Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 544−551
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2110

Implementation of Dimensionality Reduction with SVD to Improve Rating Prediction in Recommender System

M. Naufal Mu'afa, Z.K.A Baizal*

School of Computing, Informatics, Telkom University, Bandung, Indonesia
Email: [1]mnaufalmuafa@student.telkomuniversity.ac.id, [2,*]baizal@telkomuniversity.ac.id
Submitted: **15/08/2022**; Accepted: **29/08/2022**; Published: **30/08/2022**

Abstrak−Sistem pemberi rekomendasi banyak diimplementasikan di berbagai bidang. Collaborative Filtering menjadi salah satu paradigma sistem pemberi rekomendasi yang paling banyak digunakan karena mudah digunakan. Algoritma pengelompokan K-means banyak digunakan dalam Collaborative Filtering. Algoritma ini dapat memprediksi *rating* item yang akan diberikan oleh seorang *user*. *Rating* dapat diprediksi dengan menghitung rata-rata *rating* item tersebut. Performa pengelompokan algoritma ini tergolong rendah karena algoritma ini memilih *centroid* awal secara acak. Hal ini menyebabkan tingginya *error* pada prediksi rating item. Untuk menghasilkan *error* yang lebih rendah, penelitian ini menjalankan reduksi dimensi dengan Singular Value Decomposition (SVD). SVD dapat memfaktorisasi data hasil pengelompokan dan mereduksi dimensi data tersebut. Reduksi dimensi dengan SVD dapat dijalankan dengan menghapus karakteristik data yang tidak dominan. Penelitian ini menggunakan hasil faktorisasi yang sudah direduksi dimensinya untuk mengalkulasi *similarity* antar *cluster*. Nilai *similarity* antar *cluster* digunakan untuk memprediksi *rating* suatu item yang akan diberikan oleh suatu *cluster*. Hasil pengujian menunjukkan bahwa metode gabungan K-means dan SVD dapat menghasilkan RMSE sampai 8,936% lebih rendah dari metode K-means.

Kata Kunci: Sistem Pemberi Rekomendasi; K-Means; Singular Value Decomposition; Reduksi Dimensi

Abstract−Recommender system is widely implemented in various fields. Collaborative Filtering is one of the most used recommender system paradigms because it is easy to use. K-means clustering algorithm is widely use in Collaborative Filtering. This algorithm can predict the item rating that will be given by a user. Rating can be predicted by calculating the average rating of the item. The clustering performance of this algorithm is low because this algorithm selects initial centroid randomly. This causes high errors in the item rating prediction. To obtain lower error, we propose dimensionality reduction with Singular Value Decomposition (SVD). SVD is able to factorize the clustering result data and reduce dimensionality of the data. Dimensionality reduction with SVD can be carried out by removing non-dominant characteristics of the data. This study uses the result of factorization to calculate the similarity between clusters. The value of similarity between clusters is used to predict the rating of an item that will be given by a cluster. The experimental results show that the combined method of K-means and SVD can produces RMSE up to 8.936% lower than the K-means method.

Keywords: Recommender System; K-Means; Singular Value Decomposition; Dimensionality Reduction

1. INTRODUCTION

K-means clustering algorithm is widely used in Collaborative Filtering (CF). CF is a recommender system paradigm that provides recommendations of items that have been liked by other users [1]. The K-means clustering algorithm is able to cluster users or items. With K-means, the recommender system can predict the rating of an item that will be given by a user. K-means is included in Centroid Based Clustering algorithm. The clustering algorithm that included in Centroid Based Clustering clusters data points based on the distance of the data points to the nearest centroid. K-means selects the initial centroid randomly, this can cause the clustering performance to be low [2]. If the clustering performance is low, then the item rating prediction will have a large error. Therefore, this study carried out dimensionality reduction with Singular Value Decomposition (SVD) on the clustering result data (user clustering). Dimensionality reduction aims to produce a lower error. SVD is generally used to perform dimensionality reduction [3]. Dimensionality reduction in SVD is carried out by removing non-dominant data characteristics. This leaves only the dominant data characteristics. This can lead lower errors [4].

