# An Improved Collaborative Filtering Algorithm to Improve Recommendation Accuracy and Protect User Privacy

Fang Yin, Yao Song, Ao Li
*School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China*
Email: 13936421412@163.com, 546841677@qq.com, dargonboy@126.com

***Abstract*—The emergence of recommendation algorithm alleviates the problem of information overload to a certain extent, makes recommendations to users on the premise of protecting user privacy, and helps users to explore potential demands. An improved collaborative filtering algorithm is proposed for solving the problem that the relationship between recommendation time as well as accuracy cannot be weighed and how to protect user privacy. In this paper, the matrix dimension reduction technique of locally optimized singular value decomposition(SVD) and the K-means clustering technique are used to reduce the dimensions and cluster similar users in the user-item scoring matrix. The approximate difference matrix is used to represent the local structure of the scoring matrix to implement the local optimization. The locally optimized SVD technique can alleviate the problem of data sparsity and poor scalability in collaborative filtering by using fewer iterations. K-means clustering technique can greatly narrow the search range of neighbor sets and improve the recommendation speed. Therefore, collaborative filtering algorithms based on dimensionality reduction and clustering can generate recommendations in an accurate and real-time manner. The experimental results on the MovieLens dataset show that the algorithm can reduce the impact of data sparsity, protect user privacy and effectively improve the accuracy of recommendation.**

***Index Terms*—collaborative filtering algorithm, cluster, dimension reduction, local optimization, user privacy**

## I. Introduction

Human recommendation is not only costly but also sometimes violate others' privacy . For most companies, it is increasingly important to understand the needs and preferences of online users or customers. However, online users or customers often face the problem of critical information overload. The recommendation system can effectively alleviate information overload.

It personalized guides users to find attractive or satisfying objects in many possible items. The recommendation system fully improves customer satisfaction by effectively mapping customer needs with the best products. At present, recommendation system is facing serious privacy protection and security problems. The recommendation system needs to collect a large amount of user information, user behavior, etc. The richer the data, the higher the recommendation accuracy may be. However, these information may reveal users' personal privacy. For privacy and information security reasons, users may be reluctant to allow these data to be recorded and stored by recommendation systems. Since the recommendation system requires a large amount of user data for collaborative filtering, The privacy protection of data has become an urgent problem to be solved in the field of recommendation systems. However, the recommendation system depends on the recommendation algorithm [1].

Traditional recommendation algorithms are mainly divided into three categories: content-based recommendation algorithm, collaborative filtering (CF) recommendation algorithm and hybrid recommendation algorithm [2-3]. Content-based recommendation algorithms focus on user preference files and project descriptions in order to recommend items that are most similar to those with high scores in the past. The CF recommendation algorithm mainly includes memory-based CF algorithm and model-based CF algorithm. Among them, the memory-based CF algorithm includes user-based CF algorithm and item-based CF algorithm [4]. The CF recommendation algorithm mainly collects feedback from different users on items, and provides recommendation to the target users through similarities between items and items (item-based) or between users and users (user-based). The hybrid recommendation algorithm combines content-based recommendation methods with collaborative filtering recommendations to leverage their strengths to recommend more user-friendly items. Compared with the CF recommendation algorithm, the content-based recommendation algorithm does not need to obtain historical data of the user, and there is no data sparsity and cold start problem. However, the scalability and recommendation diversity of the algorithm is poor, and because of the constraint of the feature extraction method, the content-based recommendation method is not suitable for the recommendation of multimedia data. The CF recommendation algorithm has rich recommendation results and is easy to find user's interests and preferences. However, the CF recommendation algorithm has serious data sparsity and cold start problems. Hybrid recommendation algorithm improves the sparsity and cold start problem to some extent, but its framework is complex and difficult to implement.

