared and their value variation.

Table 2. Variations of parameter values in the first optimization

Parameter	Value variation		
Epochs	5	10	15
Learning rate	0.002	0.005	0.01
Regularization	0.02	0.4	0.6

Then it is optimized again with other parameters to get a more optimal value. Table 3 shows the parameters that are compared and their value variations in the second optimization. From the variation in the parameter value, the combination of parameters that gave the best RMSE score is epochs = 15, learning rate = 0.005, and regularization = 0.4 with an RMSE value of 1.045075.

Table 3. Variations parameter values in the second optimization

Parameter	Value variation		
Factors	50	100	150
Epochs	20	25	30
Learning rate	0.005	0.00075	0.01
Regularization	0.2	0.4	0.6

From the results of the second optimization, a better combination of parameters is obtained with an RMSE value of 1.041784. So, the recommendation system to be used is the last optimized SVD algorithm with the best combination of parameters, with factors = 50, epochs = 20, learning rate = 0.005, and regularization = 0.2. Compared to other algorithms, the optimized SVD algorithm is the best algorithm with the smallest RMSE value.

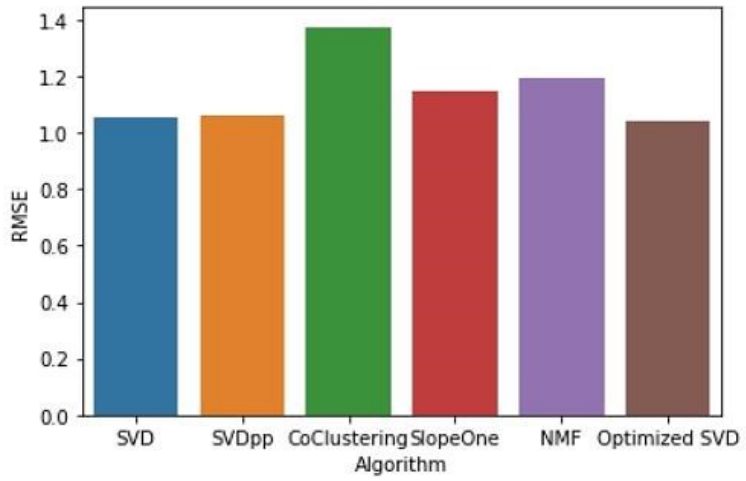

Figure 5. RMSE Comparison between optimized SVD algorithms and other matrix factorization algorithms

Following the best model parameter acquired previously, the next step is to apply the model to the recommendation system. The model was fitted into the test data, which will generate predictions of the rating the user gives the items. Table 4 shows the five best predictions of the model. From those prediction results, it can be seen that some of the predictions are equal to the original rating the user gives to the item, which means there is no error in the model for the relevant item.

Table 4. Best predictions

User ID	Item ID	Rating	Prediction	Error
kendra	153630	5.0	5.0	0
shelley	40899	5.0	5.0	0
leiafeliz22	129393	5.0	5.0	0
Jo	67022	5.0	5.0	0
kimberlysieglaff	122266	5.0	5.0	0

Table 5. Worst predictions

User ID	Item ID	Rating	Prediction	Error
R	64745	1.0	4.964602	3.964602
Ann	64921	1.0	4.764354	3.764354
April	80238	1.0	4.741017	3.741017
Andie	76049	1.0	4.738409	3.738409
Nini	153473	1.0	4.696421	3.696421

Table 6. Recommended items for user "Sarah"

Item ID	Prediction
133926	4.686671
132686	4.620667
153380	4.598588
153118	4.593400
154064	4.589076

Table 5 shows the five worst predictions of the model for comparative purposes. The prediction results show that some predictions have big differences from the actual rating, with the worst error generated being nearly 4. It is further analyzed that all of the best predictions have actual rating values of 5.0, whereas all of the worst predictions have actual rating values of 1.0. Intuitively, it is because items with 5-star ratings have much more data than those with 1-star ratings, with 54.626% and 3.527%, respectively. Aside from giving rating predictions to items, the recommendation system may generate items that are recommended for users. Table 6 shows the recommended items for user "Sarah" and their ratings that are well-matched to "Sarah" according to the model.

