Cold Start Aware Hybrid Recommender System Approach for E-Commerce Users

S Gopal Krishna Patro
Gandhi Institute of Engineering and Technology
Brojo Kishore Mishra
Gandhi Institute of Engineering and Technology
Sanjaya Kumar Panda
National Institute of Technology Warangal
Raghvendra Kumar
Gandhi Institute of Engineering and Technology
Hoang Viet Long (sonntkmath@gmail.com)
Ton Duc Thang University
David Taniar
Monash University

Research Article

Keywords: Data Sparsity, Cold Start, Hybrid Recommendation, Data Decomposition, Clustering.

Posted Date: November 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-792132/v1

Cold Start Aware Hybrid Recommender System Approach for E-Commerce users

S Gopal Krishna Patro[1], Brojo Kishore Mishra[1], Sanjaya Kumar Panda[2], Raghvendra Kumar[1], Hoang Viet Long[3,4,*], David Taniar[5]

[1]Department of Computer Science and Engineering, GIET University, India sgkpatro2008@giet.edu, bkmishra@giet.edu, raghvendra@giet.edu
[2]Department of Computer Science and Engineering, NIT Warangal, India. sanjaya@nitw.ac.in
[3]Division of Computational Mathematics and Engineering, Institute for Computational Science, Ton Duc Thang University, Ho Chi Minh City, Viet Nam.
[4]Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Viet Nam, hoangvietlong@tdtu.edu.vn
[5]Faculty of Information Technology, Monash University, Melbourne, Australia. Email: David.Taniar@monash.edu

* Corresponding author: Hoang Viet Long (hoangvietlong@tdtu.edu.vn)

Abstract: Collaborative Filtering (CF) schemes are very popular in Recommender System (RS) and offer specialized suggestions to users in e-commerce and social websites. But, they suffer from the Cold Start Problem (CSP) that occurs due to the lack of sufficient information about the new customers, purchase history, and browsing data. Moreover, data sparsity problems may arise when the interaction is made among a limited amount of items. This not only poses a negative impact on recommendation but also significantly condenses the diversity of choices available in the particular platform. To tackle these issues, a novel methodological approach called Sparsity and Cold Start Aware Hybrid Recommended System (SCSHRS) is designed to suppress data sparsity and CSP in RS. The proposed SCSHRS methodology comprises four stages. At the initial stage, the data sparsity is reduced and at stage 2, the similar users are grouped by Ant-Lion based k-means clustering. At stage 3, Higher-Order Singular Value Decomposition (HOSVD) method decomposes the data to a lesser dimension. At the final stage, the Adaptive Neuro-Fuzzy Inference System (ANFIS) uses IF-THEN rules and machine learning abilities to predict the output. The performance of the proposed SCSHRS method is tested on MovieLens-20M, Last. FM, and Book-Crossing datasets and compared with the prevailing techniques. Based on the evaluation report, the proposed SCSHRS system gives Mean Absolute Percentage Error (MAPE) of 40%, and, precision (0.16), recall (0.08), F-measure (0.1), and Normalized Discounted Cumulative Gain (NDCD) of 0.65. Hence, SCSHRS is proved to be a more efficient means of recommendation against cold start and sparsity problems.

Keywords: Data Sparsity; Cold Start; Hybrid Recommendation; Data Decomposition; Clustering.

1 Introduction

The amount of information found on the internet is growing exponentially day by day. This allows users to access any kind of data. However, it increases other complex problems in RS which is the difficulty of finding suitable items for the intended users. The supplement information and items make it more difficult to find the relevant products for a particular user. RS helps us to make decisions or find what we need. The RS is designed to predict the interest of users and suggest products that are most likely interesting to them [1]. To escalate the sales of the product, RS is a great machine learning system for online dealers [2]. The data required for RS comes from obvious user ratings when seeing the product picture, implicit searches, buying histories, or some other facts about the customers or product [3, 4]. Companies utilizing RS emphasize growing trades through highly customized suggestions and better user experience [5]. Suggestions usually haste up the search results and make it easier for the users to access contents or items that are interesting to them, as well as provide them with offers they would have never looked for [6]. Moreover, companies can attract and retain clients with movies and TV shows that match their profiles.

RS functions with two types of information: characteristic information and user-item interactions. Characteristic information gives data regarding items and users like profiles, categories, keywords, preferences, etc. [7]. Whereas, user-item interfaces include records like purchase history, likes, ratings, etc. [8, 9]. On the basis of these criteria, RS algorithms can be distinguished as (i) Content-based (CBF) RS (ii) Collaborative Filtering (CF) based RS and (iii) Hybrid RS. CBF uses characteristic information for recommendation [10] and CF systems works based on the relation among user-item [11]. Hybrid systems combine both types of information to avoid demerits that occur when using a single type [12].

Enough information (i.e., interactions among user and item) is needed for the RS to work efficiently. In an e-commerce site, it is hard to provide suggestions unless the users have shown interest in a certain amount of items. If we setup a new e-commerce site, we cannot give recommendations until users have interacted with a significant number of items. Lack of these data may lead to CSP [13, 14]. When the quantity of items increases, the user's view of all the available products cannot be expressed. So, there is always a lack of adequate product ratings. Due to this, the sparsity problem occurs which makes the accuracy of the recommendations low [15, 16].

1.1. Motivation

RS is developed in the E-commerce area to recommend accurate items to the users. However, CSP and data sparsity makes recommendation difficult for new users due to the lack of user-item interaction and unavailable ratings. These two issues affect the prediction accuracy of the RS. Various research groups are working to get an optimized and accurate prediction algorithm for RS. Among them, clustering and dimensionality reduction techniques make a huge impact in enhancing the recommendation accuracy. Some of the developed methods that are utilized for dimensionality reduction and clustering include Singular Value Decomposition (SVD), Matrix Factorization (MF), ontology, Deep Neural Network (DNN), k-means, fuzzy [17-22], etc. Among these techniques, k-means is one of the widely used clustering techniques in RS. It assigns the data points to a cluster that has higher similarity. However, the conventional k-means clustering does not have any technique for selecting appropriate initial centers. If the centers are chosen randomly, they may get stuck to a local solution which in turn decreases the recommendation accuracy. This motivates us to develop an efficient scheme for RS.

1.2 Contributions

The contributions of the proposed SCSHRS methodology are as follows:

- The users may not give feedback or opinions regarding all the items available on the platform. This lack of user ratings on a particular item reduces the recommendation accuracy. Hence, the unavailable ratings are initially predicted before the actual process.
- The Ant Lion Optimization (ALO) used in the clustering stage helps to determine the initial center of k-means and thus enhances the recommendation accuracy.
- Moreover, HOSVD decomposes higher-order data to lower dimensions. This increases the computational speed. Finally, the acquisition of knowledge by manual power is replaced by the ANFIS system. This mechanism does not depend on the individual experts rather, the parameters are determined by minimizing the error through a training process. This makes the prediction quick and accurate.

The remaining sections are structured as follows: Section 2 briefly explains the recent works in the recommendation system, section 3explains the proposed SCSHRS approach, section 4 provides the experimental setup and comparative analysis, section 5 provides some discussions regarding the proposed SCSHRS approach, and finally, section 6 concludes the paper with some possible directions for the future work.

2 Literature Review

Some of the recently proposed techniques in RS are reviewed in this section.

Kiran *et al.* [23] proposed a hybrid RS based on DNN by integrating the data regarding the user into the DNN. This scheme comprises of hidden layer in which the hyper-parameters are tuned by a modeler. Then the linear functions in these layers are determined which is followed by dropout. The final step includes batch normalization and activation function. This scheme mitigates over-fitting problems by utilizing techniques like regularization and dropout. The simulations were performed in various datasets with distinct categories and have indicated less error. However, the predictive accuracy of this scheme has not been investigated.

Natarajan *et al.* [24] introduced a CF-based RS using Matrix Factorization (MF) to address CSP. To do this, the demographic data is gathered from the cloud and is lead to the regulator. Then this data is fed to the 'query constructor' to obtain the missing data aggregated from the knowledge base. Here, the Linked Open Data (LOD) interface is utilized to provide the data to the regulator and a 'miner' acts a filter to give only the necessary data. Finally, the similarity among the items is determined using the 'similarity calculator' and the suggested product is displayed to the new users. This MF technique eliminates SP and CSP but with the cost of high computational time.

Nilashi *et al.* [25] presented a CF-based recommendation by combining dimensionality reduction and ontology schemes to deliver accurate and scalable recommendations. This method comprises two stages. The first stage involves clustering by Expectation-Maximization (EM) and then SVD is applied to get the decomposed matrix. The data is clustered based on (i) User Clustering and (ii) Item Clustering (IC). User Clustering clusters the users having similar preferences. Whereas, Item Clustering clusters the products based on similar ratings. In the next stage, the system is trained in offline mode to predict and give the recommendation to the target users. Even though this scheme mitigates scalability and sparsity, the prediction mechanism is not so accurate.

Kermany and Alizadeh [26] presented a hybrid RS by employing Adaptive Neuro-Fuzzy Inference System (ANFIS) and ontology. It incorporates sematic as well as demographic data for recommending the movies. It has two modes: (i) offline mode and (ii) online mode. The parameters are tuned in the offline mode and the real-time inquires of the users are tested in the online mode. Initially, the similarities are determined in the offline mode and KNN is employed to choose the top relevant movies in the online mode. The weighting factors are obtained from the offline mode and then fed to the ANFIS model to predict the overall ratings. This technique has reduced sparsity issues as well as CSP but, the complexity and computational time tend to be high.

Wang *et al.* [27] presented a CF-based DNN framework that includes explicit and implicit feedback. The preferences of the users are indicated by numeric values and explicit feedback which is based on ratings. In case explicit feedback is not available, the implicit feedback is taken into consideration which includes browsing records, purchase history, click events, etc. The learning efficiency may decrease due to the imbalance between the users and the objects. Therefore, the meta-paths related to object type and user type are chosen. Then the nodes that have user type are filtered and given as the input to the skip-gram model. This helps to get the knowledge about the user behavior. Even though this technique eradicates new user problems, the sparsity issue is not considered.

Herce-Zelaya *et al*. [28] introduced an RS based on random decision forest to mitigate CSP. They used the data obtained from social site such as Twitter to suggest the product to the new users. To accomplish this, a decision tree was used to categorize the users based on their profiles. The model is then trained and verified to decide the prediction results. Even though the recommendation is performed better by this scheme, the usage of personal data obtained from social media causes security and privacy issues.

Guo *et al.* [29] presented an Attribute Fused SVD model (FASVD) to predict the user's preferences. This method uses historical rating and attributes data to generate an accurate RS. They used K Nearest Neighbor (KNN) technique to determine the similarity, in which the target user is replaced with the nearest neighbor. Then SVD is utilized to determine the missed ratings. At the following stage, the MF reduces the dimensionality of the rating matrix. Here, the MF maps the rating matrix based on user-based and item-based profiles. By including the attribute data of the items, the relation between the existing and the cold start items are predicted.

Jiang L *et al.* [30] set up a Single Slope scheme based on a combination of user identical features, and trusted data, which can be distributed across multiple system configurations. This scheme includes three actions. First, the trusted data is selected. Next, the resemblance between the users is determined. Thirdly, the resemblance is added to the weight component of the improved SS scheme, and finally, the recommendation is obtained. Results of the Single Slope scheme displays that a combination of user resemblance and trusted data has significantly enriched the predictive accuracy. However, the presence of fraudster users can affect the accuracy of RS.

Kumar *et al.* [31] suggested a User Rating Prediction (URP) scheme to forecast scores for items. The design of the URP algorithm relies heavily on the same users and considers that users who share the same tastes may enjoy similar items. The pre-designed program generates a list of relevant users for

each item and then utilizes this data to predict scores for other items. URP approach is simple to implement, as it does not examine item's features and evades layers that lessen the complication of the system. But, further works are necessary regarding scalability and sparsity issues.

Wang *et al.* [32] designed innovator-CF to recommend cold products. Innovators are an exceptional group of consumers who can find cold products without the support of RS. Therefore, the cold products can be apprehended in the suggestion list by the innovators, thus obtaining the equilibrium between service and accuracy. The innovator-CF scheme is tested both offline and online. The online section can be executed on the mobile devices of the users, allowing the suggestion list to be customized in real-time and completely save the cost of communications. However, the presence of weak ties affects the innovators.

Mao *et al.* [33] designed a standard User-Item-Context model (UIAC) to suggest various summary data and high-order links for creating a multipartite hyper graph. It creates a proportionate hyper graph classification scheme to rank various kinds of items in the hypergraph. An integrated framework is recommended as a guide for executing this multi-goal software design. In the end, it is observed that the UIAC can manage changing and complex needs for e-business. It can be easily expanded by introducing new ends when the volume increases. But, we acknowledge that the present scheme may not be sufficiently scalable since moderate size matrices are required to be analyzed.

Sulthana and Ramasamy [34] set up a Neuro-Fuzzy method using rules to obtain the review context. This method routinely classifies the evaluation under each fuzzy rule. This technique classifies the reviews on the basis of the previous user-generated assessments. Here, 16 fuzzy rules are set to bring information from the reviews, and these reviews are categorized based on the rules. Even though NF seems to progress the accuracy of the RS, the user profile is not considered.

Ahuja *et al.* [35] implemented a movie RS by combining k-Means with k-Nearest Neighbor (KNN). Initially, the panda's module parts the data from screen files. Moreover, it divides user information and movies into a different database using the panda library. Later, the utility matrix determines the users who rated a certain movie. This helps to identify how often movies are rated by the customers. After obtaining the utility matrix, K-means clusters are utilized to construct a variety of basic information that displays the genre of the movie. Lastly, the KNN technique calculates the movie rating using the clustered matrix. Even though the RMSE is low in this method, sentimental analysis can be employed to progress the efficiency of the film RS.