In recent years, many studies have been carried out to improve the accuracy of recommendation, data sparsity and cold start. Many researchers use clustering technology or dimensionality reduction technology to improve the shortcomings of traditional collaborative filtering algorithm. Kim et al. [5] proposed a typical model, which integrates collaborative filtering, clustering and social network analysis techniques to improve the accuracy of recommendation system. Their model uses SNA to identify the most influential people on social networks, and then uses these people for clustering analysis. However, as the number of users increases, the model may encounter scalability problems. Koohi et al. [6] proposed a fuzzy C-means method based on user collaborative filtering, and compared its performance under different clustering methods. Although the recommendation accuracy has been improved, the problem of sparse matrix has not been improved. Le et al. [7] proposed a new matrix decomposition method, called bounded SVD, which utilizes the constraint that all evaluations in the evaluation matrix are bounded within a predetermined range. However, the learning rate is adjusted manually and only by simple rules. Zhang et al. [8]

978-1-7281-1973-1/19/$31.00 ©2019 IEEE

proposed a collaborative filtering algorithm based on user preference clustering. Considering user preferences from local and global perspectives, the similarity measurement method is improved. Lee et al. [9] proposed an extensible clustering-based CF method to provide balance by repositioning the elements in the clustering model. It can improve the system performance under the cold start problem, and also solve the first user/item problem. But it can not improve the performance of collaborative filtering system without the influence of cold start problem. Compared with the memory-based CF method, which uses the overall similarity between users or items to predict, the model-based CF method only uses a set of score data to train the model, and then predicts the user's score on the non-scored items. Memory-based CF method is easier to understand and implement. However, due to the user-item matrix is sparse, the accuracy of this method in large-scale applications begins to decline.

In order to overcome the shortcomings of traditional CF recommendation methods such as matrix sparsity and poor scalability, a collaborative filtering recommendation algorithm based on dimensionality reduction of local optimization and clustering is proposed. The purpose is to improve the performance and speed of recommendation system, and to minimize the impact of sparse user-item matrix on the accuracy of recommendation. K-means clustering technology and SVD dimension reduction technology of local optimization are used to cluster similar users in user-item matrix and reduce dimensions [10].

The structure of this paper is as follows: Chapter 2 elaborates the theoretical basis; Chapter 3 describes the proposed algorithm in detail; Chapter 4 verifies the proposed algorithm through experiments; Chapter 5 summarizes the full text.

## II. Related Theory

### A. SVD Dimension Reduction Technology of Local Optimization

SVD is one of the most important matrix decomposition methods in linear algebra, which is usually used to reduce the number of data features[11]. Sarwar et al. applied SVD technology to collaborative filtering recommendation algorithm for the first time. The recommendation system based on SVD can reduce the sparsity of user-item score matrix and improve the accuracy of recommendation as well as protect user privacy. Generally, there are two methods to solve the SVD-based recommendation algorithm: random gradient descent method and alternating least squares method.

Compared with the latter, the gradient descent method has more advantages in computation speed and implementation. However, in the process of solving the problem, the speed of error reduction decreases gradually and a large number of iterations require more training time. In this paper, the approximate difference matrix is used to represent the local structure of the scoring matrix to achieve the effect of local optimization. Because the user-item matrix is too sparse, the approximate difference matrix can be constructed by using the same user's scoring difference for the adjacent items. But dimensionality reduction will lose some information of user rating matrix to a certain extent. This paper chooses the appropriate dimension s to reduce the error caused by dimension reduction.

The locally optimized SVD theorem shows that for all matrices $C[k,n]$, where $k$ rows represent users and $n$ columns represent items, $C$ can be decomposed as follows:

$$C = U \cdot \sum \cdot V^T \tag{1}$$

Among them, $U$ is the standard orthogonal matrix of size $k \times r$, $\sum$ is the diagonal matrix of size $r \times r$ and has the singular value of $C$, $V^T$ is the standard orthogonal matrix of size $r \times n$.