Mohamed, M. H. et al. [2] proposed a recommender system using K-means. From the experimental result, this algorithm produces a high error in rating predictions. Therefore, we used SVD in this study to produce lower errors. There are several related studies that also use SVD to produce lower errors. Zarzour, H. et al. [5] proposed dimensionality reduction with SVD to improve the accuracy of the recommender system. They cluster the users using the K-means algorithm. They use SVD to factorize the clustering result data. The factorization produces three matrices, e.g., U, Σ, and V^T. They calculate the similarity between users using cosine similarity by utilizing U, Σ, and V^T matrices. The value of 'similarity between users' is used to predict item ratings. From the experimental results, it is proven that the combined method of SVD and K-means can produce a lower error than the error of the K-means method. Barathy, R. et al. [6] overcomes the sparsity problem by predicting item rating. The prediction results are used to update the empty user-item matrix elements. They perform dimensionality reduction for the user-item matrix. The experimental results prove that this method can produce lower MAE and RMSE than just using SVD. Nilashi, M. et al. [7] proposed the SVD to predict empty ratings. They also use ontologies to produce lower errors. In the experimental results, it is proven that SVD can produce lower errors. Wang, N. J. et al. [8] used SVD and Trust Factor (TF) to obtain items' latent feature. They calculate the similarity between items

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 544−551
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2110

with items' latent features. In the experimental results, it is proven that SVD and TF can produce lower RMSE than the traditional CF method.

From related study, it is proven that SVD can reduce errors in rating predictions. We propose dimensionality reduction for the cluster-item matrix with SVD to produce lower errors in item's rating prediction. We implement this method to create a movie recommender system. The system is able predict item ratings with the value of 'similarity between clusters'. System calculates the 'similarity between clusters' based on the rows of the U matrix using cosine similarity. This study aims to implement and evaluate the proposed system. The results of this study are expected to help the industry (that have CF recommender system) to consider using dimensionality reduction so that they can improve their recommender system. While the contribution of this study to computer science is the proposed system can be used to improve existing recommender system that used K-means algorithm.

2. RESEARCH METHODOLOGY

2.1 Dataset

We used the MovieLens 1M dataset to build this system. This dataset consists of 1,000,209 ratings from 6,040 users on 3,706 movies. The value of rating range 1 to 5. We use the dataset from MovieLens because the datasets from MovieLens are widely used to evaluate CF. The dataset sample is listed in Table 1.

Table 1. Dataset sample

UserID	MovieID	Rating	Timestamp
1	1193	5	978300760
1	661	3	978302109
1	914	3	978301968
1	3408	4	978300275
1	2355	5	978824291