Anitha and Kalaiarasu [36] integrated Supportive Support System (SVM) and Improved Ant Colony Optimization (IACO) to mitigate issues in CF. This SVM-IACO system has two stages. Initially, the SVM classifies the feedbacks into positive and negative comments. Secondly, the recommendation is performed only on the items with positive feedbacks. This increases the effectiveness of RS. However, this system lacks privacy and also needs to be tested in real-time.

Wang *et al.* [37] formulated an Aspect-Based Opinion Mining (ABOM) concept-based to get accurate suggestions. The ABOM model encompasses two portions. In the initial part, a neural network is used to extract important information by classifying people's opinions on various items. In the next part, tensor factorization is employed to predict the average rating. Even though the error is less, the complexity of ABOM is high.

Selvi and Sivasankar [38] proposed a CF-based Adaptive Genetic Neural Network (AGNN) with a modified k-means clustering technique for targeting online users. It separates a set of data into a predefined amount of clusters with the ultimate goal of finding the correlation between the data in clusters. As a user's taste is the same as more than one group, this enhances the recommendation accuracy and decreases the error. Since some users belong to more than one cluster, it results in quicker convergence. However, if a new user joins the system, the recommendation accuracy is reduced due to the lack of demographic data.

Li *et al.* [39] presented a Soft Co-Clustering based CF recommendation model that generates a sparse partition matrix. In the customized order, users can select items named by categories. Here, the users who handpick the same item are considered to have similar interests. Therefore, based on the behavior of the users, the preference matrix is formed by combining the item type matrix with the user-item interaction matrix. Even though this method can address sparsity and scalability issues, the CSP is not completely solved.

Nozari and Koohi [40] introduced a new CF-based recommendation method which is set up to compare participants' experiences with each other based on trust and similarity. Usually, in clusters, there are leaders who are more trustworthy than other members and have a lot of influence on the members. One interesting feature of this method is the blend of similarity measure and Fuzzy C Means (FCM) clustering to determine the participants with similar interests. In FCM, the samples are divided into defined clusters. In addition, FCM can provide samples to more than one group unlike other common techniques like k-means. However, since this FCM technique fails to determine the initial center of the cluster, the accuracy of the recommendation is low.

Ioannidis *et al*. [41] presented a Single and coupled matrix-tensor factorization (CMTF) approach for generating a recommendation. The challenges like the elimination of columns and rows from the correlation matrix are avoided by this method. Therefore, the computational complexity and the sparsity problem are reduced. But, the recommendation accuracy may decrease due to poor clustering. Wei *et al.* [42] presented a CF-based recommendation strategy to address CSP. Here, the neural network is utilized to obtain the item's content feature. They used a machine learning approach to enhance the performance. However, due to the lack of clustering phase, the computational complexity is high. Table 1 provides a comparative analysis of the existing methods.

Table 1: Comparative Analysis

References	**Technique**	**Type of RS**	**CSP**	**Sparsity**	**Advantages**	**Disadvantages**
Kiran *et al.* [23]	DNN	Hybrid	✓	-	Less error.	Predictive accuracy is low.
Natarajan *et al.* [24]	MF-LOD	CF	✓	✓	Eliminates SP and CSP	Computational time. is high
Nilashi *et al.* [25]	SVD+EM+ontology	CF	-	✓	Mitigates scalability and SP	The prediction mechanism is not so accurate.
Kermany and Alizadeh [26]	ANFIS	Hybrid	-	✓	This technique has reduced sparsity issues as well as CSP	The complexity and computational time tends to be high.
Wang *et al.* [27]	DNN	CF	✓	-	Eradicates new user problem	The sparsity issue is not considered.
Herce-Zelaya *et al*. [28]	Random Forest	CF	✓	-	Data obtained from the social site improve the performance	The usage of personal data obtained from social media cause security and privacy issues.
Guo*et al.* [29]	FA-SVD	CF	✓	✓	Reduces the dimensionality	Complexity is high
Jiang L *et al.* [30]	Single Slope scheme	CF	✓	✓	Enriched predictive accuracy.	The presence of fraudster users can affect the accuracy of RS.
Kumar *et al.* [31]	URP	CF	-	-	URP approach is simple to implement, and the complication of the system is less.	Sparsity issues occur.
Wang *et al.* [32]	Innovator-CF	CF	✓	-	Can be customized in real time and completely save the cost of	Presence of weak ties affects the innovators.

					communications.	
Mao *et al.* [33]	UIAC	CF	-	✓	UIAC can manage changing and complex needs for e-business.	Not sufficiently efficient since moderate size matrices are required to be analyzed.
Sulthana and Ramasamy [34]	Neuro-Fuzzy	CBF	-	-	It progress the accuracy of the RS.	The user profile is not considered.
Ahuja *et al.* [35]	KNN	CF	-	-	The RMSE is low.	Sentimental analysis can be employed to progress the accuracy of the film RS.
Anitha and Kalaiarasu [36]	SVM-IACO	CF	-	✓	Increases the effectiveness of RS.	This system lacks privacy and also needs to be tested in real-time.
Wang *et al.* [37]	ABOM	CF	✓	✓	Error is less	Complexity of ABOM is high
Selvi and Sivasankar [38]	k-means clustering	CF	-	-	quicker convergence	recommendation accuracy is reduced due to the lack of demographic data.
Li *et al.* [39]	Soft Co-Clustering	CF	-	✓	Reduces sparsity and scalability issues.	CSP is not completely solved.
Nozari and Koohi [40]	FCM clustering	CF	-	-	It provides samples to more than one group.	It fails to determine the initial center of the clusters.
Ioannidis *et al*. [41]	Coupled graph+ tensor factorization	CF	✓	✓	Computational complexity and the sparsity problem is reduced.	Poor clustering
Wei*et al.* [42]	DNN+SVD	CF	✓	-	Low error.	Computational complexity is high.

3 Proposed Cold Start Aware Hybrid Recommender System

RS is a tool that recommends products like electronics, movies, books, contacts, and other facilities to the users. Consumers are faced with situations where they have a lot of choices to select and need help to narrow down the choices. RS tool helps to break this gap. There are different methods used to build RS models, but the best methods are categorized into two main categories: CB and CF systems. CB system operates based on the product's content and product history that has been previously purchased by the customer. In contrast, CF provides the recommendation by studying the similarity among other users. Both of these RS require important information like features of the product, customer ratings,

etc. So, they cannot provide personalized suggestions with high quality and accuracy. The drawbacks faced by these individual systems are overcome by a novel hybrid mechanism in the proposed RS.

In our proposed SCSHRS methodology as revealed in Fig.1, we make use of the benefits from some methods like SRCF (Sparsity Resolving Collaborative Filtering), SRWCF (Sparsity Resolving Weighted Collaborative Filtering), Ant-Lion (AL) optimization, HOSVD, and ANFIS to eliminate cold start and sparsity problem. The SRCF & SRWCF is used for estimating the unavailable data. The AL is adopted based on the hunting behavior of ant lion larvae. The AL has the capabilities to pull the prey towards itself and consumes. This makes the AL a superior hunter. Therefore the AL can obtain the initial center of K means which is not considered by other existing techniques. The AL optimization creates the data in similar groups by clustering. Then the data is reduced by HOSVD in the dimensionality reduction stage. At last, the ANFIS prediction method compares the output data with the actual data to forecast an exact output. This gives a recommended and similar products to new users.

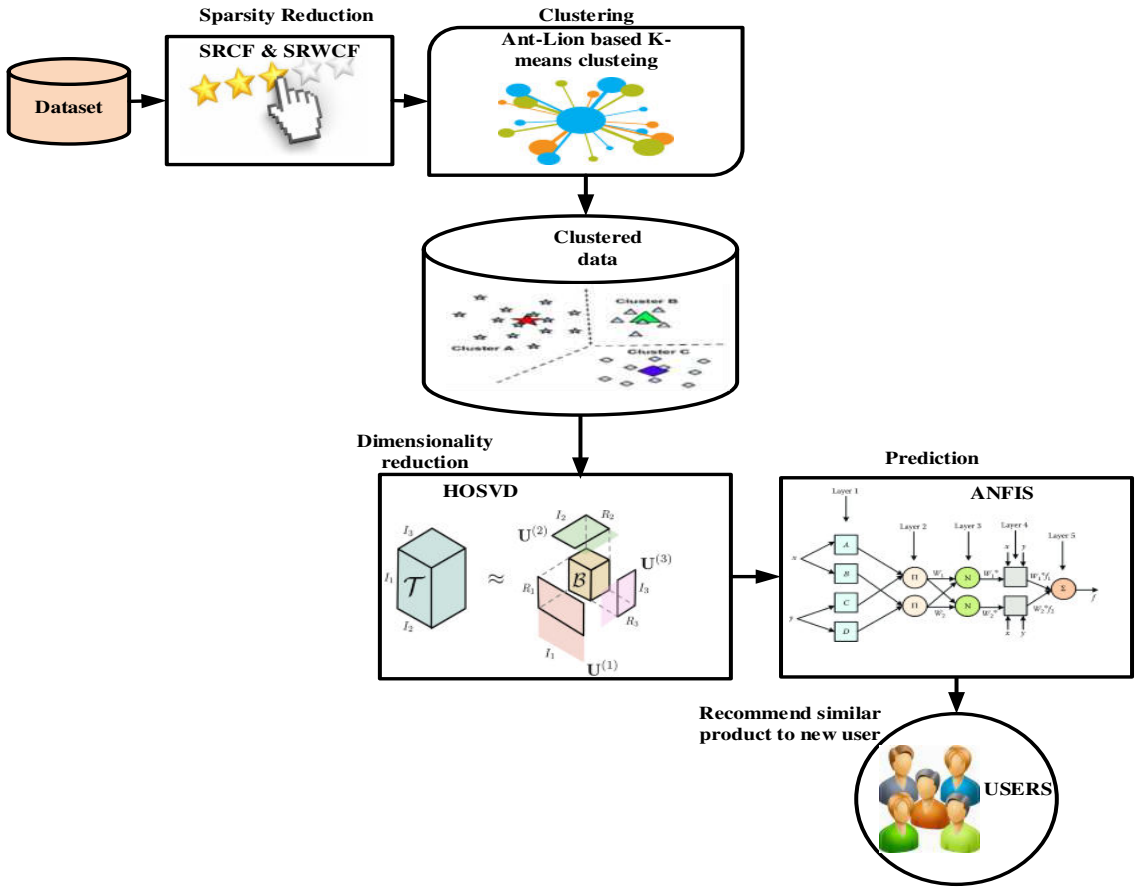


Figure 1: Architecture of the Proposed Sparsity and Cold Start Aware Hybrid Recommender System (SCSHRS)

The proposed SCSHRS method involves four major steps and is explained below.

3.1 Sparsity Reduction

SRCF and SRWCF methods are used before clustering to clear the sparsity issues that occur. It aims to achieve unavailable data (ratings) with minimum computation and consuming time.

The recommendation accuracy decreases for the sparse datasets. Therefore, we employ SRCR and SRWCF method to determine the Unavailable Ratings (UR). The sparsity range in the dataset can be found out by using Eqn. (1).

$$S_R = 1 - \frac{E_R}{E_T} \tag{1}$$

Where E_R denotes the number of rated entities and E_T denotes the total number of available entities.

For describing the neighbor users, the similarity among users is measured, which can be used for the prediction method. With respect to this, the prediction rating of user m for item k is determined in Eqn. (2).

$$PR_{(l,k)} = \overline{UR_l} + \frac{\sum_{m \in N} Sim(l,m).(UR_{m,k} - \overline{UR_m})}{\sum_{m \in N} S(l,m)} \tag{2}$$

The $\overline{UR_l}, \overline{UR_m}$ gives the mean value of ratings stated by user *l* and m. UR_{mk} is the rating of k item of every similar user m. The similarity between two user Sim(*l*,m) is calculated by using Pearson's correlation coefficient (PCC), which compares ratings of all items rated by the neighbor and the target user. The PCC between user *l* and m is given in Eqn. (3),

$$Sim(l,m) = \frac{\sum_{k \in K} (UR_{l,k} - \overline{UR_l})(UR_{m,K} - \overline{UR_m})}{\sqrt{\sum_{k \in K} (UR_{l,k} - \overline{UR_l})^2} \sqrt{\sum_{k \in K} (UR_{m,K} - \overline{UR_m})^2}} \tag{3}$$

Where K signifies the set of the items, in which $K = \{k_1, \ldots k_a\}$. The range of metric is in between 1(similarity) and -1 (dissimilarity). A negative value decreases prediction accuracy, so it is rejected.

The UR of user *l* for the product k in the SRCR method is determined by using the Eqn. (4.)

$$UR_{l,k} = \overline{UR_l} + \frac{\sum_{m \in T} (UR_{m,k} - \overline{UR}_m)}{|T|} \tag{4}$$

Where $\overline{UR}_m$ signifies the average rating of user l, $UR_{m,k}$ is user m rating on item k, $|T|$ is the number of users rated item k.

In the SRWCF method, the UR of user m for the product k is calculated by using the Eqn. (5).

$$UR_{l,k} = \overline{UR_l} + \frac{\sum_{m \in T} X_m * (UR_{m,k} - \overline{UR_m})}{\sum_{m \in T} X_m} \quad (5)$$

Where $\overline{UR}_l$ signifies the average rating of user l, $UR_{m,k}$ is user m rating on item k and X_m is estimated by using the Eqn. (6).

The rate estimating process is simple by utilizing X_m, which can find the more rated items.

$$X_m = \frac{I_{R,m}}{I_T} \quad (6)$$

$I_{R,m}$ is the number of user m-rated items, I_T are the total items.

This method takes less computation time and it can be run offline. Therefore, the sparsity problem can be solved without reducing the dimension of the data set.

