Journal of Advances in Computer Research
Quarterly ISSN: 2008-6148
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 12, No. 1, February 2021), Pages: 13-25
www.jacr.iausari.ac.ir

# Provide a Video Recommendation System Using Collaborative Filtering and Data Mining Methods

**Reza molaee fard***
*Department Of Computer Engineering - Dezful Branch of Azad University, Khuzestan, Iran*
Email: rezamolae4@gmail.com

**Received: 2021/01/13; Accepted: 2021/05/01**

**Abstract**

*Due to the growing number of videos available on the web, it seems necessary to have a system that can extract users' favorite videos from a huge amount of information that is increasing day by day. One of the best ways to do this is to use referral systems. In this research, a method is provided to improve the recommender systems in the field of film recommendation to the user. In this research, DBSCAN clustering algorithm is used for data clustering. Then we will optimize our data using the cuckoo algorithm, then the genetic algorithm is used to predict the data, and finally, using a recommender system based on participatory refinement, a list of different movies that can be of interest to the user is provided. The results of evaluating the proposed method indicate that this recommender system obtained a score of 99% in the accuracy of the system and a score of 95% in the call section Suggest the user's favorite videos correctly to the user.*

***Keywords:*** *Recommender system, DBSCAN algorithm, cuckoo algorithm, genetic algorithm, participatory filtering.*

## 1. Introduction

Referrer or bidder systems are systems that, by taking limited information from the user and features such as what information the former user has searched for and what privileges he has given to the goods, can provide appropriate suggestions to the user that may be of interest to him. Recommending systems today offer attractive items to the user in many different applications. One of the most important applications of recommendation systems in the field of web, digital libraries, restaurant industry, tourism industry, film recommendation and other environments where there is a large amount of information, can provide appropriate suggestions to the user. One of the most important and popular systems that users have been interested in from the beginning until now, has been the recommendation system related to movie suggestion systems. Due to the large volume of videos available on the web, which is increasing day by day, the existence of a system that can extract users' favorite videos and can suggest to the user seems necessary. Recommender systems allow movies to be suggested to the user by features such as movie title, movie type, actor or writer name, or other features related to a movie. The most important type of recommendation system for recommending videos to the user is the recommendation system based on participatory filtering. This type of recommendation system starts by finding users who have preferences and purchase history of subscriptions with the current user, then collects information provides a list of

recommendations for the group's favorite items and the removal of items previously purchased by the user. This group of systems is called participatory refinement systems, which are among the most widely used systems for generating recommendations to users. The main mechanism of the participatory refinement algorithm is that using the similarity criterion of individuals, the preferences of large groups of users are recorded. Next, users who have the same priorities as the current user are selected as their neighbors, then the average of the priorities is calculated and the final priority function tries to recommend an item that the user has not scored. In this research, a method is provided to improve the recommendation systems in the field of film suggestion to the user. Previous research had drawbacks that this study tried to address. Among these problems were problems with how to cluster. Previous research has mostly used hierarchical methods such as k-means or C-means algorithms. In this research, we tried to increase the clustering accuracy by using the DBSCAN algorithm. Another issue is the use of genetic algorithms for the prediction part, which was able to achieve better performance than other existing methods. Another disadvantage of the previous methods was the lack of a method to optimize the results, which we optimize in this study using the cuckoo algorithm. The method is that after collecting the videos that have already been searched by different users and given ratings to these videos, as well as information about users and their tastes, then we process this information, then we cluster our data, then to produce Prediction We will use a genetic algorithm and in the last step, using a recommendation system based on participatory filtering, a list of recommendations that can be of interest to the user will be provided to the user.

## 2. Research background

In their 2020 article, Leo and Seda provided a way to recommend the film to users. In this study, which was conducted with the aim of discovering different groups to recommend movies to users, they tried to solve the problems of users to find their favorite movies. To develop the proposing system, the researchers used several algorithms to obtain user grouping, such as the k-means clustering algorithm, the mean change algorithm, the scatter algorithm, and the spectral clustering algorithm. In this study, in order to group the films based on the film genre and film labels, the K-means clustering method was used, which resulted in better and more practical recommendations to the user [1].

In their 2020 article, Leung and Girua presented a way to improve film recommendation based on users' emotions. The researchers used users' emotional vectors and movie emotion vectors as components of the recommendation system. In this study, a platform consisting of five recommenders based on content-based filtering algorithms and participatory filtering was used to classify the emotional indexes of the film by reviewing the film based on the number of repetitions of the observed videos. Users were then formulated from the movie watching history using the average of all the movies the user had watched. In evaluating this system, many smart and useful recommendations were observed [2].