2.2 Research Stages

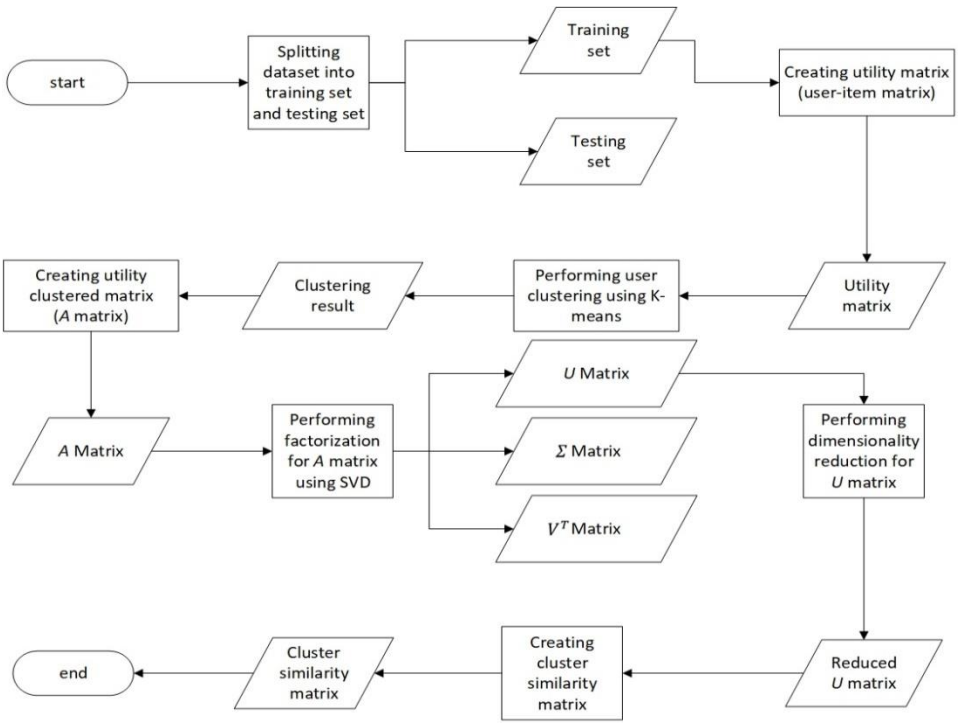

Figure 1. Cluster similarity matrix creation flow

We built a movie recommender system. The recommender system can predict the item rating that will be given by a cluster. To predict item rating, the system needs a cluster similarity matrix as shown below,

$$\begin{pmatrix} s_{1,1} & s_{1,2} & s_{1,3} & s_{1,4} & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & s_{2,3} & s_{2,4} & \cdots & s_{2,n} \\ s_{3,1} & s_{3,2} & s_{3,3} & s_{3,4} & \cdots & s_{3,n} \\ s_{4,1} & s_{4,2} & s_{4,3} & s_{4,4} & \cdots & s_{4,n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & s_{n,3} & s_{n,4} & \cdots & s_{n,n} \end{pmatrix}$$

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 544−551
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2110

both rows and columns represent clusters. The notation $s_{a,b}$ represents the similarity of a-th cluster and b-th cluster. Figure 1 shows the workflow of the system in creating a cluster similarity matrix. First, system split dataset into training set and testing set. Training set is used in the next stage, while testing set is used during testing. Then, the system creates utility matrix (user-item matrix) based on training set. Based on utility matrix, system performs user clustering using K-means clustering algorithm. Then, system create a utility clustered matrix (cluster-item matrix) using the clustering result. System factorizes the utility clustered matrix using SVD. This factorization produces three matrices, e.g., U, Σ, and V^T. Only matrix U is used in the next stage. Thus, system performing dimensionality reduction for U matrix. The rows of U matrix represents the cluster. So, system calculate the value of 'similarity between cluster' based on the rows of U matrix. System used the value of 'similarity between cluster' to create cluster similarity matrix.

2.2.1 Dividing dataset into training set and testing set

We split the dataset into two parts, e.g., training set and the testing set. The percentages of training set and testing set are 80% and 20% respectively. The splitting of the dataset is done randomly. The training set is used in the next stage, while the testing set is used during testing.

2.2.2 Creating utility matrix (user-item matrix)

The system creates a utility matrix as shown below,

$$\begin{pmatrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & r_{2,3} & r_{2,4} & \cdots & r_{2,n} \\ r_{3,1} & r_{3,2} & r_{3,3} & r_{3,4} & \cdots & r_{3,n} \\ r_{4,1} & r_{4,2} & r_{4,3} & r_{4,4} & \cdots & r_{4,n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{m,1} & r_{m,2} & r_{m,3} & r_{m,4} & \cdots & r_{m,n} \end{pmatrix}$$

rows represent users and columns represent items. The notation $r_{m,n}$ represents the rating of the n-th item given by the m-th user. The system creates a utility matrix of all user-item pairs in the training set. If an item does not have a rating from a user, its rating is considered as 0 [9].

2.2.3 Performing user clustering using K-means

The system clusters the users using the K-means algorithm based on the row of utility matrix (the row of utility matrix represent user).

2.2.4 Creating utility clustered matrix (A matrix)

The system creates a utility clustered matrix as shown below,

$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} & \cdots & a_{2,n} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} & \cdots & a_{3,n} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} & \cdots & a_{4,n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & a_{m,3} & a_{m,4} & \cdots & a_{m,n} \end{pmatrix}$$

rows represent clusters and columns represent items. The notation $a_{m,n}$ is the average rating of the n-th item given by users in the m-th cluster. This matrix is hereinafter referred as A matrix.