Because the reduction of singular value is very fast, this paper approximates the matrix with the singular value of the first $s$, and obtains the lower order approximation of $C$ as follows.

$$C_s = U_s \cdot \sum_s \cdot V_s^{\ T} \tag{2}$$

$$X = U_s \cdot \sum_s^{\frac{1}{2}}, \quad Y = V_s \cdot \sum_s^{\frac{1}{2}} \tag{3}$$

$X, Y$ are user feature matrix and item feature matrix, Dimensions are $s$, $C_s = X \cdot Y^T$ .The difference matrix $D$ is used to represent the local information of the user-item scoring matrix.

$$D = C \cdot S^T \tag{4}$$

$$D = \begin{pmatrix} C_{11} - C_{12} & \cdots & C_{1(\mathrm{n}-1)} C_{1n} \\ \vdots & \ddots & \vdots \\ C_{k1} - C_{k2} & \cdots & C_{k(\mathrm{n}-1)} - C_{kn} \end{pmatrix}$$

$$C = \begin{pmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & \ddots & \vdots \\ C_{k1} & \cdots & C_{kn} \end{pmatrix} \quad S = \begin{pmatrix} 1 & -1 & 0 & \\ 0 & 1 & -1 & \cdots \\ 0 & 0 & 1 & \end{pmatrix}$$

$D_{ij} = C_{ij} - C_{i(\mathrm{j}+1)}$ indicates the score difference of user $i$ for item $j$ and $j+1$, which also reflects the popularity of the project.

It is not easy to get the difference matrix from the sparse matrix graded by user. Although there is a general problem of scoring standard among users, it can be approximated that the vast majority of users can distinguish good items or not. This paper considers that different users have the same score difference on the same item, and then get the approximate difference matrix $\hat{D}$ .

$$\widehat{D_{ij}} = \left|M_j\right|^{-1} \sum\nolimits_{p \in M_j} (\mathrm{C}_{pj} - \mathrm{C}_{p(\mathrm{j}+1)}) \tag{5}$$

Among them, $M_j$ represents a set of users scoring item $j$ and item $j+1$ at the same time.

### B. K-means Clustering Technology

With the wide application of recommendation system, the number of users and items increases exponentially, so the computation of the algorithm is also increasing, and the search time of neighborhood set will be longer. This algorithm uses clustering technology to reduce the searching time of user's neighborhood set, and then improves the speed of recommendation and the real-time performance of recommendation. K-means method is one of the most popular clustering algorithms. It was proposed by Macqueen in 1967. It has been widely used in data mining and most recommendation system industries [12].

K-means algorithm has the advantages of high efficiency, easy implementation and well scalability. The basic idea is to divide the data set into $k$ clusters by iteration, so that the similarity of the data in the same category is greater, and the similarity of the data in different categories is smaller. However, the number of clustering $k$ needs to be determined in advance. In this paper, the upper bound of the optimal solution $k \le \sqrt{n}$ is calculated to narrow the scope of the optimal solution, and then the value of $k$ is determined by a further experiment. Firstly, $k$ users are randomly selected as $k$ centers. Secondly, the remaining users are allocated to the nearest cluster according to their distance to each center. Pearson similarity is used to calculate distance. The similarity $sim(\mathrm{i,j})$ between user $i$ and user $j$ are as follows:

$$sim(\mathrm{i,j}) = \frac{\sum_{p\in I_{ij}} (\mathrm{C}_{ip} - \overline{C_i})(\mathrm{C}_{jp} - \overline{C_j})}{\sqrt{\sum_{p\in I_{ij}} (\mathrm{C}_{ip} - \overline{C_i})^2 (\mathrm{C}_{jp} - \overline{C_j})^2}} \qquad (6)$$

Among them, $I_{ij}$ is the set of items scored jointly by $i$ and $j$ .Besides, $C_{ip}$ indicates the score of user $i$ on item $p$ , $\overline{C_i}$ and $\overline{C_j}$ respectively represent the average score of user $i$ and $j$ on common scoring items.