3.2 Ant-lion Based k-means Clustering

Clustering is used to cluster the users on the basis of similarity in ratings. Therefore, for clustering the similar users, Ant-Lion-based K-means algorithm is used. At this stage,k-means clusters the users by finding the distance between two data and the Ant-Lion algorithm helps to determine the initial center of k-means.

3.2.1k-means Algorithm

The unsupervised k-mean algorithm is a simple technique that can be used for clustering. This basically finds the local optimal solution by iteration process. Here, k-means is employed to determine the best cluster centers for each cluster.

Step 1: Initialize k number of cluster centers randomly.

Step 2: Assign the data or object to the nearest cluster center.

Step 3: Update the cluster center by computing the mean value.

Step 4: Terminate when the cluster center stops moving further or when it touches the maximum number of iterations.

3.2.2Ant lion optimization

ALO algorithm works based on the hunting characteristics of the ant-lion larvae. It makes a hole (conical shaped) in the sand using its jaw. The hole act as a trap to hunt the prey (initial cluster centers). In this research, the optimal cluster center is considered as prey. Therefore, the main objective is to hunt the optimal cluster centers. Once the trap is set up, the larvae hide inside the sand and wait for the prey to appear. As soon as the prey gets trapped, the ant lion catches and drags the prey into the hole. Finally, it consumes the prey using its jaws. Therefore for clustering similar users, Ant Lion's hunting behavior is adapted to obtain the initial value of cluster center. This enhances the recommendation accuracy. The ALO method includes five stages to complete its action:(i) Random walk (ii) making traps (iii) trapping ants in hole (iv) hunting prey (v) reconstructing trap.

Search space of ant's walking movement can be determined by using Eqn. (7).

$$Y(p) = [0, cum_{sum}(2rand(p_1) - 1), cum_{sum}(2rand(p_2) - 1), ..., cum_{sum}(2rand(p_n) - 1)] \quad (7)$$

Where cum_{sum} represents the cumulative set, p signifies the steps in random walks, n is maximum repetitions and rand(p) is the random function, which is created using Eqn. (8).

$$rand(p) = \begin{cases} 1 & R > 0.5 \\ 0 & R \le 0.5 \end{cases} \quad (8)$$

Where R is the random value, which is distributed in uniform in [0,1], p denotes the repetition in the random walk stage.

The positions of ants are stored in the matrix, MX_{ant} and given in Eqn. (9).

$$MX_{ant} = \begin{bmatrix} L_{1,1} & L_{1,2} \ldots\ldots\ldots L_{1,e} \\ L_{2,1} & L_{2,2} \ldots\ldots\ldots L_{2,e} \\ \ldots & \ldots \quad \ldots \quad \ldots \\ L_{n,1} & L_{n,2} \ldots\ldots\ldots L_{n,e} \end{bmatrix} \quad (9)$$

Where MX_{ant} is the matrix for saving the location of every ant. $L_{i,j}$ denotes the j-th variable (dimension) of i-th ant, n denotes the number of ants, e signifies the variable number.

The fitness function can be determined by using Eqn. (10).

$$F = \sum_{i=1}^{k} \sum_{ob_e \in ce_i} E_{dis}(ob, ce_i) \tag{10}$$

Where $E_{dis}(ob, ce_i)$ represents the Euclidean distance between object e and i-th cluster center ce_i. Allocate each object closer to the center. In our work, the objective function is to minimize fitness function. The minimum fitness value is chosen for clustering similar users.

For calculating every ant, a fitness function is used in optimization, the fitness of each ant is stored in the matrix, M_{antF} and given in Eqn. (11)

$$M_{antF} = \begin{bmatrix} F([L_{1,1}, L_{1,2.} \ldots\ldots\ldots L_{1,e}]) \\ F([L_{2,1}, L_{2,2} \ldots\ldots\ldots L_{2,e}]) \\ \vdots \\ F([L_{n,1}, L_{n,2} \ldots\ldots\ldots L_{n,e}]) \end{bmatrix} \tag{11}$$

F denotes an objective function, L_{ij} denotes the j-th dimension value of the i-th ant, n represents the number of ants.

It is assumed that the ant lions are hidden in search space (SS). Hence, the position of the ant lions is stored in the matrix, $MX_{ant}lion$ and given in Eqn. (12).

$$MX_{ant}lion = \begin{bmatrix} LB_{1,1} & LB_{1,2} \ldots\ldots\ldots LB_{1,e} \\ LB_{2,1} & LB_{2,2} \ldots\ldots\ldots LB_{2,e} \\ \ldots & \ldots \quad \ldots \quad \ldots \\ LB_{n,1} & LB_{n,2} \ldots\ldots\ldots LB_{n,e} \end{bmatrix} \tag{12}$$

$LB_{i,j}$ denotes the jth variable of ithAL, n signifies the number of ant lions, e represents the variable number (dimension).

The fitness of each AL is stored in matrix $M_{antlionF}$ and is given in Eqn. (13).

$$M_{antlionF} = \begin{bmatrix} F\left(\left[LB_{1,1}, LB_{1,2} \ldots\ldots\ldots LB_{1,e}\right]\right) \\ F\left(\left[LB_{2,1}, LB_{2,2} \ldots\ldots\ldots LB_{2,e}\right]\right) \\ \vdots \\ F\left(\left[LB_{n,1}, LB_{n,2} \ldots\ldots\ldots LB_{n,e}\right]\right) \end{bmatrix} \quad (13)$$

Where F denotes an objective function, L_{ij} denotes the j[th] variable of the i-thAL, n represents the number of ant lion.

(i) Random Walk

Update the position of ants at every step of optimization with a random walk. Then, there will be a range of variables for every search space. Then Eqn. (7) cannot be directly employed for updating the ant's location. For keeping the Random Walk (RW) in the boundary of SS, RW is controlled using Eqn. (14).

$$Y_j^p = \frac{(Y_j^p - s_p) \times (u_p - t_j^p)}{u_j^p - s_j} + t_j \quad (14)$$

Where s_p and u_p denotes the least and highest RW of p[th] variable, t_j^p and u_j^p signifies the iteration at j of minimum and maximum of p[th] variable.

(ii) Making traps and trapping ants in a hole

The trapped ant's slipping and escaping action denote the behavior of ants. The radius of the hyper-sphere of RW is decreased. This behavior is represented mathematically in Eqn. (15) and (16).

$$t^p = t^p / I \quad (15)$$

$$u^p = u^p / I \quad (16)$$

Where t^p signifies the least value for every variable at p-th iteration, u^p denotes the highest value for every variable, I denote a ratio and is determined using Eqn. (17).

$$I = 10^w \frac{p}{P}, \quad (17)$$

Where p signifies the present iteration, P denotes the highest number of iteration, w is a constant value.

(iii) Hunting prey and reconstructing trap

The last step of the hunt is when an ant is in the lowermost pits and the AL catches the ant with its jaw. After this stage, the AL pulls the ant into the sand and eats its body. An ant lion needs to update its latest position of hunted prey for improving its chance to catch its prey. The location of the chosen j-th ant lion at p-th iteration $An_{Li}{}_j^p$ is given in Eqn. (18).

$$An_{Li}{}_j^p = Ant_i^j \qquad if\ F(Ant_i^j) > F(An_{Li}{}_j^p) \tag{18}$$

Where p denotes the present iteration, Ant_i^j denotes the location of i-th and j-th iteration.

(iv) Elitism

It is the main stage in the evolutionary algorithm. It permits us to find the best solution found at any step of the optimization method. In this method, the best AL attained is saved at each iteration and the best solution is named as elite. The fittest ant lion is elite. During iteration, it should be able to influence the movements of all ants. Hence, it is assumed that every ant randomly bypasses the designated AL using the roulette wheel and elite at the same time. The location of i-th ant at p-th iteration, Ant_i^p is given in Eqn. (19).

$$Ant_i^p = \frac{RN_W^p + RN_E^p}{2} \tag{19}$$

Where RN_W^p is is the RW around the AL chosen by the roulette wheel at p-th iteration and RN_E^p is the RW around the elite at p-th iteration. Fig.2 provides the flowchart for the ALO technique.

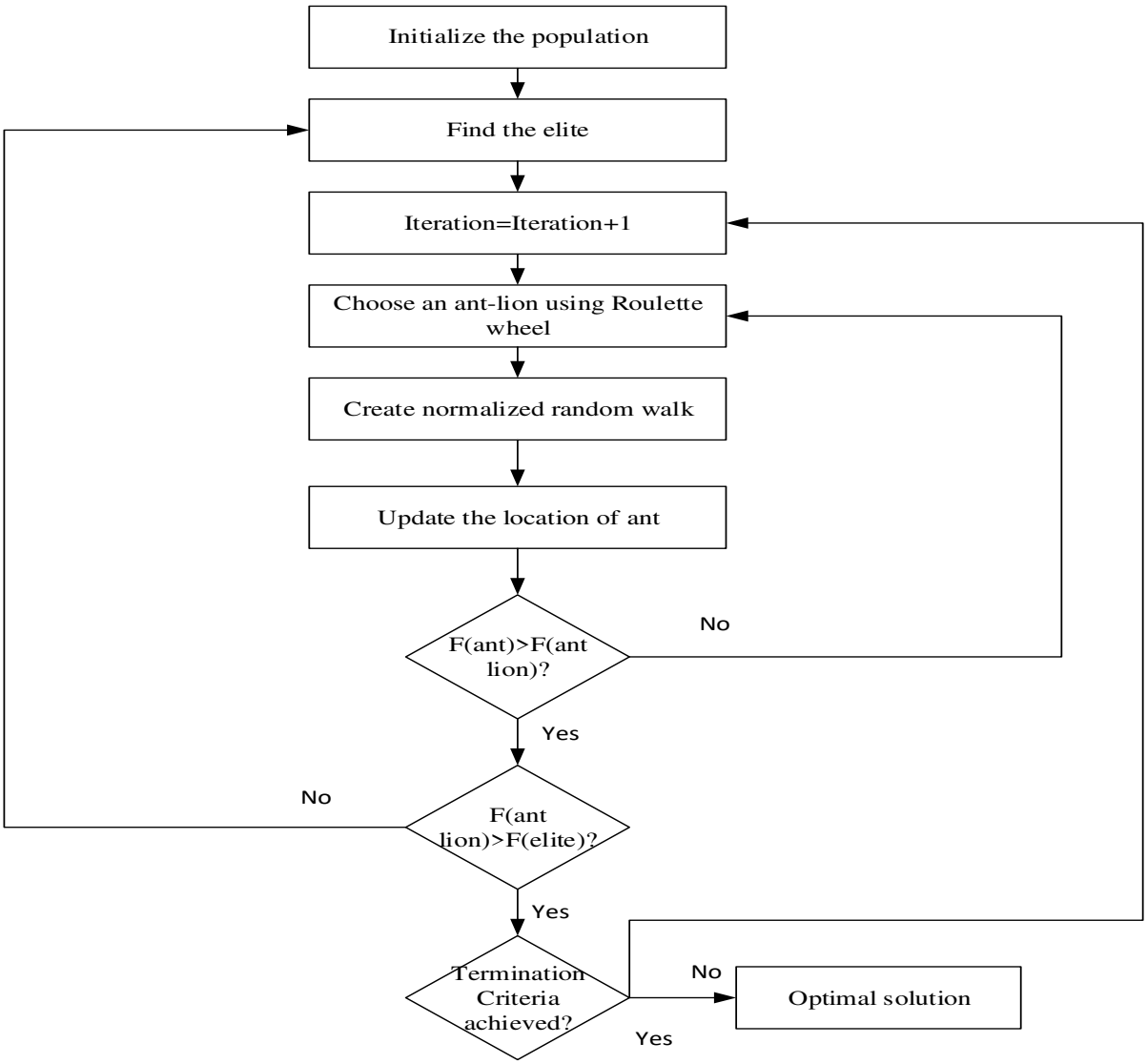


Figure 2: Flow chart of Ant lion optimization

3.3 Dimensionality reduction

Dimensionality reduction is the process in which the bigger dataset is dimensionally reduced into a lower dimension. To accomplish this, we employ by HOSVD technique. The benefits of HOSVD include matching, visualization, and reducing the computational or processing time. Hence, HOSVD is considered an efficient scheme for tensor decomposition.

By using HOSVD, the tensor $T(O \times W \times Q)$ as shown in Fig.3 is decomposed into a matrix form $UR \in K^{O \times O}, PR \in K^{W \times W}, CT \in K^{Q \times Q}, S \in K^{O \times W \times Q}$.

Here, S, PR, CT, UR, and denote the central tensor, products, context, and users respectively.

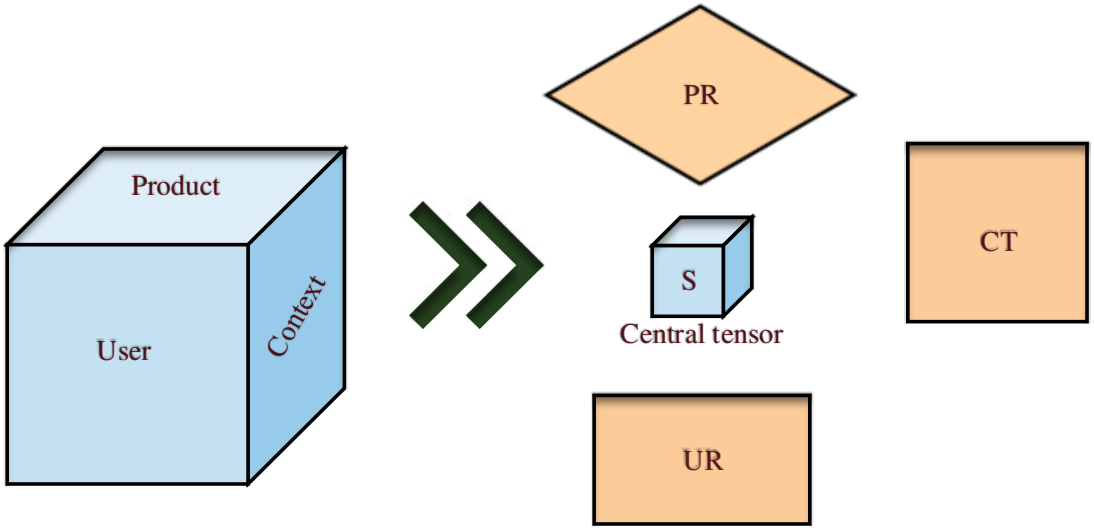


Figure 3: Structure of HOSVD

3.4Prediction by ANFIS

Neural Networks and Fuzzy Logic are useful in resolving various real-world problems. Yet, both schemes also have restrictions that prevent them from obtaining effective solutions. In fuzzy logic, it is often challenging to define the membership functions and the right set of rules. In addition, tuning a solution is more complex and takes a longer time. On the other hand, NNs are imperfect and hard to understand than the fuzzy system. A proper blend of these two techniques which is termed a Neuro-Fuzzy system can solve the problems of fuzzy logic and neural networks effectively. The acquisition of knowledge by manual power can be replaced by the Neuro-Fuzzy system. Therefore, this mechanism does not depend on the individual experts rather, the parameters are determined by minimizing the error through a training process.