2.2.5 Performing factorization for A matrix using SVD

Matrix A is factorized using SVD. The factorization of A matrix produces three matrices, e.g., U, Σ, and V^T. The SVD theorem is found in equation (1).

$$A_{m,n} = U_{m\times r}\Sigma_{r\times r}{V^T}_{r\times n} \quad (1)$$

Figure 2 shows the SVD procedure for factorize A matrix. The c_n notation represents the n-th cluster. So, the rows in the U matrix represent clusters. The U matrix is also known as the left singular vector. The notation f_r represents the r-th eigenvector of the matrix $A \times A^T$. The i_p notation represents the p-th item. So, the columns in the V^T matrix represent items. V^T matrix is also known as right singular vector. The notation g_r represents the r-th eigenvector of the matrix $A^T \times A$ [10]. The Σ matrix is a diagonal matrix containing the square root of the eigenvalues of the $A \times A^T$ matrix or the $A^T \times A$ matrix in descending order. The square root of the eigenvalues of the $A \times A^T$ matrix or the $A^T \times A$ matrix is called as rank. Rank is denoted by σ. The number of ranks in the Σ matrix is denoted by r [11].

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 544−551
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2110

2.2.6 Performing dimensionality reduction for U matrix

Dimensionality reduction with SVD can be carried out by considering only (100- t)% of the first column of the U matrix, (100- t)% of the first rank, and (100- t)% of the first row of the V^T matrix [7]. The notation t represents the percentage of dimensionality reduction. In the next stage, only the U matrix is used in calculating the similarity between clusters. So, the system only considers the (100- t)% of the first column in the U matrix. The system removes the last t% of the column in the U matrix.

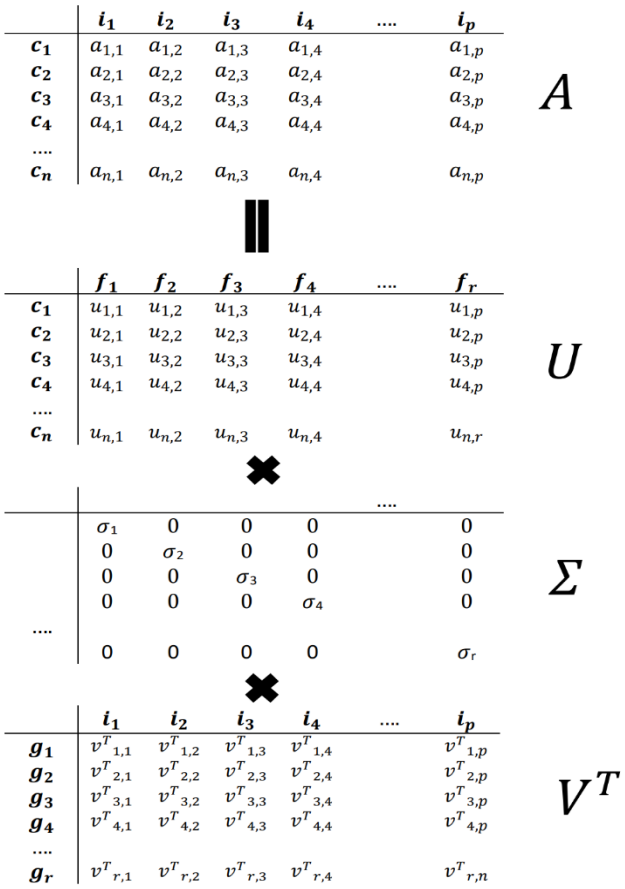

Figure 2. SVD Procedure in factorize matrix

2.2.7 Creating cluster similarity matrix

The system calculates similarity between clusters using equation (2). The x and y notations represent the x-th cluster vector and the y-th cluster vector respectively. The x-th and y-th cluster vectors are obtained from the x-th and y-th rows of U matrix respectively. The notations $\|x\|$ and $\|y\|$ represent the length of the x and y vectors respectively. The similarity results are used to create a cluster similarity matrix.