In 2020, Leo provided a way to recommend the film to the user. The researcher believed that with the development of the entertainment and film industry, users could have a better chance of

accessing their favorite movies. In this research, ranking and mainly the average score of users have been used to recommend videos. He categorized movies such as romance, comedy, drama, and horror movies based on movie tagging and recommended movies based on users' viewing tastes. Finally, some recommendation methods based on machine learning such as factorization, attention factorization machine, extensive learning and deep factorization device were used, which had many advantages [3].

In his 2020 paper, Gupta proposed a way to improve film recommendation systems. In this research, a film proposing framework called Big Screen has been used. In this method, a content-based screening approach that analyzes the data used by the customer is then recommended to the user to videos that may be of interest to the user. The list of recommended videos with the ratings provided by previous customers is offered to the user. The proposed method framework presents the proposals using different types of information about customers, available items and past equations in the modified databases, then the customer can review the offer and decide to select his favorite movie [4].

## 3. Recommender system

Recommending systems have become very important in recent years. The goal of any proposing system is for consumers to be able to find new goods or services, such as movies, articles, the Web, books, music, restaurants, or even people, based on information about the consumer or recommended [6, 5]. The recommender system is a system that jointly recommends items to a group of users according to the user's preferences [7]. Recommender Systems emerged after that web was transformed to an interactive media by allowing users to provide their feedbacks [16]. Referral systems are systems that help the user find interests in situations of over-information. Where the user's preferences are estimated based on the behavior observed in the past and can provide the user with a ranked list of suggestions [8].

## 4. Proposed method

In the proposed method, a new method is presented in order to improve the recommendation systems in the field of offering video recommendations to users. In this method, after collecting information about users as well as rated videos by users, we must perform data preprocessing operations on our data. We perform data preprocessing operations because our data is raw data and cannot be injected raw into data mining algorithms. After data processing, in order to be able to extract similar users as well as more videos viewed by users, we need to cluster our data and thus obtain the weight of the obtained pages for the user's interest in these items. After performing the clustering operation, we need to optimize our data for a better output using the cuckoo algorithm. Then, in order to produce an accurate prediction, we evaluate our data using a genetic algorithm. Then, in the last step, using a recommendatory system based on participatory filtering, we provide the user with information that this information can be of interest to the user. Figure (1) shows an overview of the proposed method.