The ANFIS method as depicted in Fig. 4 introduces a Fuzzy Logic and Neural Network system. The Fuzzy Logic is used for making a proper decision, and Neural Networks have learning abilities that help to address real-world difficulties. In this ANFIS system, training is executed with the "MovieLens" dataset to predict the result that is closer to the authentic output. Here, the ANFIS utilizes the rule base taken by the Fuzzy Logic to perform the forecasting task. The rules are produced by training the system to obtain exact predictions. In this, the prediction mechanism works by two fuzzy rules:

Rule 1: If the demographic data of the new user is similar or matches the user 'y', then the products adored or bought by the user 'y' will be recommended to the new user.

Rule 2: If the demographic data of the new user do not match the user 'y', then the products adored or bought by the user 'y' will not be recommended to the new user.

In this way, the similarity between the new user and all other customers in the database will be found out and a final prediction is made based on the rules.

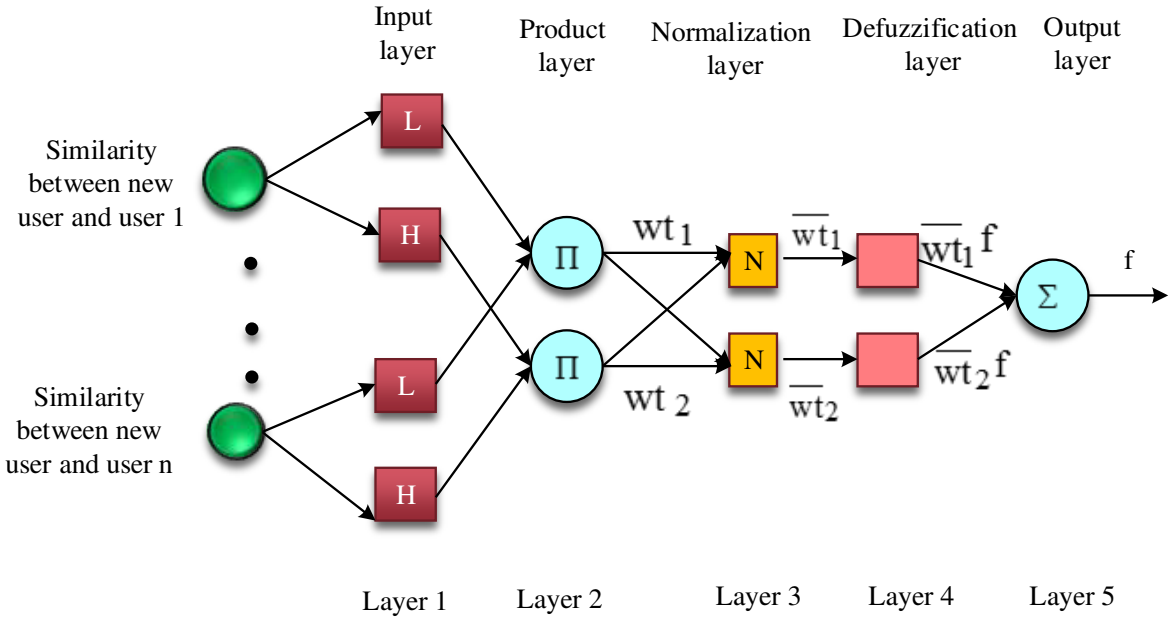


Figure 4: ANFIS based Prediction Mechanism

Layer 1: Fuzzification is done for the variables (given atthe input). Here, the membership grades are created at the output of this layer as given in Eqn. (20).

$$O = \mu G_m^T(x), \quad m = 1,2 \text{ and } T = 1,2...n^{th} user \tag{20}$$

Where $\mu G_m^t(x)$ signifies the Membership Function (MF) in which, m=1,2 signifies low similarity and high similarity respectively and T signifies the resemblance score between the new user& the T^{th} user.

$$\mu G_m^T(x) = \frac{1}{1+\left[\left(\frac{x-L_m}{J_m}\right)\right]^{K_m}} \tag{21}$$

Where J_m, K_m, and L_m are considered as parameter sets that can alter the form of the MF.

Layer 2: It pre-defined quantity of nodes and the output is computed by Eqn. (22).

$$Wg_m = G_m^1(x) \times G_m^2(x) .... \times G_m^n(x) \tag{22}$$

Layer 3: This layer computes the Normalized Firing Strength (NFS) and the output of the third layer is defined in Eqn. (23).

$$\overline{Wg}_m = \frac{Wg_m}{Wg_1 + Wg_2} \tag{23}$$

Layer 4: The fourth layer determines the consequent rule parameters. Each node in this layer produces an output by multiplying normalized firing strength with a polynomial as given in Eqn. (24).

$$\overline{Wg}_m \mathrm{f}_i = (R_m + S_m + U_m)\overline{Wg_m} \tag{24}$$

Where R_m, S_m, and U_m are the parameter set and $\overline{Wg}_m$ signifies the output of layer 3.

Layer 5: This layer computes the final output by Eqn. (25) and the flowchart for SCSHRS is provided in Fig. 5.

$$\sum_m \overline{Wg}_m \mathrm{f}_i = \frac{\sum_i Wg_m f_i}{\sum_i Wg_m} \tag{25}$$

Fig.5 gives the flowchart for the proposed recommender system.

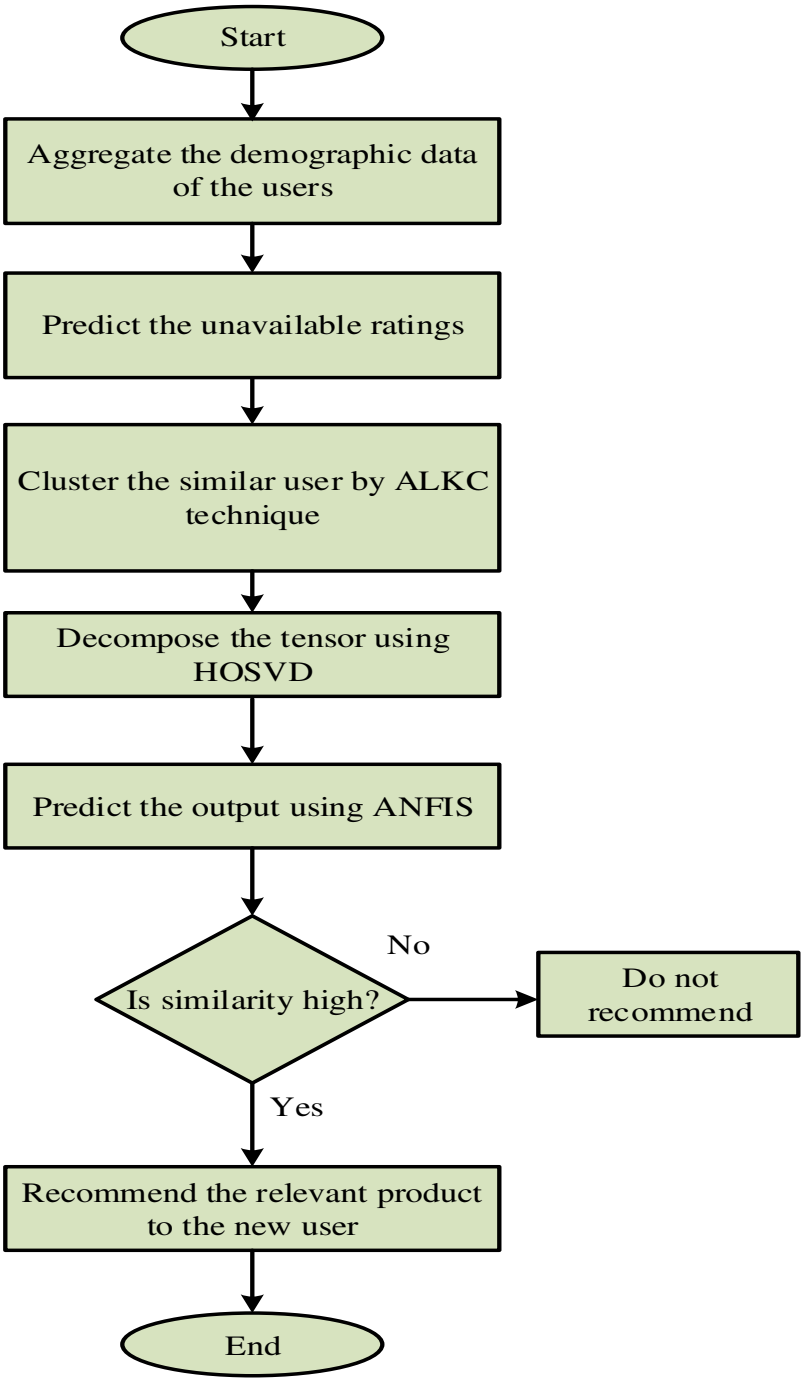


Figure 5: Flowchart for the proposed SCSHRS recommender system

3.5 Example for the proposed SCSHRS system

The first ten users from the Movie Lens dataset are chosen to illustrate the proposed method and is provided in Table 2. The dataset includes demographic details such as gender, age, and occupation of the various users.

Table 2: User profile

Users	Gender	Age	Occupation
U1	M	24	Technician
U2	F	53	Other
U3	M	23	Writer
U4	M	24	Technician
U5	F	33	Other
U6	M	42	Executive
U7	M	57	Administrator
U8	M	36	Administrator
U9	M	29	Student
U10	M	53	Lawyer

Based on the demographic data provided in Table 2, the similarity among the users is determined by calculating the distance between them. Based on this calculation, the similarity score obtained for the first 10 users in the Movie-Lens dataset is provided in Table 3.

Table 3: Similarity Score

User	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
U1	0	36.39	12.33	0	26.95	18.90	16.12	17.75	19.40	11.32
U2	36.39	0	27.79	34.34	4	33.06	20.40	24.15	21.03	13.88
U3	12.33	27.79	0	13.37	24.65	24.65	14.06	17.05	14.66	16.82
U4	0	34.34	13.37	0	32.60	11.73	19.88	15.41	16.92	19.59
U5	26.95	4	24.65	32.60	0	24.68	25.39	38.67	30.81	28.04
U6	18.90	33.06	24.65	11.73	24.68	0	14.40	17.99	14.42	15.00
U7	16.12	20.40	14.06	19.88	25.39	14.40	0	3	16.07	12.33
U8	17.75	24.15	17.05	15.41	38.67	17.99	3	0	16.28	17.13
U9	19.40	21.03	14.66	16.92	30.81	14.42	16.07	16.28	0	10.11
U10	11.32	13.88	16.82	19.59	28.04	15.00	12.33	17.13	10.11	0

After finding the similarity score, the users who are more similar to each other are clustered using the ALO based k-means technique. For example: From Table 3, user 1 and user 4 have a difference of '0' which indicates that they are very much similar. Since user 1 and user 4 are similar to each other they will be clustered together. In this way, all the similar users are clustered together.

After clustering, the tensor is decomposed into lower dimensions by using the HOSVD technique. This helps to lessen the computational time. Finally, the ANFIS system recommends the product to the new users based on machine learning and fuzzy rule system. According to Rule 1 and Rule 2 of ANFIS, if the demographic data of the new user is similar or matches the user 'y', then the products purchased or rated by the user 'y' will be recommended to the new user. Else, the item will not be recommended. For example, let us consider user 1 as a new user. According to the similarity score, since the new user (i.e., user 1) has a profile similar to user 4, the items purchased or liked by user 4 will be recommended to the new user.

4. Experimental Setup and Result Analysis

The proposed recommender system is simulated in the MatLab platform. To make a fair comparison, the proposed and the baseline techniques are evaluated using three similar datasets (MovieLens-20M, Last.FM, and Book-Crossing) with similar testing and training data i.e., 80% is utilized for training and the remaining 20% is used for testing.

4.1 Datasets

The datasets used for evaluation are MovieLens-20M, Last.FM, and Book-Crossing.

- **MovieLens-20M:** This dataset is collected from Movie-Lens website (https://grouplens.org/datasets/movielens/20m/). It has around 20 million movie ratings ranging from 1 to 5.
- **Last.FM:** It is an online music system and has music artist listening data collected (https://grouplens.org/datasets/hetrec-2011/).
- **Book-Crossing:** This dataset has 1 million ratings of books collected from book-crossing community. The ratings in this dataset range from 0 to 10 (http://www2.informatik.uni-freiburg.de/~cziegler/BX/). The statistical data of these datasets are provided in Table 4.

Table 4: Statistics of the dataset

Statistics	MovieLens-20M	Last.FM	Book-Crossing
Items	27,278	17,632	271,379
Users	138,000	1,892	278,858
Ratings	20 million	42,346	1,149,780

These datasets contain the demographic data of numerous users. It contains the user's ID, age, and occupation. Occupation includes administrator, artist, doctor, lawyer, and technician.