$$sim(x, y) = \frac{x \cdot y}{\|x\| \, \|y\|} \tag{2}$$

2.3 Rating Prediction

After cluster similarity is formed, the system can predict the rating of an item that will be given by a cluster by using equation (3) [12]. The notation $\hat{r}_{ci}$ represents the predicted rating of item i from cluster c, the notation $\bar{r}_c$ represents the average rating of cluster c, the notation N represents the set of neighbours of cluster c, and the notation r_{ni} represents the rating item i from cluster n.

$$\hat{r}_{ci} = \bar{r}_c + \frac{\sum_{n \in N} s_{c,n} |r_{ni} - \bar{r}_n|}{\sum_{n \in N} s_{c,n}} \tag{3}$$

3. RESULTS AND DISCUSSION

3.1 Training Set and Testing Set

Table 2 listed the sample of training set. Training set consists of 800,167 ratings from 6,040 users on 3,670 movies. Table 3 listed the sample of testing set. Testing set consists of 200,042 ratings from 6,038 users on 3,464 movies.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 544−551
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2110

There are users on training set but not on testing set. And there are items on training set but not on testing set. We remove those users and items from testing set.

Table 2. Training set sample

User id	Movie id	Rating
1676	2746	2
4834	2702	3
481	2944	4
1164	535	3
2909	1957	5

Table 3. Testing set sample

User id	Movie id	Rating
1	595	5
1	2018	4
1	527	5
1	2340	3
1	1097	4

3.2 Utility Matrix

The matrix below is the first six rows and first six columns of utility matrix,

$$\begin{pmatrix} 5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

rows represent user and column represent item. This matrix is very sparse. The size of the utility matrix is 6,040 × 3,670 because the training set consist of 6,040 users and 3,670 movies.

3.3 Clustering Result

The system performs user clustering based on the rows of utility matrix (because the rows of utility matrix represent user). Table 4 listed clustering result sample for k value of 10.

Table 4. Clustering result sample

User id	Cluster
1	4
2	6
3	4
4	4
5	9

3.4 Utility Clustered Matrix

The matrix below is the first six rows and first six columns of utility clustered matrix,

$$\begin{pmatrix} 4.329 & 3.000 & 2.581 & 2.714 & 2.583 & 3.686 \\ 4.128 & 3.405 & 2.750 & 1.000 & 2.625 & 3.863 \\ 4.105 & 2.932 & 2.958 & 2.500 & 2.727 & 3.793 \\ 3.994 & 3.348 & 3.203 & 2.692 & 3.071 & 4.120 \\ 4.043 & 2.965 & 2.607 & 2.615 & 2.649 & 3.891 \\ 4.074 & 3.345 & 3.054 & 2.667 & 3.242 & 3.928 \end{pmatrix}$$

rows represent cluster and column represent item. This matrix is very sparse. The size of the utility matrix is 10 × 3,670 because there are 10 clusters.

3.5 U Matrix

Only matrix U is used to calculate similarity between cluster. The matrix below is the U matrix produced by SVD factorization,

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 544−551
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2110

$$\begin{pmatrix} -0{,}337 & 0{,}242 & -0{,}022 & 0{,}043 & -0{,}033 & 0{,}815 & -0{,}283 & -0{,}112 & -0{,}291 & 0.801 \\ -0{,}302 & -0{,}609 & 0{,}484 & -0{,}514 & 0{,}082 & 0{,}147 & -0{,}082 & 0{,}026 & -0{,}045 & 0{,}047 \\ -0{,}317 & 0{,}228 & 0{,}302 & 0{,}245 & -0{,}033 & 0{,}149 & -0{,}552 & -0{,}304 & 0{,}428 & -0{,}305 \\ -0{,}334 & 0{,}113 & 0{,}074 & 0{,}151 & -0{,}398 & 0{,}613 & 0{,}556 & 0{,}025 & 0{,}004 & -0{,}057 \\ -0{,}317 & 0{,}240 & -0{,}183 & -0{,}276 & 0{,}091 & -0{,}071 & -0{,}024 & -0{,}317 & -0{,}633 & -0{,}466 \\ -0{,}293 & -0{,}354 & -0{,}547 & -0{,}093 & -0{,}585 & -0{,}259 & -0{,}137 & -0{,}088 & 0{,}211 & -0{,}008 \\ -0{,}317 & -0{,}153 & -0{,}254 & 0{,}300 & 0{,}263 & 0{,}183 & -0{,}268 & 0{,}703 & -0{,}157 & -0{,}169 \\ -0{,}304 & -0{,}371 & 0{,}027 & 0{,}554 & 0{,}357 & -0{,}287 & 0{,}334 & -0{,}368 & -0{,}025 & 0{,}058 \\ -0{,}323 & 0{,}263 & -0{,}313 & -0{,}411 & 0{,}480 & 0{,}007 & 0{,}246 & 0{,}023 & 0{,}509 & 0{,}094 \\ -0{,}319 & 0{,}304 & 0{,}415 & -0{,}010 & -0{,}235 & -0{,}621 & 0{,}190 & 0{,}394 & 0{,}014 & -0{,}032 \end{pmatrix}$$