The third step is to calculate the mean of cluster to define a new center. The fourth step is to recalculate the distance for each user to update the cluster to which the user belongs. Finally, the user is redistributed according to the distance between the user and the center until the termination condition is satisfied.

### *C. The Collaborative Filtering Algorithm Based on Locally Optimized SVD and K-means*

For protecting user privacy, Collaborative filtering algorithm based on K-means improves the disadvantage that real-time recommendation of the traditional memory-based collaborative filtering algorithm is poor. But when the data is very sparse, the recommendation accuracy of K-means-based collaborative filtering algorithm is low, while the locally optimized SVD dimensionality reduction technology has a better effect in dealing with high-dimensional sparse data. Therefore, this paper proposes a collaborative filtering recommendation algorithm based on local optimization SVD dimensionality reduction technology and K-means clustering technology.

*Step 1:* reduce the dimension of the sparse user-item scoring matrix to get the user feature matrix.

Locally optimized SVD algorithm uses the gradient descent method to iterate. The termination condition of iteration can be set as the maximum number of iterations or the difference between the two mean square errors be less than a certain threshold. The algorithm first iterates at learning rate $sr_1$ ,when the difference of the two mean square error (MSE) is less than threshold $\beta$ , the algorithm iterates at a smaller learning rate $sr_2$ . The steps of locally optimized of SVD algorithm are as follows:

*1) Initialization:* $PMSE = 0$ , $Sum = 0$ , $sr_1 = 0.003$ , $sr_2 = 0.00005$ , $\lambda = 0.12$ , $\beta = 0.0003$ ;

*2) For user item set* $(\mathrm{i,j})$ *in training set D :*

*a) Calculate user* $i$ *'s score on item* $j$ *:* $\widehat{C_{ij}} = X_i Y_j^T$ ;

*b) Calculating the error between the predictive rating and the real:* $r_{ij} = C_{ij} - \widehat{C_{ij}}$ ; $Sum{+}= r_{ij} \cdot r_{ij}$ ;

*c) For all features* $f (1 \le f \le s)$ *, solution by gradient descent method:* Delete the author and affiliation lines for the extra authors.

$$X_{if} = X_{if} - sr_1(\mathrm{r}_{ij} \cdot \mathrm{X}_{if} + \lambda \mathrm{Y}_{jf}) \qquad (7)$$

$$Y_{jf} = Y_{jf} - sr_1(\mathrm{r}_{ij} \cdot \mathrm{Y}_{jf} + \lambda \mathrm{X}_{if}) \qquad (8)$$

$sr_1$ is the learning rate and $\lambda$ is the regularization parameter;

*3) if* $\left| PMSE - \frac{Sum}{|D|} \right| < \beta$ ,go to step 4;

*else* $PMSE = \frac{Sum}{|D|}$ , $Sum = 0$ , iterative step 2;

*4) For user-item set* $(\mathrm{i,j+1})$ *:*

*a) Calculate the error of the approximate difference matrix and the difference matrix:*

$$r_{ij} = D_{ij} - \widehat{D_{ij}} = (X_i Y_j^T - C_{i(\mathrm{j+1})}) - \widehat{D_{ij}} \; ;$$

$$Sum{+}= r_{ij} \cdot r_{ij} \; ;$$

*b) For all features* $f (1 \le f \le s)$ *, solution by gradient descent method:*

$$X_{if} = X_{if} - sr_2 \cdot \mathrm{r}_{ij} \cdot \mathrm{X}_{if} \qquad (9)$$

$$Y_{jf} = Y_{jf} - sr_2 \cdot \mathrm{r}_{ij} \cdot Y_{jf} \qquad (10)$$

Among them, $sr_2$ is the learning rate and $sr_2 << sr_1$ ;

*5) If the termination condition is satisfied, the iteration will be terminated, otherwise step 4 will be iterated.*

*Step 2:* apply clustering technology to the user feature matrix to get the clustering of similar users.