4.2 Evaluation Metrics

The metrics such as MAPE, NDCG, precision, recall, accuracy, and F-measure are used to evaluate the proposed SCSHRS system. The descriptions of these metrics with their mathematical equations are explained below.

(i) NDCG: The quality of the recommendation can be determined by using Eqn. (26).

$$\mathrm{NDCG}@\mathrm{L} = \frac{1}{\mathrm{IDCG}} \times \sum_{q=1}^{L} \frac{2^{rev_q} - 1}{\log_2(q+1)} \tag{26}$$

$$\text{Where IDCG @ L} = \sum_{q=1}^{|REV|} \frac{2^{rev_q} - 1}{\log_2(q+1)} \quad (27)$$

Where rev_q signifies the relevance of an item at position q for a particular user.

(ii) Precision: This measure signifies the exactness of the operation i.e. it evaluates whether the generated recommendations are relevant to the new users.

$$\text{Precision} = \frac{Rc_{rev}}{Tot_{rec}} \quad (28)$$

Where Rc_{rev} symbolizes the recommendations that are relevant to the new user, Tot_{rec} signifies the total quantity of recommended items.

(iii) Recall: It is the measurement of the sum of the recommendations that are relevant to the new user to authentic quantity recommendations that are actually relevant.

$$\text{Recall} = \frac{Rc_{rev}}{Act_{rev}} \quad (29)$$

Where Rc_{rev} symbolizes the recommendations that are relevant to the new user and Act_{rev} represents the actual amount of relevant suggestions or recommendations.

(iv) F-measure: It reveals the accuracy of the experiment on the basis of precision and recall measure and is determined by Eqn. (30).

$$\text{F-measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{precision} + \text{recall}} \quad (30)$$

(v) MAPE: It determines the percentage of deviation from the actual value and is computed using Eqn. (31).

$$MAPE = \frac{100}{z} \sum_{j=1}^{2} \frac{b_j - \hat{b}_j}{b_j}$$

(vi) Accuracy: The accuracy of recommendation is determined by Eqn. (31).

$$Acc = \frac{TN + TP}{FN + FP + TP + TN}$$

Where TP and TN signify true positive and true negative respectively whereas, FP and FN signify false positive and false negative respectively.

4.3 Baselines

Clustering and dimensionality reduction are the most commonly used method in the recommendation system for avoiding data sparsity and cold start issues. Our proposed method uses both of these techniques at different stages to enhance the recommendation accuracy. To make a fair comparison, the proposed method is compared with some of the existing dimensionality reduction and clustering techniques as follows:

(i) Dimensionality reduction techniques

MF+LOD [24] are a matrix factorization model developed to provide personalized recommendations in social media and electronic business. This technique avoids data sparsity and CSP. SVD+EM+Ontology [25] is a CF method that finds the relevancy among the users and items by reducing the dimension of the data. FA-SVD [29] is a CF-based matrix decomposition model that alleviates CSP by using the historical rating matrix and the attribute data of the items. Coupled graph+ tensor factorization [41] model provides the correlation from multiple repositories in higher-order tensor and graph. This model leverages CSP and data sparsity issues.

(ii) Clustering techniques

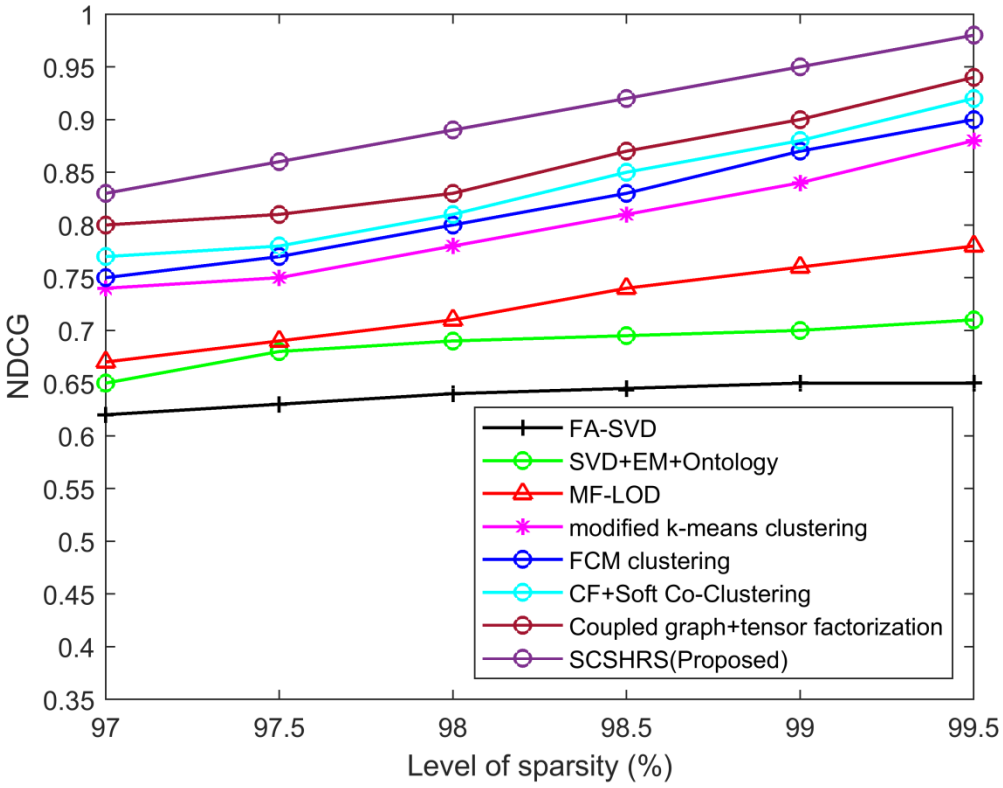
CF+k-means clustering [38] is a k-mean clustering-based recommender system developed for targeting online users. Here, the data points in the cluster are located by a using neural network. CF+Soft Co-Clustering [39] is a recommendation model that generates a sparse partition matrix. This model addresses sparsity and scalability issues. Fuzzy C- means clustering [40] model finds the users with similar interests by finding the similarity measure. This is one of the widely used clustering techniques in RS.

4.4 Comparative Analysis

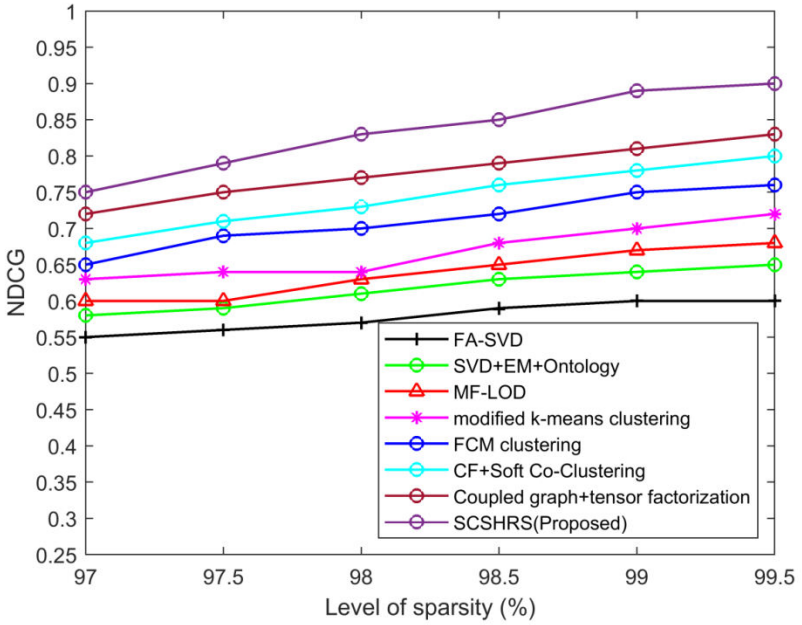
The proposed SCSHRS is implemented in the MovieLens-20M, LastFM, andBook-Crossing datasets and it is compared with the MF-LOD [24], SVD+EM+Ontology [25],FA-SVD [29], Coupled graph+tensor factorization [41], modified k-means clustering [38], CF+Soft Co-Clustering [39], and FCM clustering [40] techniques.

4.4.1NDCG analysis

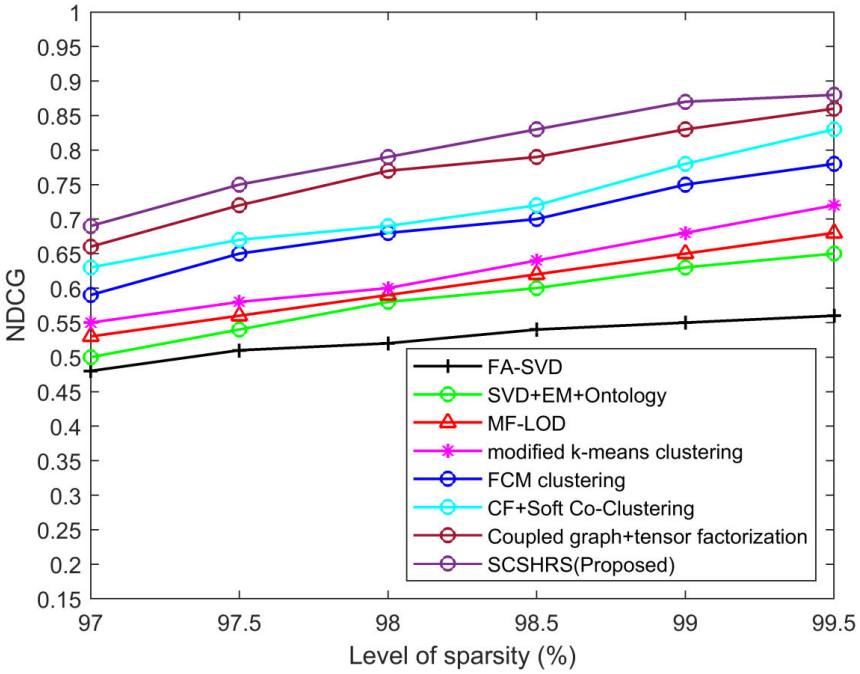
Fig.6 illustrates the sparsity level analysis which is obtained by computing NDCG in various datasets such as MovieLens-20M, Last.FM, and Book-Crossing datasets. NDCG of our proposed SCSHRS method is compared with baseline methods such as MF-LOD [24], SVD+EM+Ontology [25], FA-SVD [29], Coupled graph+tensor factorization [41] techniques, modified k-means clustering [38], CF+Soft Co-Clustering [39], and FCM clustering [40]. From the results, it is perceived that the proposed SCSHRS method has obtained better NDCG results than the existing techniques for varying sparsity levels ranging from 97% to 99.55.This is because, SRCF & SRWCF helps to levitate thesparsity problem and thus the NDCG is enhanced.



(a)



(b)

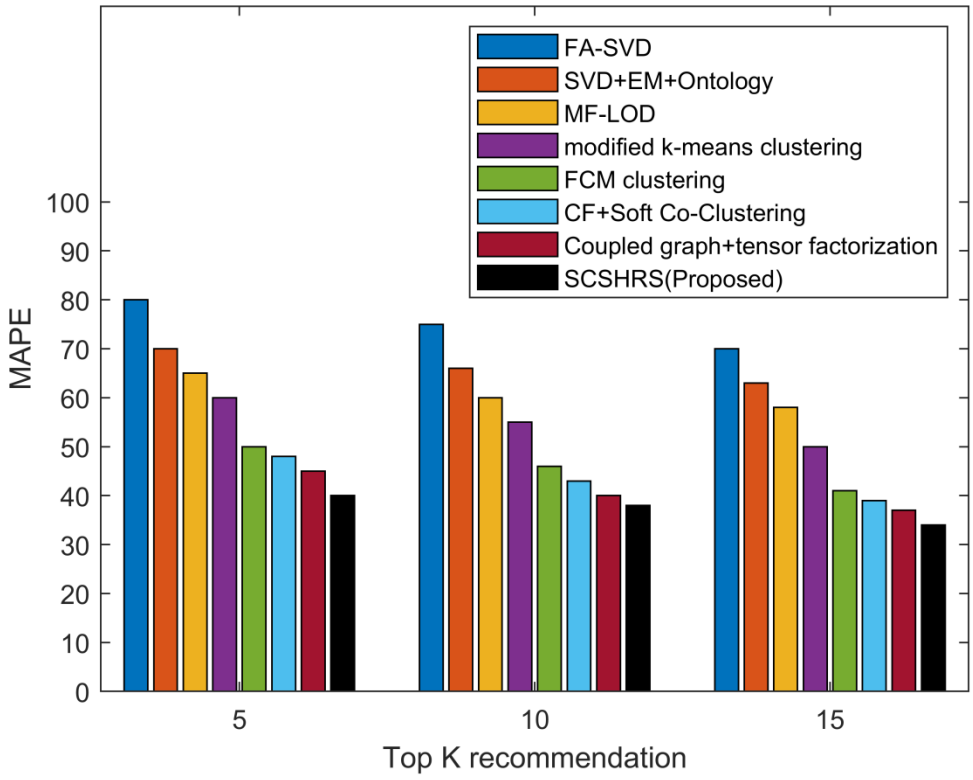


(c)