rows represent cluster and columns represent eigenvectors of $A \times A^T$ matrix. The size of the utility matrix is 10 × 10 because there are 10 clusters.

3.6 Reduced U matrix

The matrix below is the reduced U matrix with t value of 10,

$$\begin{pmatrix} -0{,}337 & 0{,}242 & -0{,}022 & 0{,}043 & -0{,}033 & 0{,}815 & -0{,}283 & -0{,}112 & -0{,}291 \\ -0{,}302 & -0{,}609 & 0{,}484 & -0{,}514 & 0{,}082 & 0{,}147 & -0{,}082 & 0{,}026 & -0{,}045 \\ -0{,}317 & 0{,}228 & 0{,}302 & 0{,}245 & -0{,}033 & 0{,}149 & -0{,}552 & -0{,}304 & 0{,}428 \\ -0{,}334 & 0{,}113 & 0{,}074 & 0{,}151 & -0{,}398 & 0{,}613 & 0{,}556 & 0{,}025 & 0{,}004 \\ -0{,}317 & 0{,}240 & -0{,}183 & -0{,}276 & 0{,}091 & -0{,}071 & -0{,}024 & -0{,}317 & -0{,}633 \\ -0{,}293 & -0{,}354 & -0{,}547 & -0{,}093 & -0{,}585 & -0{,}259 & -0{,}137 & -0{,}088 & 0{,}211 \\ -0{,}317 & -0{,}153 & -0{,}254 & 0{,}300 & 0{,}263 & 0{,}183 & -0{,}268 & 0{,}703 & -0{,}157 \\ -0{,}304 & -0{,}371 & 0{,}027 & 0{,}554 & 0{,}357 & -0{,}287 & 0{,}334 & -0{,}368 & -0{,}025 \\ -0{,}323 & 0{,}263 & -0{,}313 & -0{,}411 & 0{,}480 & 0{,}007 & 0{,}246 & 0{,}023 & 0{,}509 \\ -0{,}319 & 0{,}304 & 0{,}415 & -0{,}010 & -0{,}235 & -0{,}621 & 0{,}190 & 0{,}394 & 0{,}014 \end{pmatrix}$$

only nine columns left. The system removes last one column of U matrix.

3.7 Cluster Similarity Matrix

The matrix below is the cluster similarity matrix,

$$\begin{pmatrix} 1 & -0{,}058 & 0{,}412 & 0{,}059 & 0{,}715 & 0{,}010 & 0{,}216 & -0{,}075 & -0{,}138 & 0{,}033 \\ -0{,}058 & 1 & 0{,}014 & 0{,}002 & 0{,}025 & 0{,}001 & 0{,}008 & -0{,}003 & -0{,}005 & 0{,}011 \\ 0{,}412 & 0{,}014 & 1 & -0{,}015 & -0{,}178 & -0{,}002 & -0{,}054 & 0{,}019 & 0{,}034 & -0{,}008 \\ 0{,}059 & 0{,}002 & -0{,}014 & 1 & -0{,}025 & -0{,}001 & -0{,}008 & 0{,}003 & 0{,}005 & -0{,}001 \\ 0{,}715 & 0{,}025 & -0{,}018 & -0{,}026 & 1 & -0{,}004 & -0{,}094 & 0{,}032 & 0{,}060 & -0{,}014 \\ 0{,}010 & 0{,}001 & -0{,}002 & -0{,}001 & -0{,}004 & 1 & -0{,}001 & 0{,}001 & 0{,}001 & -0{,}001 \\ 0{,}216 & 0{,}008 & -0{,}054 & -0{,}008 & -0{,}093 & -0{,}001 & 1 & 0{,}010 & 0{,}018 & -0{,}004 \\ -0{,}075 & 0{,}002 & -0{,}019 & 0{,}003 & 0{,}032 & 0{,}004 & 0{,}010 & 1 & -0{,}006 & 0{,}001 \\ -0{,}138 & -0{,}004 & 0{,}034 & 0{,}005 & 0{,}060 & 0{,}001 & 0{,}018 & -0{,}006 & 1 & 0{,}002 \\ 0{,}033 & -0{,}001 & -0{,}008 & -0{,}001 & -0{,}014 & -0{,}001 & -0{,}004 & 0{,}001 & 0{,}003 & 1 \end{pmatrix}$$