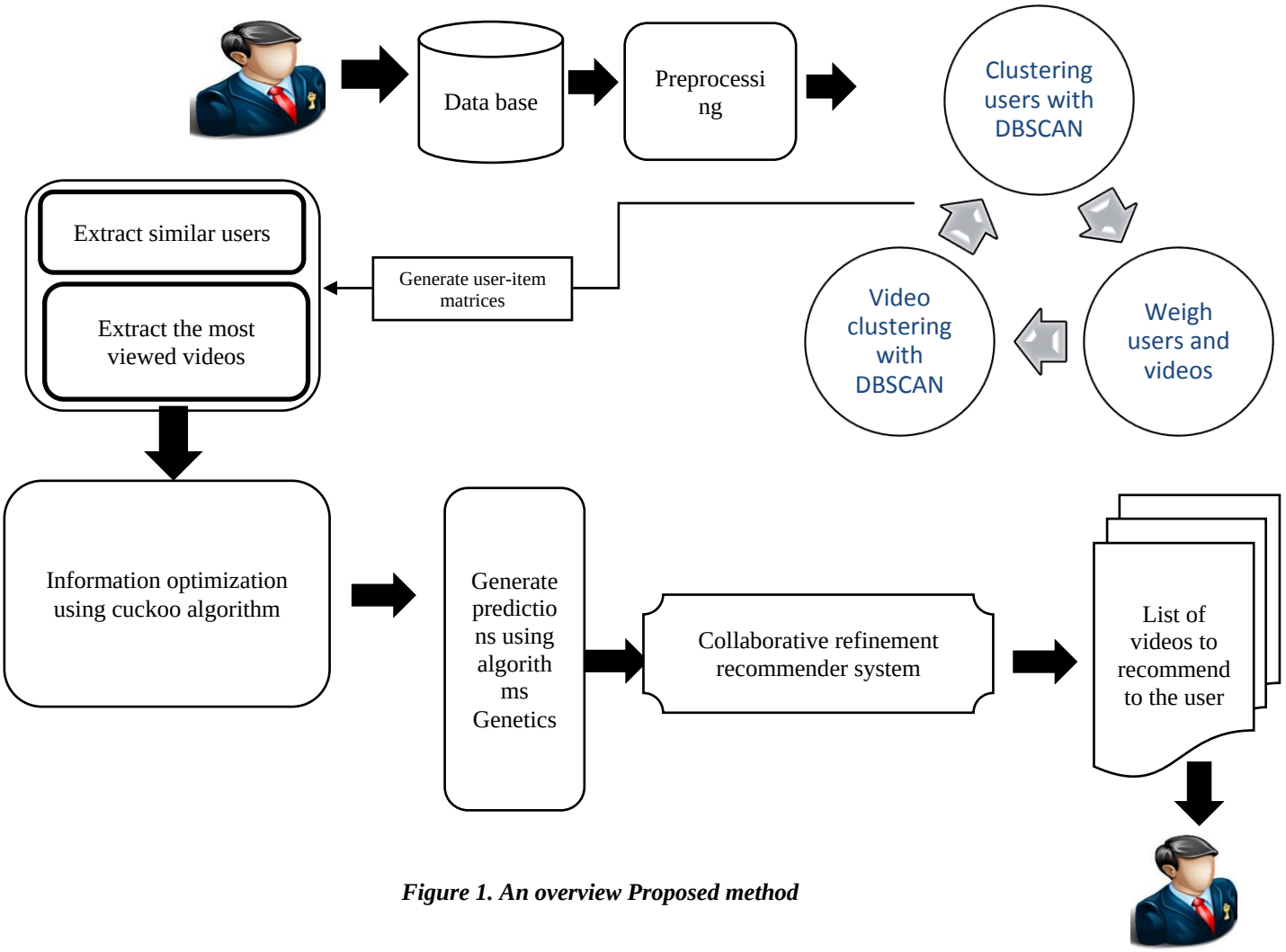


***Figure 1. An overview Proposed method***

### *4.1 . Data processing*

In the first step of the proposed method, we must first perform the data preprocessing operation, because it is usually not possible to extract the data in raw form into data algorithms. In order to prepare the data, it is necessary to take them out of their original form and state and transform them into a form that is suitable for the algorithm. If different data are pre-processed, the same reliable and effective performance will occur in all datasets [9]. Also, the available data usually have different extras that may confuse the algorithm. In data mining we also need to remove extra data that does not help the problem and the algorithm. Data preprocessing operations are usually performed before the main operation of data mining algorithms and facilitate and assist the algorithms. Data processing is an important step towards successful data mining.

### *4.2. Clustering of users and videos*

Next we need to cluster our data. The preferred method for data clustering is the use of the DBSCAN clustering algorithm. The way this algorithm works is that DBSCAN starts with a desired starting point that has not been visited. The range of this point is

extracted using the epsilon distance (all points in the distance ε are group points or neighbors). It should be noted that the algorithm uses the Euclidean distance to find neighbors in a two-dimensional and three-dimensional space. Neighborhood is thus defined by the smallest distance from the principal point. If there are enough min points in this range, the clustering process starts (border point) and the current data point becomes the first point of the cluster in the new cluster, otherwise the point is considered as noise (later this)[10] . The noise point may be part of the cluster. In both cases, the point is specified as visited. For this point in the new cluster, the points in the range ε are also part of a cluster. This method is used to construct all points in the ε group belonging to the same cluster and then it is repeated for all new points that are only added to the cluster group. This process is repeated until all points in the clusters are entered, i.e. all points in the ε range of the clusters are visited and tagged [11, 12].

### *4.3. Extract similar users*

Once a new user of a cluster or class has been identified, its neighbors, which include users in that cluster, are extracted. The comments of these neighbors are effective in the final offer of the film to the new user; But not all neighbors are alike similar to the new user, and a similarity criterion should be used for closer neighbors. Assume system users as a set U= $\{u_1, u_2, u_3, \dots, u_m\}$ with properties D= $\{d_1, d_2, d_3, \dots, d_n\}$ and the set of movies should be defined as I=$\{i_1, i_2, i_3, \dots, i_k\}$ Then the similarity of a new user and each of the neighbors is calculated based on the following equation.