4. CONCLUSION

This paper uses the SVD algorithm to create an e-commerce recommendation system. SVD is a matrix factorization-based algorithm that makes a recommendation system according to user similarity. The results show that SVD yields the best accuracy compared to other matrix factorization-based algorithms with RMSE 1.055586. Then, the SVD algorithm is optimized by changing the hyperparameters of the algorithm to generate better accuracy and yield a model with an RMSE value of 1.041784. The model was constructed using SVD algorithm with the best combination of parameters, with factors = 50, epochs = 20, learning rate = 0.005, and regularization = 0.2. There are two points to be conducted in future works: the improvement of model accuracy and further analysis of inference/recommendation results. The authors used an empirical and quasi-random approach to determine the value variation of several parameters in this research. In that regard, a more systematic approach might have the potential to obtain a better accuracy of the model. For the second point, we expect a further evaluation of the model so that the model gives a proper result even to infer other datasets.

REFERENCES

[1] Y. X. Qing, "An Intelligent E-Commerce Recommendation Algorithm Based on Collaborative Filtering Technology," in *2014 7th International Conference on Intelligent Computation Technology and Automation*, Changsha, 2014, pp. 80-83.

[2] X. Wang and C. Wang, "Recommendation system of e-commerce based on improved collaborative filtering algorithm," in *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, 2017, pp. 332-335.

[3] J. H. Wu, Q. Liu, and S. W. Luo, "Clustering Technology Application in e-Commerce Recommendation System," in *2008 International Conference on Management of e-Commerce and e-Government*, Nanchang, 2008, pp. 200-203.

[4] J. Zhou, F. Wan, and R. Jing, "Model and Implementation of E-commerce Recommendation System Based on User Clustering," in *2020 International Conference on Advance in Ambient Computing and Intelligence (ICAACI)*, Ottawa, 2020, pp. 197-200.

[5] J. Xia, "E-Commerce Product Recommendation Method Based on Collaborative Filtering Technology," in *2016 International Conference on Smart Grid and Electrical Automation (ICSGEA)*, Zhangjiajie, 2016, pp. 90-93.
[6] (2021) "Content-based Filtering," Google Developers. [Online]. Available: https://developers.google.com/machine-learning/recommendation/content-based/basics
[7] (2021) "Collaborative Filtering," Google Developers. [Online]. Available: https://developers.google.com/machine-learning/recommendation/collaborative/basics
[8] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-Based Sytems*, vol. 46, pp. 109-132, 2013.
[9] N. Akter, R. Mustafa, M. S. Chowdhury, and A. S. Hoque, "Accuracy analysis of recommendation system using singular value decomposition," in *2016 19th International Conference on Computer and Information Technology (ICCIT)*, Dhaka, 2016, pp. 405-408.
[10] (2021), V. Kumar, "Singular Value Decomposition (SVD) & Its Application In Recommender System," Analytics India Magazine. [Online]. Available: Singular Value Decomposition (SVD) & Its Application In Recommender System
[11] (2021) M. Waskom, "seaborn: statistical data visualization," Seaborn. [Online]. Available: https://seaborn.pydata.org/
[12] N. Hug, "Surprise: a Python scikit for recommender systems," Surprise. [Online]. Available: http://surpriselib.com/
[13] X. Guan, C.-T. Li, and Y. Guan, "Matrix Factorization With Rating Completion: An Enhanced SVD Model for Collaborative Filtering Recommender Systems," *IEEE Access*, vol. 5, pp. 27668-27678, 2017.
[14] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R, 2nd ed.*, New York: Springer, 2021.
[15] M. Kuhn and K. Johnson, *Applied Predictive Modeling, 1st ed. 2013, Corr. 2nd printing 2018 Edition ed.*, New York: Springer, 2018.