both row and column represent cluster. All value in the diagonal of cluster similarity matrix is 1 because those is the similarity of the same vectors.

3.8 Implementation

System evaluation was carried out using the RMSE method. RMSE is commonly used to evaluate recommender systems, especially those using CF. RMSE is used to measure the closeness between the actual rating and the rating predicted by the system (predicted rating). The smaller the RMSE value, the better the system built [13]. Let r_{ci} is the actual rating of cluster c to item i, $\hat{r}_{ci}$ is the predicted rating of item i to cluster c, and n is the number of rating-prediction pairs, RMSE can be calculated using equation (4) [14]. The actual rating is obtained from the testing set, while the predicted rating is calculated by the system using equation (3) [15].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(r_{ci}-\hat{r}_{ci})^2}{n}} \quad (4)$$

The neighbours of a cluster that used to calculate the predicted rating are all other clusters in the cluster similarity matrix. Clustering with K-means was carried out five times with k values of 10, 20, 30, 40, and 50 respectively. Dimensionality reduction with SVD was carried out three times with t values of 10 (*K-means-svd-10*), 20 (*K-means-svd-20*), and 30 (*K-means-svd-30*) respectively.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 544−551
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2110

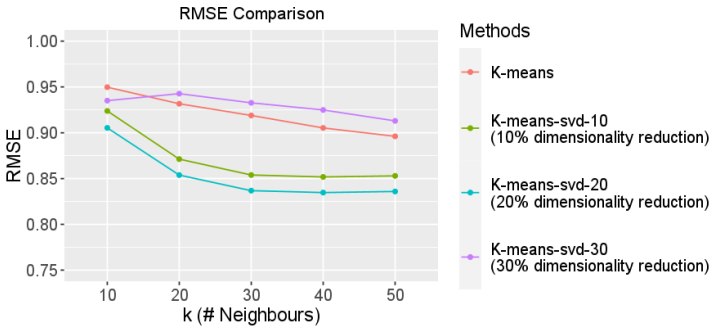

Figure 3. RMSE comparison between K-means method, *K-means-svd-10* method, *K-means-svd-20* method, and *K-means-svd-30* method

Figure 3 shows the RMSE comparison of the K-means, K-means-svd-10, K-means-svd-20, and K-means-svd-30 methods. For all k values, the K-means-svd-20 method produces lowest RMSE value than the other methods. So, the K-means-svd-20 method is the most optimal method in this case. The K-means-svd-10 method produces a lower RMSE value than the K-means method. The K-means-svd-30 method produces a lower RMSE value than the K-means method only when the value of $k = 10$. So, the K-means-svd-30 method is not suitable to be applied to the value of $k > 10$. Table 5 shows the RMSE percentage difference from the K-means method and the *K-means-svd-20* method. Testing of the *K-means-svd-20* method can produce RMSE 4.677% to 8.936% lower than the K-means method.