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (Xi - yi)} \qquad (1)$$

### *4.4. Generate user-item matrices*

To do this, consider C as all users and S as all items (videos) that can be suggested to the user. The u utility function expresses the utility of the s item for c users. The total set of orders is denoted by R, which we define as $C \times R \rightarrow R$. Then, for each $c \in C$ user, we define a clause such as $s \in S$ that maximizes user usability as a contract according to the following relation.

$$\forall_c \in \boldsymbol{All\ user}(\boldsymbol{C}),\ \boldsymbol{Items}_c(\boldsymbol{S}) = \arg\max \mathbf{Utility\ function}(\mathbf{U})\big(\boldsymbol{All\ user}, \boldsymbol{items}\big) \qquad (2)$$

In the recommender system we have to create a matrix of users and items. In this method, each user votes for an item, that point is stored in the desired cell in the matrix. Once a user's vote is determined, it can be used to determine a bid for similar users. Obviously, this matrix will be thin and solitary. A recommending system should anticipate these dispersions and suggest them to the user if the prediction score is high. The most widely used criteria for evaluating the system are the use of MSE and MAE criteria. Items that two users voted for (user to user) or all user ratings from two items (item to item) are compared.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i)^2 \quad (3)$$

$$MAE = \frac{\sum(u.i) \in |Rtest| R_{u.i} - R_{u.i}|}{|Rtest|} \quad (4)$$

### *4.5. How to score*

This section deals with the concept of points and the point's matrix. Since user feedback plays a key role in the participatory filtering technique, it is necessary to design methods and templates for collecting it. In the literature of recommending systems, different methods have been introduced to collect user feedback, but the common method used by most refining-based recommending systems is to consider a numerical interval (for example, 1 to 5) for each item, defining the meaning of each. These numbers (for example, 1: very bad, 2: bad, 3: average, 4: good, 5: very good) and ask the user to map one of these numbers to each of the items he sees. These numbers are called scoring systems in the literature and this method is called scoring.
The following table shows the score matrix of a recommendation system based on participatory multi-movie filtering:
You can see how to rate movies in Table 1.

***Table1. How to rate movies***

| Film 1 | Film 2 | Film 3 | Film 4 | User ID |
|---|---|---|---|---|
| 5 | 2 | 5 | 2 | **User 1** |
| 1 | 3 | 4 | 4 | **User 2** |
| 3 | 5 | 1 | 2 | **User 3** |
| 4 | 1 | 5 | 5 | **User 4** |

CF Techniques use a database of preferences for items by users to estimate additional topics or create new users. In a typical CF scenario, there is a list of m users $u1, u2, u3..um$ {as well as a list of n items {$i1, i2, i3...in$}and each Ui user has a list of items. Has lui that the user has rated or those whose preferences have been inferred through their behavior. The rank can be explicit references and ... which is on a scale of 1 to 5 or it can also be an implicit reference.

### *4.6. The nearest neighbor-based algorithm*

This type of algorithm uses the scores given by a similar user to predict a user's interest in a particular item. These similar users are called user neighbors. If n is the same

as user u, n is said to be a neighbor of u. To predict user's interest in item i, the average score given by u neighbors (including user n) to i should be calculated equal to (1.2). In this equation, rni is the score that user n gave to item i. We see how to calculate this algorithm in the following relation.

$$\boldsymbol{pred(u.i)} = \frac{\sum \mathbf{ncneighbors(u)\,r\,ni}}{\mathbf{number\,of\,neighbors}} \tag{5}$$

The prediction of the score that user u will give to item i is obtained by calculating the weighted sum of user points u on similar item i.

$$\boldsymbol{Pred(u,i)} = \frac{\sum_{j \in rateditems(u)} \boldsymbol{itemsim(i,j).rui}}{\sum_{j \in rateditems(u)} \boldsymbol{itemsim(i,j)}} \tag{6}$$

Information optimization using cuckoo algorithm

In the next step, we have to optimize the patterns obtained from the previous step to improve the results and increase the accuracy. The proposed optimization method is to use the cuckoo algorithm.

In the next step, we will optimize the node energy using the cuckoo meta-heuristic algorithm. The cuckoo algorithm is one of the newest and most powerful evolutionary optimization methods. This algorithm uses the lifestyle of a bird named Face. This algorithm starts with an initial population. This population of cuckoos has a number of eggs that they will lay in the nest of a number of host birds. Some of these eggs are more similar to the eggs of the host bird, more likely to be the eggs of the host bird, more likely to grow into an adult cuckoo [13, 14]. Other eggs identified by the host bird are destroyed. The amount of eggs grown indicates the suitability of the nests in that area. The more eggs that can live and survive in an area, the more profit will be made to that area; Therefore, the situation in which the largest number of eggs are saved will be a parameter that the cuckoos intend to optimize [15, 16].