This paper chooses K-means clustering technology to cluster users. K-means algorithm can ensure that users of the same class have the same preferences.

*Step 3:* Predict the rating of the target user on the user test set.

Firstly, the nearest clustering center is calculated according to formula (5) Pearson similarity, and its category is determined, that is, the nearest neighbor set. The target user is then predicted to score on the unrated item.

There are three classical predictive formulas, Assume that the user set is $U = \{u_1, u_2, \cdots u_m\}$ ,the item set is $I = \{i_1, \mathrm{i}_2, \cdots i_n\}$ , $P_{u,i}$ is the predictive rating of user $u$ for item $i$ , and based on the score of neighbor users for item $i$ . Assuming that $N$ is the neighbor set of user $u$, the following formula can be used to predict $P_{u,i}$ .

$$P_{u,i} = \frac{1}{n}\sum_{c\in N} P_{c,i} \tag{11}$$

Formula (11) takes the average score of target user's neighbor set as the predictive rating. Although the calculation is simple, it does not consider the influence of different users in the neighborhood concentration on the target users.

$$P_{u,i} = k\sum_{c\in N} sim(\mathrm{u,c})P_{c,i}\ ,\ k = \frac{1}{\sum_{c\in N}\left|sim(\mathrm{u,c})\right|} \tag{12}$$

$sim(\mathrm{u,c})$ represents the similarity between user $u$ and user $c$, and $k$ is the standardization factor. Because the higher the similarity, the greater the influence of the neighbor users on the target users. Formula (12) on the basis of Formula (11) takes into account the influence of users of different neighborhood sets on target users. However, formula (12) does not take into account the different user scoring standard.

$$P_{u,i} = \overline{P_u} + \frac{\sum_{c\in N} sim(\mathrm{u,c})(P_{c,i} - \overline{P_c})}{\sum_{c\in N}\left|sim(\mathrm{u,c})\right|} \tag{13}$$

$\overline{P_u}$ is the average score of users, that is, the rating habits of user $u$. Formula (13) solves the problem of scoring standard on the basis of Formula (12), and the accuracy of prediction is higher. Therefore, this paper chooses formula (13) as the prediction formula.

*Step 4:* Select the $top-N$ item with the highest score according to the prediction result to recommend.

The flow chart of the algorithm is as follows:

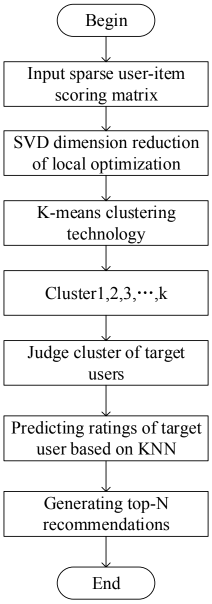
Begin
Input sparse user-item scoring matrix
SVD dimension reduction of local optimization
K-means clustering technology
Cluster1,2,3,⋯,k
Judge cluster of target users
Predicting ratings of target user based on KNN
Generating top-N recommendations
End

Fig. 1. The Algorithm Flow Chart in This Paper

## III. Experimental Results and Analysis

### A. Experimental Data Set

In order to evaluate the effectiveness of the method, the experiment used the MovieLens 100K dataset collected by the GroupLens research project at the University of Minnesota [13]. Among them, It includes 100,000 ratings of 1,682 online movies by 943 users, and each user rated at least 20 movies. The sparseness of the data set is approximately 93.7%. The rating value is an integer between 1 and 5, the size of which indicates how much the user likes the movie. In the experiment, the data set was randomly divided into a training set and a test set with a ratio of 4:1, and the final result was averaged.