Table 5. Percentage of RMSE difference from K-means and *K-means-svd-20* methods

k	**RMSE**		**RMSE percentage difference**
	K-means	***K-means-svd-20***	
10	0.950	0.905	4.677%
20	0.932	0.854	8.361%
30	0.919	0.837	8.936%
40	0.905	0.835	7.795%
50	0.896	0.836	6.722%

4. CONCLUSION

This study evaluated a method for CF. This method not only focuses on providing recommendations based on user clusters, but also focuses on producing lower errors in item rating predictions. To cluster users, this method uses the K-means algorithm. Meanwhile, to get a lower error in item rating predictions, this method performs dimensionality reduction with SVD. Dimensionality reduction with SVD can be carried out by removing some non-dominant data characteristics, so that only the more dominant data characteristics remain. With dimensionality reduction, the error in item rating predictions can be lower. Based on the experimental results, it is proven that this method can produce a lower RMSE item rating prediction than the method that only uses K-means. This method can produce RMSE up to 8.936% lower than the method that only uses K-means. However, the experiment was carried out with a relatively small dataset when compared to the dataset used in the related research. It is hoped that in the future this method can be tested with larger datasets. It is also expected that the calculation of cosine similarity is replaced with adjusted cosine similarity so that the system can find out the scheme of a cluster when rate an item.

REFERENCES

[1] 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 and Knowledge Engineering*, vol. 128, 2020, doi: 10.1016/j.datak.2020.101813.

[2] M. H. Mohamed, M. H. Khafagy, and M. H. Ibrahim, "Two recommendation system algorithms used SVD and association rule on implicit and explicit data sets," *International Journal of Scientific and Technology Research*, vol. 9, no. 1, 2020.

[3] N. Salem and S. Hussein, "Data dimensional reduction and principal components analysis," in *Procedia Computer Science*, 2019, vol. 163. doi: 10.1016/j.procs.2019.12.111.

[4] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, 2020, doi: 10.38094/jastt1224.

[5] H. Zarzour, Z. Al-Sharif, M. Al-Ayyoub, and Y. Jararweh, "A new collaborative filtering recommendation algorithm based on dimensionality reduction and clustering techniques," in *2018 9th International Conference on Information and Communication Systems, ICICS 2018*, 2018, vol. 2018-January. doi: 10.1109/IACS.2018.8355449.

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 544−551
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2110

[6] R. Barathy and P. Chitra, "Applying Matrix Factorization in Collaborative Filtering Recommender Systems," 2020. doi: 10.1109/ICACCS48705.2020.9074227.
[7] M. Nilashi, O. Ibrahim, and K. Bagherifard, "A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques," *Expert Systems with Applications*, vol. 92, 2018, doi: 10.1016/j.eswa.2017.09.058.
[8] J. Wang, P. Han, Y. Miao, and F. Zhang, "A Collaborative Filtering Algorithm Based on SVD and Trust Factor," 2019. doi: 10.2991/cnci-19.2019.5.
[9] R. Ahuja, A. Solanki, and A. Nayyar, "Movie recommender system using k-means clustering and k-nearest neighbor," 2019. doi: 10.1109/CONFLUENCE.2019.8776969.
[10] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Computer Science Review*, vol. 40. 2021. doi: 10.1016/j.cosrev.2021.100378.
[11] S. M. Atif, S. Qazi, and N. Gillis, "Improved SVD-based initialization for nonnegative matrix factorization using low-rank correction," *Pattern Recognition Letters*, vol. 122, 2019, doi: 10.1016/j.patrec.2019.02.018.
[12] G. Geetha, M. Safa, C. Fancy, and D. Saranya, "A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System," in *Journal of Physics: Conference Series*, 2018, vol. 1000, no. 1. doi: 10.1088/1742-6596/1000/1/012101.
[13] A. Davoudi and M. Chatterjee, "Social trust model for rating prediction in recommender systems: Effects of similarity, centrality, and social ties," *Online Social Networks and Media*, vol. 7, 2018, doi: 10.1016/j.osnem.2018.05.001.
[14] J. Zhang, Y. Wang, Z. Yuan, and Q. Jin, "Personalized real-time movie recommendation system: Practical prototype and evaluation," *Tsinghua Science and Technology*, vol. 25, no. 2, 2020, doi: 10.26599/TST.2018.9010118.
[15] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 Relation Between Training and Testing Sets : A Pedagogical Explanation," *Departmental Technical Reports (CS)*, 2018.