- To solve an optimization problem, it is necessary to form the values of the problem variables in the form of an array.
- In the cuckoo algorithm, the array is called Habitat.
- in the next $N_{var}$ optimization problem, a Habitat will be an array var $1 \times_{var}$ that represents the current living position of the cuckoos. This array is defined as follows.

$$\boldsymbol{Habitat} = [x_1, x_2, x_3, ., \ldots, x_{Nvar}] \tag{7}$$

- The appropriateness of the current Habitat is obtained by evaluating $(f_p)$ in the Habitat, therefore:

$$\boldsymbol{Profit} = \boldsymbol{f_p}\ \ \boldsymbol{habitat} = \boldsymbol{f_p}\left(x_1, x_2, x_3, ., \ldots, x_{Nvar}\right) \tag{8}$$

- As can be seen, the cuckoo algorithm is an algorithm that maximizes the profit function.

- To use the cuckoo algorithm to solve the minimization problems, it is enough to multiply a negative sign in the cost function.
- To start the optimization algorithm, we generate a Habitat matrix with size $N_{pop} \times N_{var}$.
- A number of random eggs are then assigned to each of these habitats.
- In nature, each cuckoo lays between 5 and 20 eggs. These numbers are used as the upper and lower limits of each cuckoo egg allocation in different iterations.

$$ELR = a \times \frac{Number\ of\ current\ cuckoos\ eggs}{Total\ number\ of\ eggs} \times \left(Var_{hi} - Var_{low}\right) \quad (9)$$