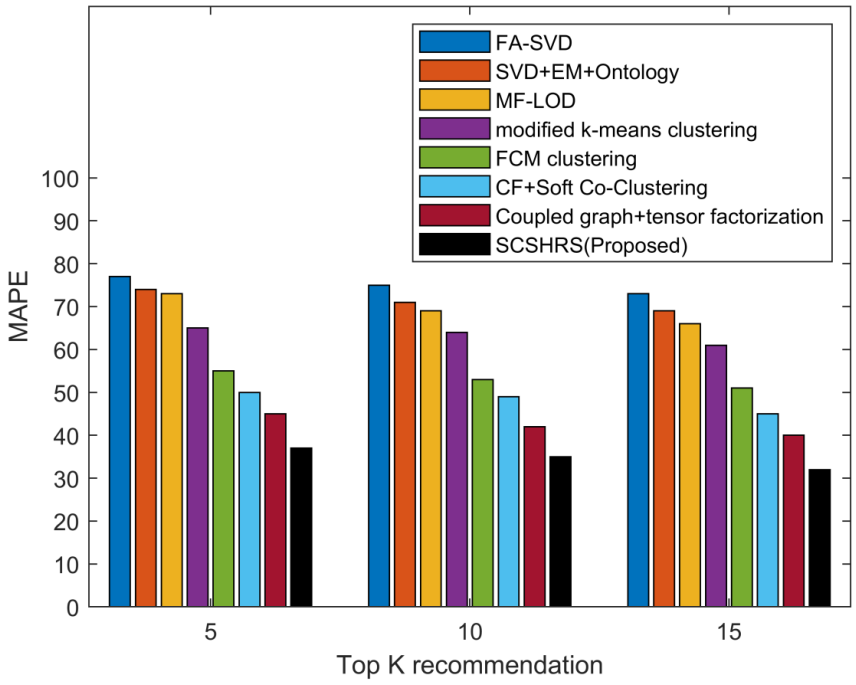
Figure 6: NDCG with different sparsity level for various datasets(a) MovieLens-20M (b) Last.FM (c) Book-Crossing

4.4.2 MAPE analysis

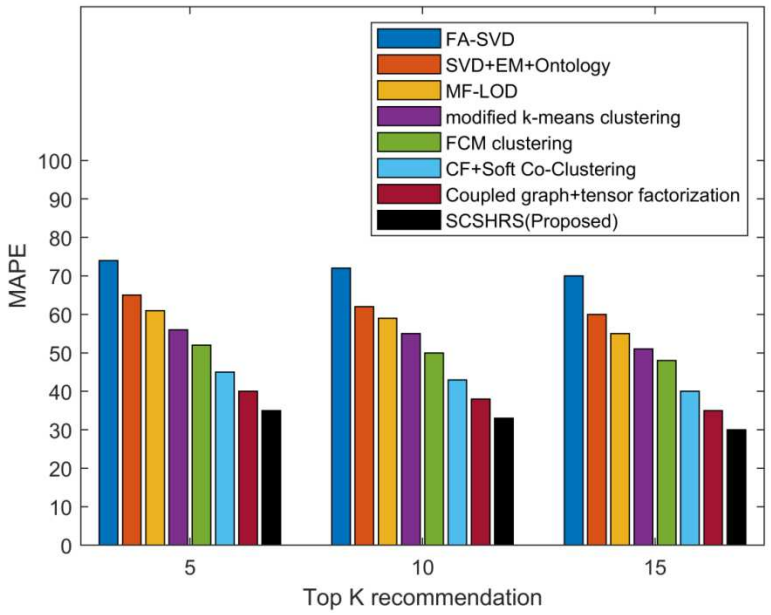
MAPE obtained for the proposed SCSHRS system is compared with some recent prevailing techniques in Fig. 7. From this measure, the MAPE measure recorded for the top 5, 10, and 15 recommendations for all the three datasets are very high for the baseline methods whereas, the proposed SCSHRS technique has produced less MAPE when tested in all the three datasets such asMovieLens-20M (40%), Last.FM(36%), and Book-Crossing(35%). This is because of the efficient clustering of similar users by the ALO-based k means clustering and HOSVD based dimensionality reduction technique employed at different stages of the recommender system. Due to this, the error is significantly reduced in our proposed SCSHRS method.



(a)



(b)

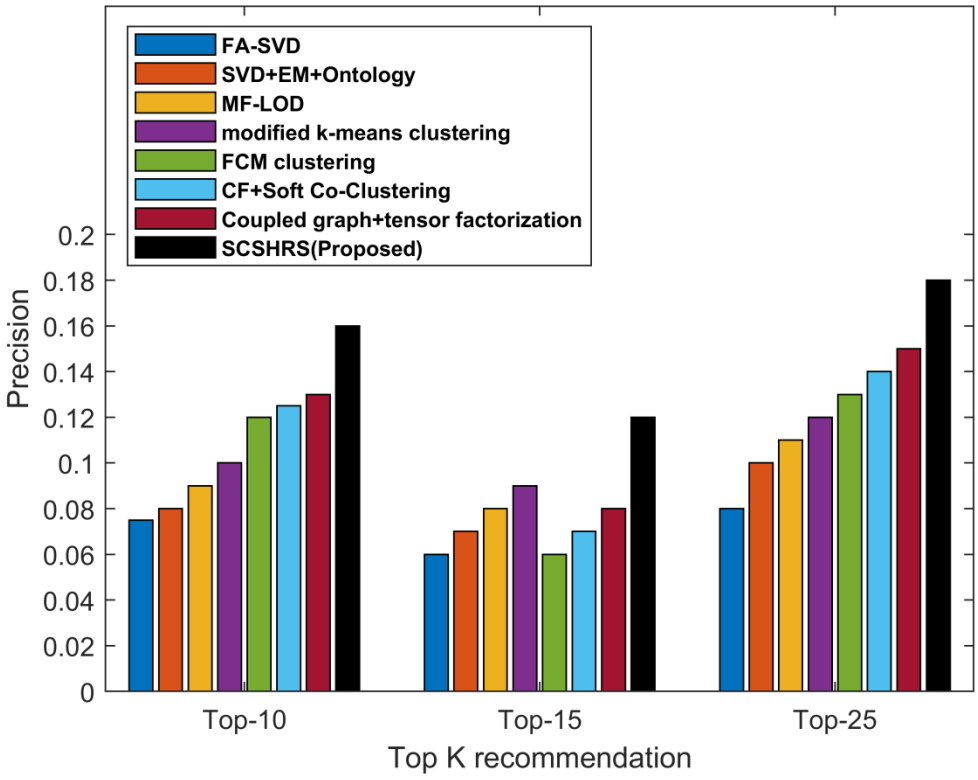


(c)

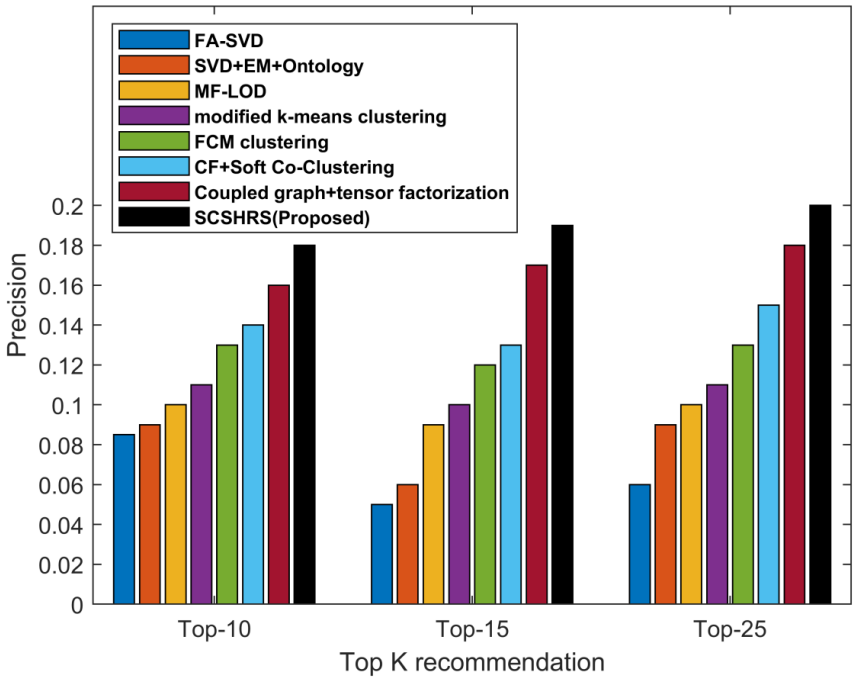
Figure 7: MAPE analysis for various datasets (a) MovieLens-20M (b) Last.FM(c) Book-Crossing

4.4.3Precision Analysis

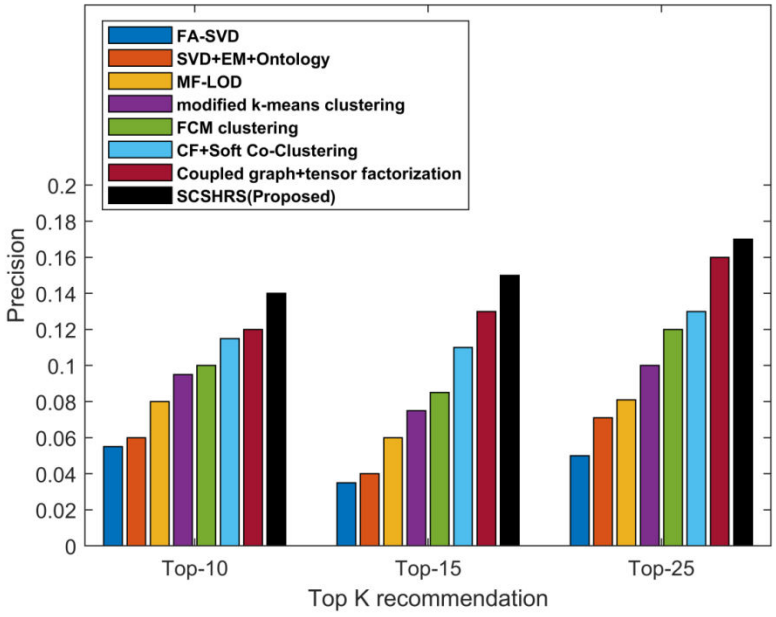
Fig.8 depicts the precision analysis conducted for the proposed method which is evaluated in three datasets that include MovieLens-20M, Last.FM, and Book-Crossing datasets. When testing the proposed method is MovieLens-20M for top 10 recommendation, it is observed that the proposed SCSHRS system gives higher precision (0.16) than the baseline techniques such as Coupled graph+tensor factorization (0.13), SVD+EM+Ontology (0.08), MF-LOD(0.09), k-means clustering(0.1), FCM clustering(0.11), CF+Soft Co-Clustering(0.12), FA-SVD(0.07) and the proposed SCSHRS. Similarly, the proposed method has also produced better precision values when tested in Last.FM and Book-Crossing datasets for top 15, and top 25 recommendations. This is because, the prediction of missed ratings by SRCF, SRWCF schemes, and accurate prediction by the ANFIS system gives precise results.



(a)



(b)

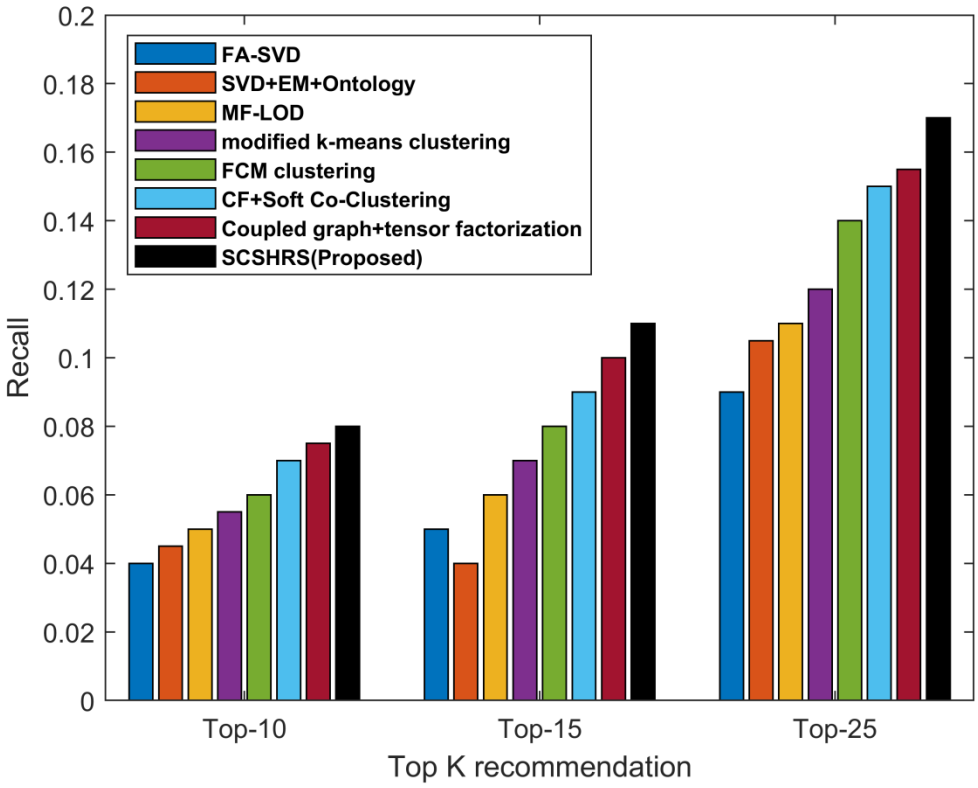


(c)