### B. Algorithmic evaluation criteria

Mean Absolute Error (MAE) and Precision (Precision) were respectively used as prediction accuracy and classification error measurement[14]. MAE is a common measurement standard in recommendation system. It evaluates the algorithm by calculating the difference between the predicted and actual scores of the target users. The smaller the MAE is, the closer the predicted score is to the user's actual score will be, the better the performance of the algorithm. MAE is defined as follows:

$$MAE = \frac{1}{S}\sum_{i=1}^{S}\left|P_{u,i} - r_{u,i}\right| \tag{14}$$

Among them, $P_{u,i}$ is the predicted score of user $u$ for item $i$, $r_{u,i}$ is the actual score and $S$ is the total score.

Precision indicators indicate the proportion of $Top-N$ recommendations that are accurate. Precision is defined as follows:

$$\Pr ecision = \frac{Test \cap Top-N}{N} \tag{15}$$

Among them, $Test$ is the number of items in the test set; $Top-N$ is the number of items recommended to users. The greater the Precision is, the higher the accuracy of the recommendation will be.

### C. Experimental Results and Analysis

In order to verify the authenticity and superiority of the proposed algorithm, the proposed algorithm is compared with the collaborative filtering algorithm based on SVD, and the collaborative filtering algorithm based on K-means clustering. On the test set, experiment selects 40 percent of the reserved energy of the matrix and cluster number 16.

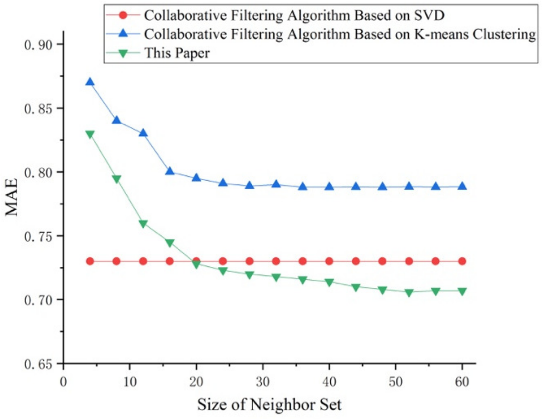

Fig. 2. Compare The MAE of Different Collaborative Filtering Algorithms

As shown in figure 2, the MAE of the algorithm in this paper is always lower than the collaborative filtering algorithm based on K-means clustering. Although the size of the neighbor set is less than 20, the performance of this algorithm is not as good as the collaborative filtering algorithm based on SVD. However, as the size of the neighbor set increases, the algorithm is superior to the other three algorithms. It is not comprehensive to verify the algorithm based on only one standard. Therefore, this paper further compares the three algorithms with Precision as the evaluation standard. The comparison results are shown in Figure 3.

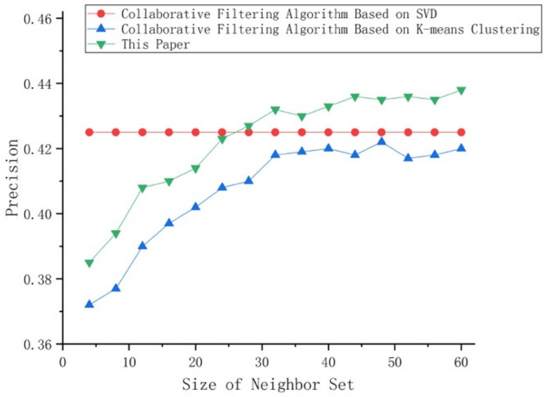

Fig. 3. Compare The Precision of Different Collaborative Filtering Algorithms

It can be seen from figure 3 that the accuracy of the algorithm in this paper is higher than the other two algorithms when the nearest neighbor number is greater than 25 and tends to be stable when the neighbor size is 45. When the nearest neighbor number is less than 25, the accuracy of the algorithm is between the SVD-based collaborative filtering algorithm and the other. By comparing the MAE and Precision of different collaborative filtering algorithms, the experimental results show that the proposed algorithm based on local optimization SVD and K-means has better prediction accuracy.