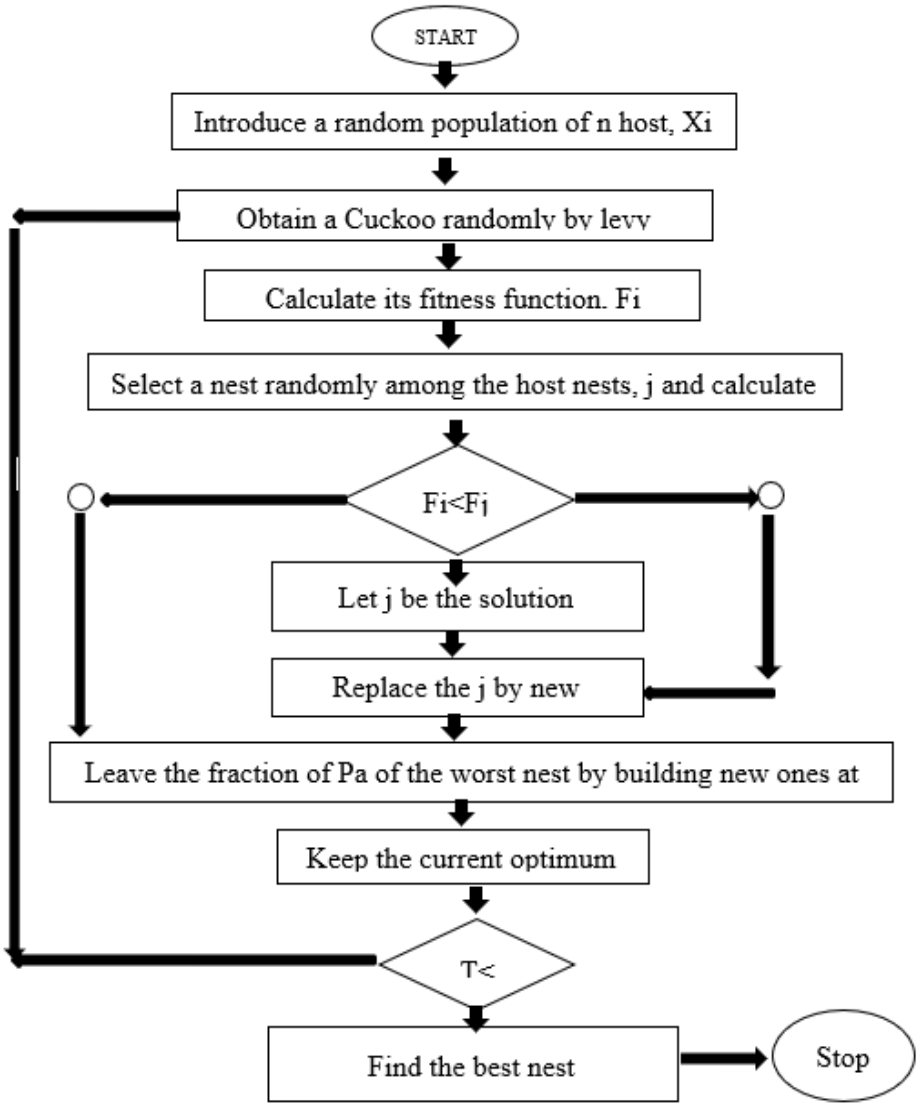


***Figure 2. The framework of the cuckoo algorithm.***

### *4.7. Generate predictions using genetic algorithms*

We use genetic algorithms to generate predictions. In order to be able to solve a problem with a genetic algorithm, we have to convert them to the special form required by this algorithm. In this process, we must define the required solution to the problem in such a way that it is represented by a chromosome. The method of work is that first the most suitable member of each community is selected. Then the element with the most fitting number is selected. As the average fit of society increases, so does the weight of choice. This method is useful when the set contains elements that have a large number of fit and only small differences distinguish them. Calculating the temporal complexity of a genetic algorithm depends on many factors, including population size, genotype length, algorithm termination conditions, population selection type, fitness function, crossover function, and mutation function [17]. In features selection, we take the number of each feature into account as a criterion for selecting that feature. In this case, the features which their numbers are on the chromosome will participate in data classification and others are not included [18]. A subset of the attributes is selected and the members of that set compete with each other, eventually the selected attributes will be used to generate a prediction.

## 5. Collaborative refinement recommender system

In the last step, we will store the data obtained from the previous steps in a recommender system based on participatory filtering and provide it to the user for recommendation. This type of recommender system starts with finding users who have preferences and purchase history of subscriptions with the current user, then collects information about the group's favorite items and deletes those items previously purchased by the user. This group of systems is called participatory refinement systems, which are among the most widely used systems in order to generate recommendations to users. The main mechanism of the participatory refinement algorithm is that using the similarity criterion of individuals, the preferences of large groups of users are recorded. Next, users who have the same priorities as the current user are selected as their neighbors, then the average of the priorities is calculated and the final priority function tries to recommend an item that the user has not scored.

## 6. Evaluate the proposed method

We use MSE and MAE criteria to evaluate the user-item matrix. You can see the results obtained from these two criteria in the following tables.

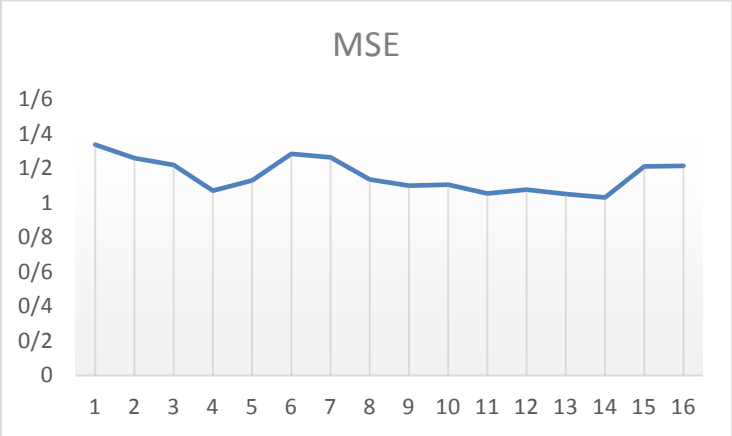


***Figure 3. MSE value of the proposed method***

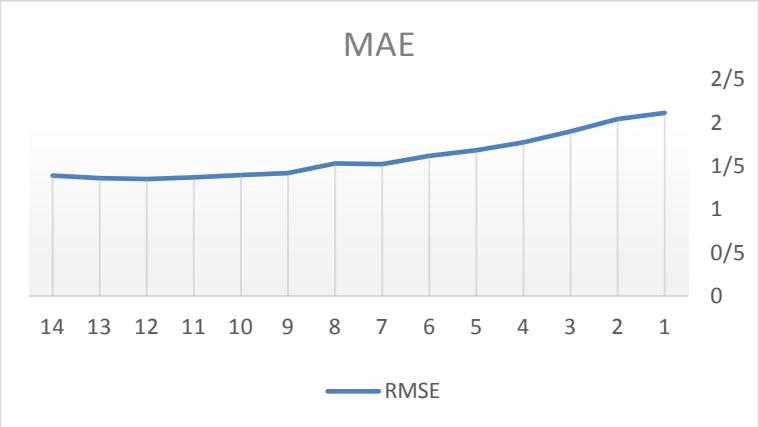


***Figure 4. MAE value of the proposed method***

It is often used to validate recommender systems such as system accuracy and item recall. In this research, these criteria have been used to evaluate the system. Accuracy and recall in recommender systems are calculated using the following two equations. Accuracy is calculated using the following equation.