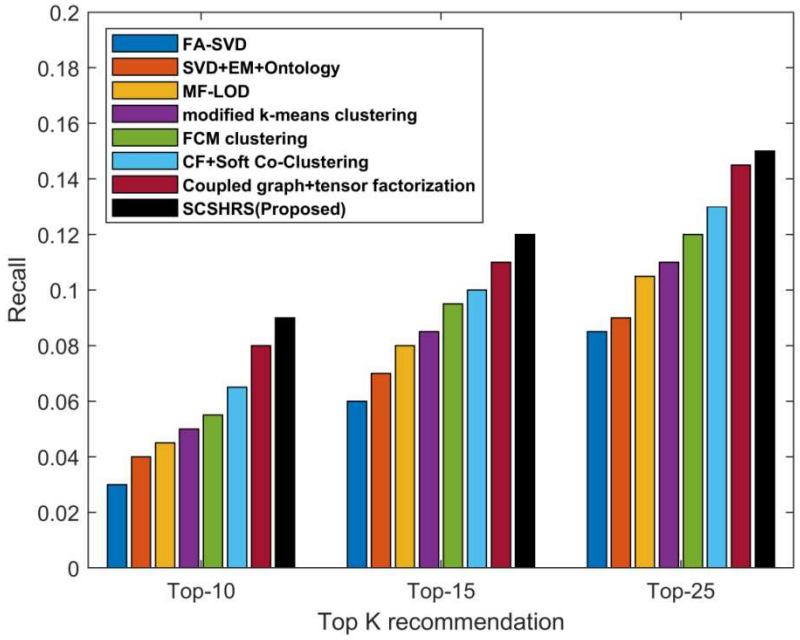
Figure 8: Precision analysis for various datasets (a) MovieLens-20M (b) Last.FM (c) Book-Crossing

4.4.4 Recall analysis

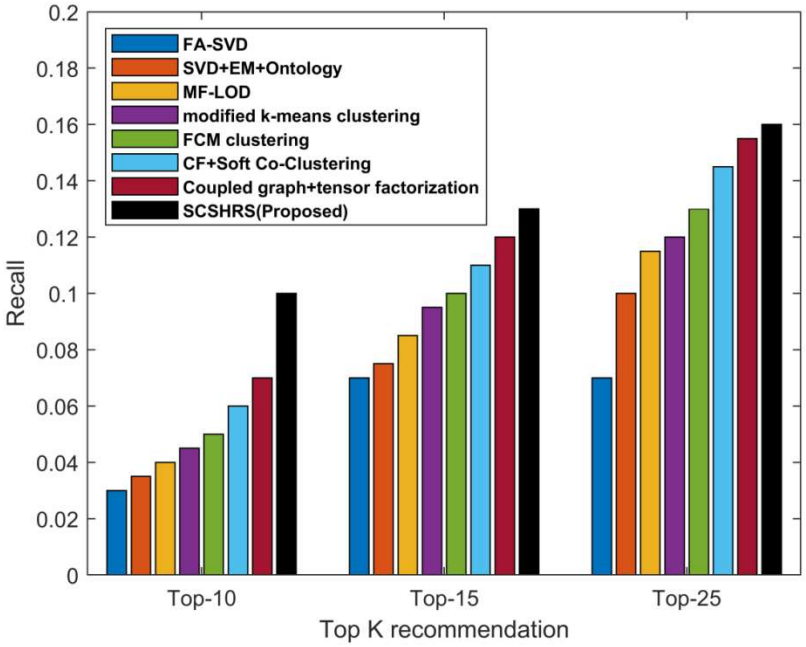
Fig.9 provides the recall score for proposed and existing ones. When evaluating the method in MovieLens 20M dataset for top 10 recommendation, the recall score obtained for Coupled graph+tensor factorization, SVD+EM+Ontology, MF-LOD, k-means clustering, FCM clustering, CF+Soft Co-Clustering, FA-SVD, and the proposed SCSHRS are0.09, 0.05, 0.06, 0.065, 0.07, 0.08, 0.04, and 0.1respectively. This analysis conveys that the proposed SCSHRS method gives better results due to the reduction in the dimension of data. This helps to accurately identify the similarities among users.



(a)



(b)

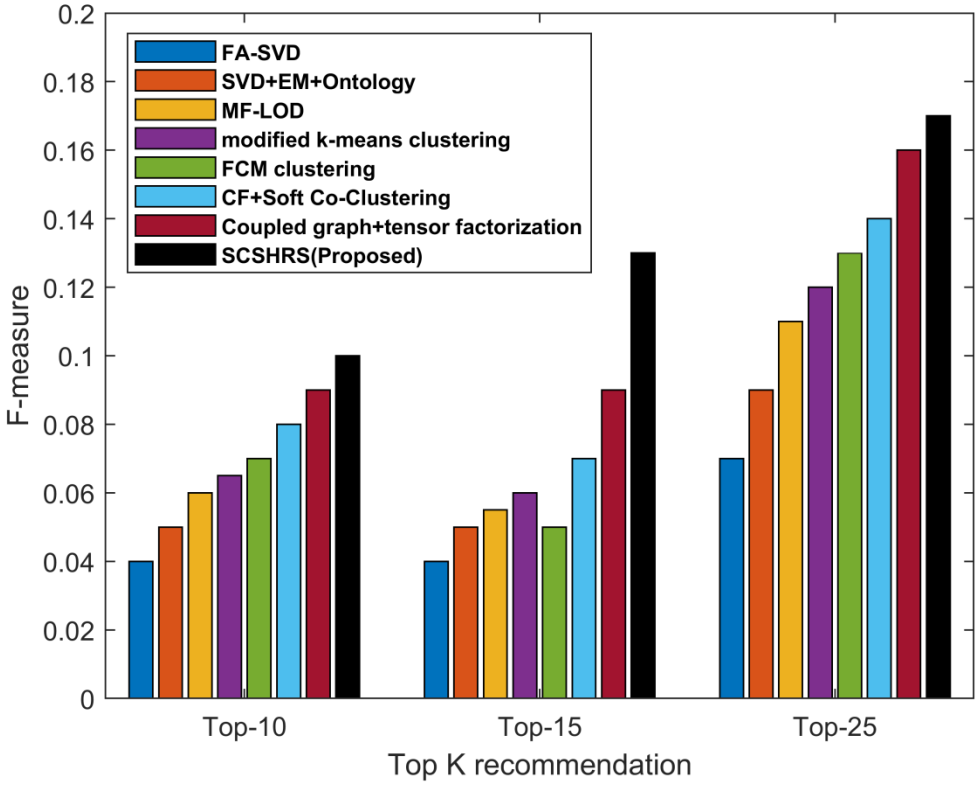


(c)

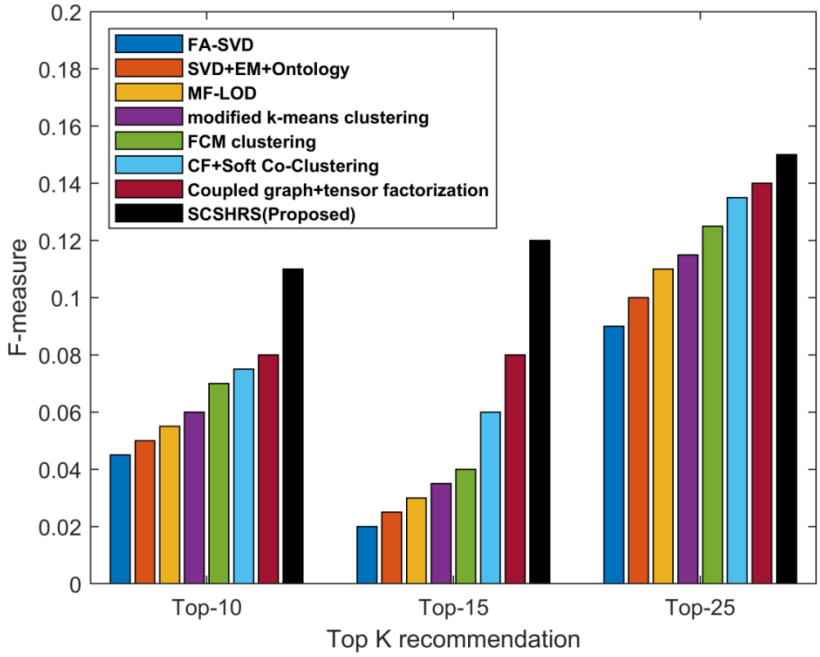
Figure 9: Recall analysis for various datasets (a) MovieLens-20M (b) Last.FM (c) Book-Crossing

4.4.5F-Measure analysis

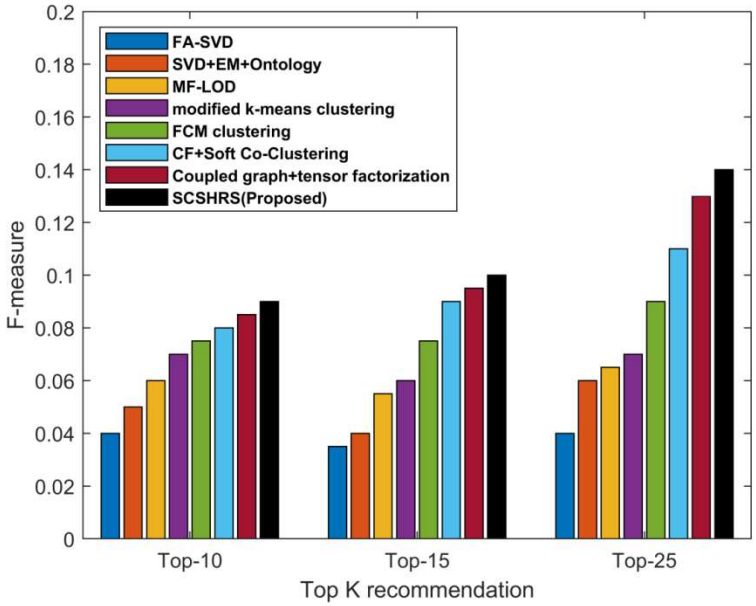
Fig.10 gives the analysis in terms of F-measure. When computing F-measure in three different datasets, the proposed SCSHRS scheme provides better F-measure than the existing systems for Top 5, Top 15, Top 20 recommendations. This is influenced by the inclusion of ANFIS for the prediction task and the HOSVD technique for dimensionality reduction. Since the ANFIS combines machine learning and fuzzy techniques, a higher F-measure score is achieved.



(a)



(b)

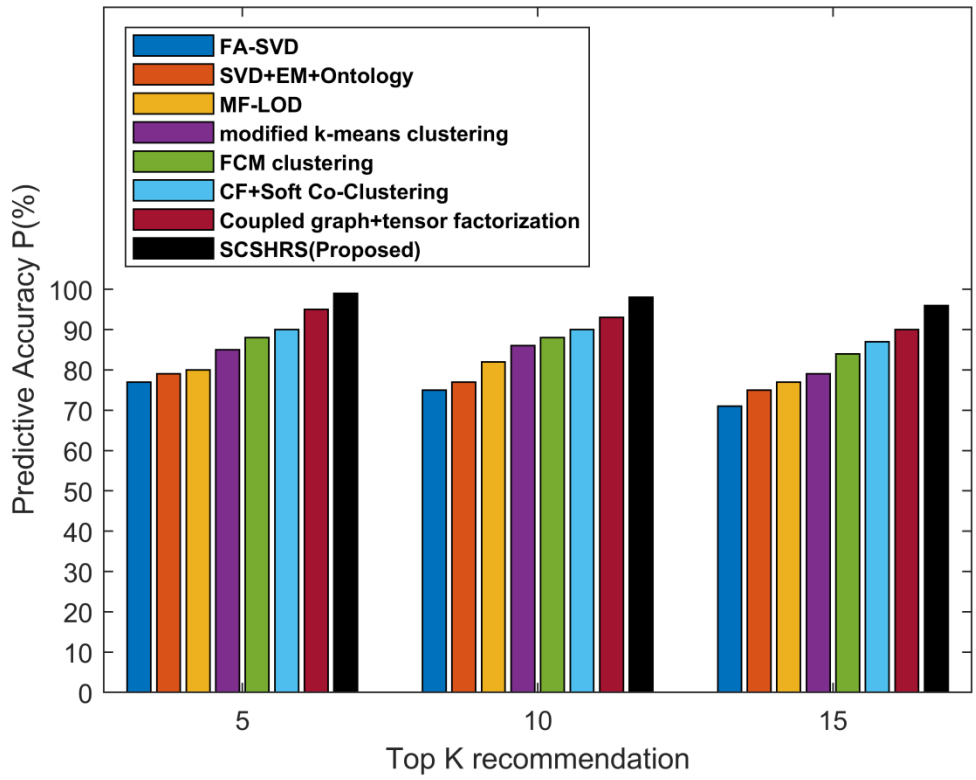


(c)

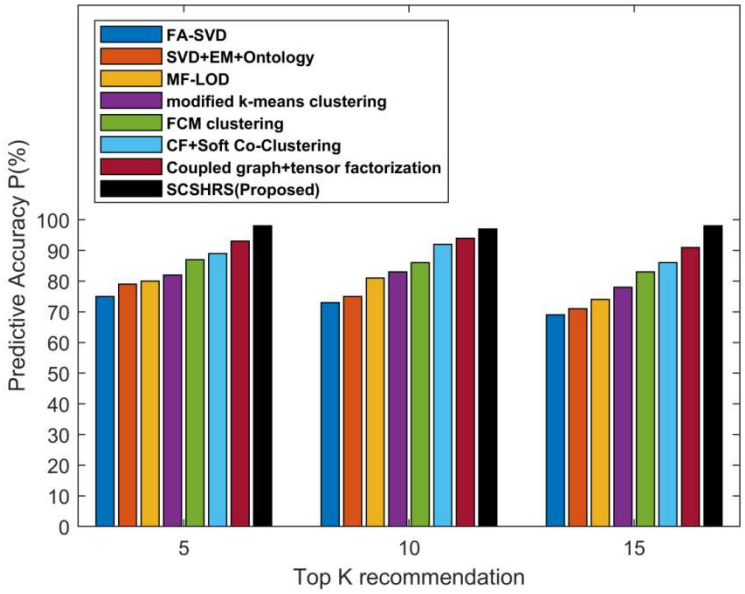
Figure 10: F-Measure analysis for various datasets (a) MovieLens-20M (b) Last.FM (c) Book-Crossing

4.4.6Accuracy analysis

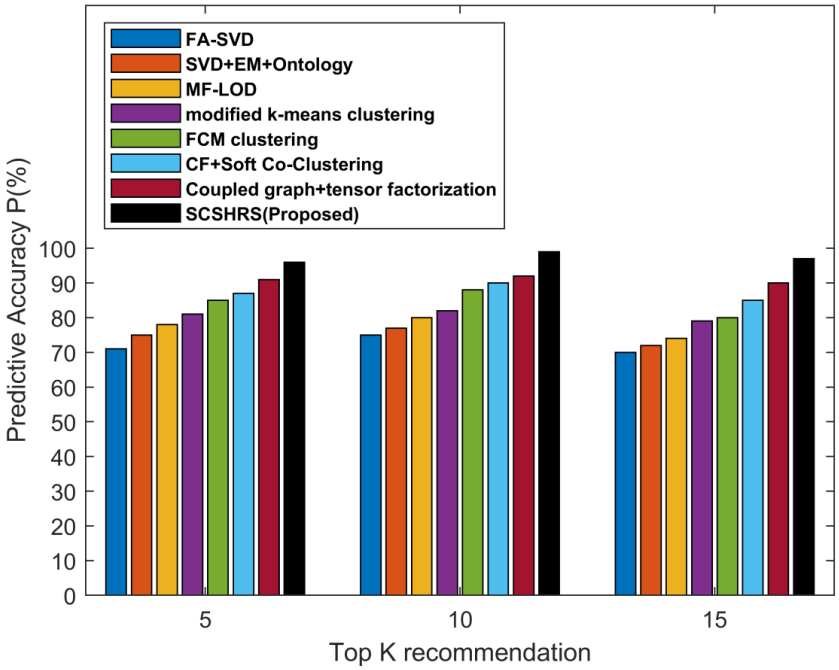
As depicted in Fig. 11, the accuracy of the proposed SCSHRS is higher than existing techniques. High accuracy is achieved by predicting the missed ratings before the clustering process. This helps to avoid sparsity problems. Moreover, the cold start problem is mitigated by the proposed hybrid mechanism that encompasses clustering, dimensionality reduction, and prediction stages. Due to this, the overall recommendation accuracy is increased. Table 5 provides statistical results obtained for the proposed method for top 10 recommendations.