## IV. Conclusion

This paper proposes a collaborative filtering algorithm based on dimension reduction of local optimization and clustering. The SVD dimension reduction technology based on local optimization can improve the impact of the sparseness problem of traditional collaborative filtering algorithms on recommendation accuracy. K-means-based clustering technology can reduce the search time of nearest neighbors and has good scalability. The experimental results show that compared with the the collaborative filtering algorithm based on Pearson correlation and K-means clustering, the proposed method significantly improves the recommended performance. And with the increase of the number of neighbors, the prediction accuracy of this method is gradually higher than that of SVD-based collaborative filtering algorithm. The paper not only protects user privacy but also has good scalability and realizes real-time and accurate recommendation to users.

## Acknowledgment *(Heading 5)*

This paper is supported by the research project of University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province(Grant UNPYSCT-2018203), Natural Science Fundation of Heilongjiang Province(Grant YQ2019F011) and Fundamental Research Foundation for University of Heilongjiang Province(Grant LGYC2018JQ013).

## References

[1] M. Kunaver, T. Požrl, “Diversity in recommender systems- A survey,” Knowledge-Based Systems, vol. 123, pp. 154–162, 2017.

[2] X. Ye, P. Yuan, X. Guo, and Z. Yan, “Collaborative filtering recommendation algorithm based on user interest and project cycle,” Journal of Nanjing University of Science and Technology, vol. 42, no. 4, pp. 392, 2018.

[3] M. Fu, H. Qu, Z. Yi, “A Novel Deep Learning-Based Collaborative Filtering Model for Recommendation System,” IEEE Transactions on Cybernetics, pp. 1–13, 2018.

[4] A. Javari, J. Gharibshah, M. Jalili, “Recommender systems based on collaborative filtering and resource allocation,” Social Network Analysis & Mining, vol. 4, no. 1, pp. 234, 2014.

[5] K. Kim, H. Ahn, “Recommender systems using cluster-indexing collaborative filtering and social data analytics,” International Journal of Production Research, pp. 1–13, 2017.

[6] H. Koohi, K. Kiani, “User based Collaborative Filtering using fuzzy C-means,” Measurement, vol. 91, pp. 134–139, 2016.

[7] B. H. Le, K. Q. Nguyen, “Thawonmas R. Bounded-SVD: A Matrix Factorization Method with Bound Constraints for Recommender Systems,” Journal of Information Processing, vol. 24, no. 2, pp. 23–26, 2015.

[8] Z. Jia, Y. Lin, M. Lin, “An effective collaborative filtering algorithm based on user preference clustering,” Applied Intelligence, vol. 45, no. 2, pp. 230–240, 2016.

[9] O. J. Lee, M. S. Hong, J. J. Jung, J. Shin, and P. Kim, “Adaptive Collaborative Filtering Based on Scalable Clustering for Big Recommender Systems,” Acta Polytechnica Hungarica, vol. 13, no. 2, pp. 179–194, 2016.

[10] X. Guan, C. T. Li, Y. Guan, “Matrix Factorization With Rating Completion: An Enhanced SVD Model for Collaborative Filtering Recommender Systems,” IEEE Access, vol. 5, no. 99, pp. 27668–27678, 2017.

[11] F. Cacheda, C. Víctor, F. Diego, “Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems,” Acm Transactions on the Web, vol. 5, no. 1, pp. 1–33, 2011

[12] Agnivesh and R. Pandey, “Elective Recommendation Support through K-Means Clustering Using R-Tool,” IEEE International Conference on Computational Intelligence and Communication Networks, 2016, pp. 851–856.

[13] F. M. Harper, J. A. Konstan, “The MovieLens Datasets: History and Context,” ACM, 2015.

[14] J. B. Schafer, F. Dan, J. Herlocker, “Collaborative Filtering Recommender Systems,” Acm Transactions on Information Systems, vol. 22, no. 1, pp. 5–53, 2004.