$$\boldsymbol{Precision} = \frac{|\{\boldsymbol{relevant\ item} \cap \{\boldsymbol{retrved\ item}\}|}{|\{\boldsymbol{retrived\ item}\}|} \quad (10$$

The call is calculated using the following equation.

$$\boldsymbol{Recall} = \frac{|\{relevant\ item \cap \{retrved\ item\}|}{|\{relevant\ item\}|} \quad (11$$

To evaluate the accuracy and convenience of the system, a comparison was made between the proposed method and the algorithms of gray wolf, ant colony and PSO, the results of this comparison can be seen in the following diagrams.

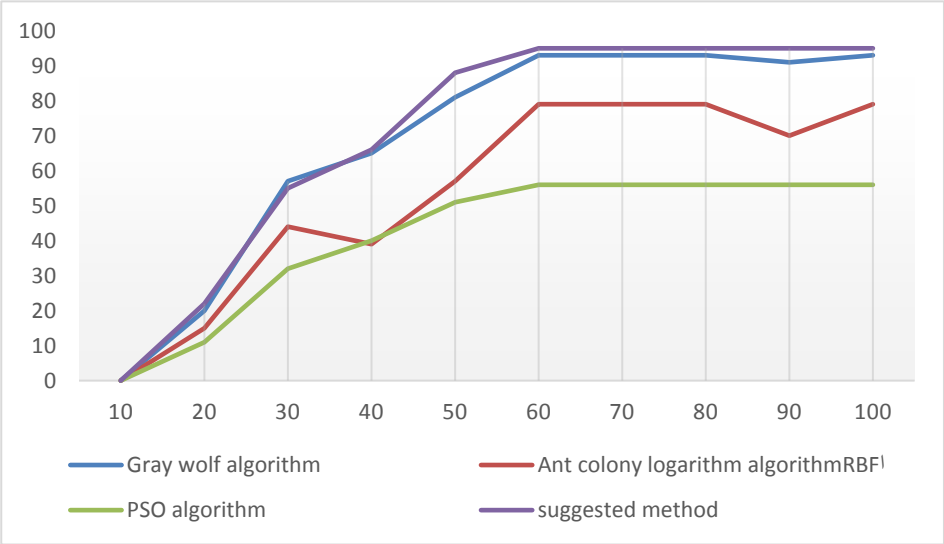


***Figure 5. Comparison diagram of the proposed method with other methods***

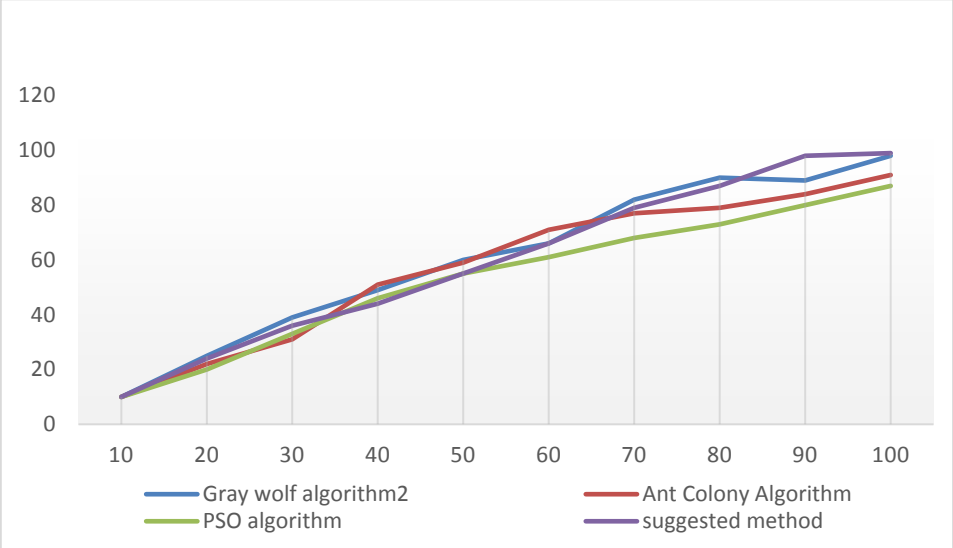


***Figure 6. Diagram of comparing the accuracy of the proposed method with other methods***

## 7. Conclusion

In this study, a new method was introduced to improve the recommendation systems in the field of film to users. Due to the growing content and content of web-based videos, it seems necessary to have a system that can extract users' favorite videos on the web and suggest them to the user. To do this we need to personalize our systems. One of the best ways to do this is to use referral systems. Referrer systems are systems that can provide the user with a list of items that may be of interest to the user by obtaining limited information from the user and features such as items searched by a past user. In this research, using a recommending system based on participatory filtering and using data mining methods, an attempt was made to design a system that can solve the problems of

previous systems and provide appropriate suggestions to the user. In this system, after collecting the data related to the user's favorite movies, first the data processing operation was performed on the desired data. We then clustered our data to evaluate the interest and similarity of the items using the DBSCAN clustering algorithm. The results of evaluating the efficiency of the DBSCAN algorithm showed that this clustering method was more efficient than other existing methods. Then, we optimized the obtained data by Cuckoo algorithmic algorithm, and finally, using genetic algorithm to generate predictions, and finally, using a recommendatory system based on participatory filtering, we extracted the user's interests and targeted the user. The results of the evaluation of the proposed method showed the accuracy and recall of the proposed method compared to other available methods, and according to the obtained statistics, it can be said that the proposed method can offer up to 90% of the user's desired information.

**References**

[1] Cintia Ganesha Putri, D., Leu, J. S., & Seda, P. (2020). Design of an Unsupervised Machine Learning-Based Movie Recommender System. *Symmetry*, *12*(2), 185.

[2] Leung, J. K., Griva, I., & Kennedy, W. G. (2020). Text-based Emotion Aware Recommender. *arXiv preprint arXiv:2007.01455*.

[3] Lu, B. (2020). Deep learning based movie recommender system.

[4] Gupta, A. (2020). A Movie Recommender System: BIG SCREEN. *Journal of Advanced Research in Information Technology, Systems & Management*, *4*(1), 9-13.

[5] Alashkar, T., Jiang, S., Wang, S, and Fu, Y., 2017, "Examples-Rules Guided Deep Neural Network for Makeup Recommendation," Proc. 31st AAAI Conference on Artificial Intelligence, .pp.941–947.

[6] Gupta, K. D. (2019). A Survey on Recommender System. *International Journal of Applied Engineering Research*, *14*(14), 3274-3277.

[7] Dara, S., Chowdary, C. R., & Kumar, C. (2020). A survey on group recommender systems. *Journal of Intelligent Information Systems*, *54*(2), 271-295.

[8] Jannach, D., Manzoor, A., Cai, W., & Chen, L. (2020). A Survey on Conversational Recommender Systems. *arXiv preprint arXiv:2004.00646*.

[9] Alexandropoulos, S. A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. *The Knowledge Engineering Review*, *34*.

[10] George,seif. the 5 clustering algorithms data scientsts need to know-a36d125ef6..Toward Data Science-2018.

[11] Siddharth Agrawal. Machine learning-DBSCAN. Toward Data Science-2019.

[12] Zatni , abdelkarim .2018 .Document text Detection in video frames acquired by a smartphone based on line segment detector and DBSCAN clustering .Journal of engineering science and technology,vol.13,no.2,540-557.

[13] Boveiri, H. R. (2020). An enhanced cuckoo optimization algorithm for task graph scheduling in cluster-computing systems. *Soft Computing*, *24*(13), 10075-10093.

[14] Cai, X., Niu, Y., Geng, S., Zhang, J., Cui, Z., Li, J., & Chen, J. (2020). An under- sampled software defect prediction method based on hybrid multi- objective cuckoo search. *Concurrency and Computation: Practice and Experience*, *32*(5), e5478.

[15] Inci, M., & Caliskan, A. (2020). Performance enhancement of energy extraction capability for fuel cell implementations with improved Cuckoo search algorithm. *International Journal of Hydrogen Energy*.

[16] Nemati, Y., & Khademolhosseini, H. (2020). Devising a Profit-Aware Recommender System using Multi-Objective GA. *Journal of Advances in Computer Research*, *11*(3), 109-120.

[17] Monireh Taheri Sarvtamin, (2019). Static Task Allocation in Distributed Systems Using Parallel Genetic Algorithm. *Journal of Advances in Computer Research*,11(4),57-72

[18] Shokripoor Bahman Bigloo, I. (2019). A Parallel Genetic Algorithm Based Method for Feature Subset Selection in Intrusion Detection Systems. *Journal of Advances in Computer Research*, *10*(2), 1-16.