(a)



(b)



(c)

Figure 11: Accuracy analysis for various datasets (a) MovieLens-20M (b) Last.FM (c) Book-Crossing

Table 5: Perform analysis for Top 10 recommendation

Technique	**Dataset**	**MAPE (%)**	**Precision**	**Recall**	**F-score**	**Accuracy (%)**
FA-SVD	MovieLens-20M	80	0.07	0.04	0.04	72

	Last.FM	76	0.085	0.03	0.042	70
	Book-Crossing	73	0.05	0.03	0.039	70
SVD+EM+ Ontology	MovieLens-20M	70	0.08	0.045	0.05	73
	Last.FM	72	0.06	0.04	0.051	71
	Book-Crossing	65	0.06	0.039	0.05	72
MF-LOD	MovieLens-20M	65	0.09	0.05	0.06	80
	Last.FM	71	0.09	0.045	0.058	74
	Book-Crossing	60	0.08	0.04	0.059	76
k-means clustering	MovieLens-20M	59	0.1	0.059	0.065	83
	Last.FM	65	0.11	0.050	0.06	76
	Book-Crossing	54	0.09	0.045	0.07	77
FCM clustering	MovieLens-20M	50	0.11	0.06	0.07	84
	Last.FM	55	0.13	0.059	0.074	80
	Book-Crossing	52	0.1	0.05	0.079	
Soft Co-Clustering	MovieLens-20M	44	0.12	0.07	0.08	87
	Last.FM	50	0.14	0.063	0.079	88
	Book-Crossing	45	0.11	0.06	0.08	82
Coupled graph+tensor factorization	MovieLens-20M	42	0.13	0.078	0.09	88
	Last.FM	44	0.16	0.079	0.08	89
	Book-Crossing	39	0.12	0.070	0.082	84
Proposed SCSHRS	MovieLens-20M	40	0.16	0.08	0.1	91
	Last.FM	36	0.18	0.09	0.105	90
	Book-Crossing	35	0.14	0.1	0.085	91

5 Discussions

Two major concerns that affect the accuracy of recommendations are data sparsity and cold start problems. Therefore, to solve these issues, a novel SCSHRS recommendation system is proposed. This system comprises of sparsity reduction stage, clustering, dimensionality reduction, and prediction stages. Initially, the sparsity level of the data is determined by Eqn. (1). From this equation, if the level of sparsity is found to be high, then the unavailable ratings are predicted using SRCF and SRWCF methods. Then clustering and dimensionality reduction is performed followed by ANFIS based prediction. Finally, the evaluation of the proposed technique is conducted on the basis of good or bad recommendations as given in Table.6.The proposed SCSHRS system is evaluated on the Movie Lens 20M, Last.FM, Book-crossing datasets and compared with the baseline algorithms such as MF-LOD [24], SVD+EM+Ontology [25], FA-SVD [29], Coupled graph+tensor factorization [41], modified k-means clustering [38], CF+Soft Co-Clustering [39], and FCM clustering [40] technique. The evaluation is performed by analyzing the results obtained for Top K recommendations. The observation outcomes depict that the proposed SRWCF technique has produced better results despite having sparse data and CSP.

Table 6: Confusion matrix

Recommendation	**Not Recommended**	**Recommended**
Bad recommendation	True Negative	False Positive
Good recommendation	False Negative	True Positive

6 Conclusion and Future Works

RS has become important and essential in social networking and business applications such as Flipkart, YouTube, Amazon, etc., and has already used filtering systems to provide specific products and services to its customers. Timely and helpful recommendations can improve the user's engagement and reliability. In the RS field, some general users only articulate ideas and rate items for a few items. Moreover, new users and new products are constantly appearing every day. These issues cause CSP and data sparsity. Hence, offering the right recommendations with very little data remains a major dispute in RSs. To mitigate these challenges, the SCSHRS approach is introduced for an accurate recommendation. Initially, the unavailable ratings are predicted to reduce data sparsity. Then the similar users are grouped depending on their demographic data such as gender, age, and profession by ALO based k-means technique. The tensor is then decomposed by HOSVD and finally, ANFIS recommends the product to the new user by using its fuzzy and machine learning abilities. The results obtained from the analysis illustrate that the SCSHRS approach has a great recommendation score in terms of error and performance investigation. However, this method has limitations such as scalability and privacy issues. Therefore, future works may focus on resolving these issues. Moreover, analyzing interests from tweets posted by the user in social media can progress the accuracy and relevancy of RS. This helps the users to decide more quickly and also saves the user time.

Ethics declarations

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Authorship contributions

All authors contributed equally to this work.

Data availability

Data can be shared if needed.

References

1. Katarya R, Verma OP. An effective collaborative movie recommender system with cuckoo search. Egyptian Informatics Journal. 2017 Jul 1;18(2):105-12.
2. Hwangbo H, Kim YS, Cha KJ. Recommendation system development for fashion retail e-commerce. Electronic Commerce Research and Applications. 2018 Mar 1;28:94-101.

3. Greenstein-Messica A, Rokach L. Personal price aware multi-seller recommender system: Evidence from eBay. Knowledge-Based Systems. 2018 Jun 15;150:14-26.
4. Qiu J, Lin Z, Li Y. Predicting customer purchase behavior in the e-commerce context. Electronic commerce research. 2015 Dec 1;15(4):427-52.
5. Pappas IO, Kourouthanassis PE, Giannakos MN, Lekakos G. The interplay of online shopping motivations and experiential factors on personalized e-commerce: A complexity theory approach. Telematics and Informatics. 2017 Aug 1;34(5):730-42.
6. Isinkaye FO, Folajimi YO, Ojokoh BA. Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal. 2015 Nov 1;16(3):261-73.
7. Subramaniyaswamy V, Manogaran G, Logesh R, Vijayakumar V, Chilamkurti N, Malathi D, Senthilselvan N. An ontology-driven personalized food recommendation in IoT-based healthcare system. The Journal of Supercomputing. 2019 Jun 1;75(6):3184-216.
8. Qiu L, Gao S, Cheng W, Guo J. Aspect-based latent factor model by integrating ratings and reviews for recommender system. Knowledge-Based Systems. 2016 Oct 15;110:233-43.
9. Guo J, Deng J, Wang Y. An intuitionistic fuzzy set based hybrid similarity model for recommender system. Expert Systems with Applications. 2019 Nov 30;135:153-63.
10. Cami BR, Hassanpour H, Mashayekhi H. User preferences modeling using dirichlet process mixture model for a content-based recommender system. Knowledge-Based Systems. 2019 Jan 1;163:644-55.
11. Sneha V, Shrinidhi KR, Sunitha RS, Nair MK. Collaborative filtering based recommender system using regression and grey wolf optimization algorithm for sparse data. In2019 International conference on communication and electronics systems (ICCES) 2019 Jul 17 (pp. 436-441). IEEE.
12. Walek B, Fojtik V. A hybrid recommender system for recommending relevant movies using an expert system. Expert Systems with Applications. 2020 May 13:113452.
13. Camacho LA, Alves-Souza SN. Social network data to alleviate cold-start in recommender system: A systematic review. Information Processing & Management. 2018 Jul 1;54(4):529-44.
14. Silva N, Carvalho D, Pereira AC, Mourão F, Rocha L. The Pure Cold-Start Problem: A deep study about how to conquer first-time users in recommendations domains. Information Systems. 2019 Feb 1;80:1-2.
15. Cheng J, Zhang L. Jaccard coefficient-based bi-clustering and fusion recommender system for solving data sparsity. InPacific-Asia Conference on Knowledge Discovery and Data Mining 2019 Apr 14 (pp. 369-380). Springer, Cham.

16. Idrissi N, Zellou A. A systematic literature review of sparsity issues in recommender systems. Social Network Analysis and Mining. 2020 Dec;10(1):1-23.
17. Yuan X, Han L, Qian S, Xu G, Yan H. Singular value decomposition based recommendation using imputed data. Knowledge-Based Systems. 2019 Jan 1;163:485-94.
18. Wang R, Cheng HK, Jiang Y, Lou J. A novel matrix factorization model for recommendation with LOD-based semantic similarity measure. Expert Systems with Applications. 2019 Jun 1;123:70-81.
19. Pujahari A, Sisodia DS. Pair-wise Preference Relation based Probabilistic Matrix Factorization for Collaborative Filtering in Recommender System. Knowledge-Based Systems. 2020 Mar 24:105798.
20. Lv G, Hu C, Chen S. Research on recommender system based on ontology and genetic algorithm. Neurocomputing. 2016 Apr 26;187:92-7.
21. Ahamed MT, Afroge S. A Recommender System Based on Deep Neural Network and Matrix Factorization for Collaborative Filtering. In2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) 2019 Feb 7 (pp. 1-5). IEEE.
22. Jain A, Gupta C. Fuzzy logic in recommender systems. InFuzzy Logic Augmentation of Neural and Optimization Algorithms: Theoretical Aspects and Real Applications 2018 (pp. 255-273). Springer, Cham.
23. Kiran R, Kumar P, Bhasker B. DNNRec: A novel deep learning based hybrid recommender system. Expert Systems with Applications. 2020 Apr 15;144:113054.
24. Natarajan S, Vairavasundaram S, Natarajan S, Gandomi AH. Resolving data sparsity and cold start problem in collaborative filtering recommender system using linked open data. Expert Systems with Applications. 2020 Jul 1;149:113248.
25. Nilashi M, Ibrahim O, Bagherifard K. A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques. Expert Systems with Applications. 2018 Feb 1;92:507-20.
26. Kermany NR, Alizadeh SH. A hybrid multi-criteria recommender system using ontology and neuro-fuzzy techniques. Electronic Commerce Research and Applications. 2017 Jan 1;21:50-64.
27. Wang H, Amagata D, Maekawa T, Hara T, Hao N, Yonekawa K, Kurokawa M. A DNN-based Cross-Domain Recommender System for Alleviating Cold-Start Problem in E-commerce. IEEE Open Journal of the Industrial Electronics Society. 2020 Jul 28.

28. Herce-Zelaya J, Porcel C, Bernabé-Moreno J, Tejeda-Lorente A, Herrera-Viedma E. New technique to alleviate the cold start problem in recommender systems using information from social media and random decision forests. Information Sciences. 2020 May 28.
29. Guo X, Yin SC, Zhang YW, Li W, He Q. Cold start recommendation based on attribute-fused singular value decomposition. IEEE Access. 2019 Jan 14;7:11349-59.
30. Jiang L, Cheng Y, Yang L, Li J, Yan H, Wang X. A trust-based collaborative filtering algorithm for E-commerce recommendation system. Journal of Ambient Intelligence and Humanized Computing. 2019 Aug 1;10(8):3023-34.
31. Kumar P, Kumar V, Thakur RS. A new approach for rating prediction system using collaborative filtering. Iran Journal of Computer Science. 2019 Jun 1;2(2):81-7.
32. Wang CD, Deng ZH, Lai JH, Philip SY. Serendipitous recommendation in e-commerce using innovator-based collaborative filtering. IEEE transactions on cybernetics. 2018 Jun 21;49(7):2678-92.
33. Mao M, Lu J, Han J, Zhang G. Multiobjective e-commerce recommendations based on hypergraph ranking. Information Sciences. 2019 Jan 1;471:269-87.
34. Sulthana AR, Ramasamy S. Ontology and context based recommendation system using Neuro-Fuzzy Classification. Computers & Electrical Engineering. 2019 Mar 1;74:498-510.
35. Ahuja R, Solanki A, Nayyar A. Movie recommender system using K-Means clustering and K-Nearest Neighbor. In2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) 2019 Jan 10 (pp. 263-268). IEEE.
36. Anitha J, Kalaiarasu M. Optimized machine learning based collaborative filtering (OMLCF) recommendation system in e-commerce. Journal of Ambient Intelligence and Humanized Computing. 2020 Jun 26:1-2.
37. Wang K, Zhang T, Xue T, Lu Y, Na SG. E-commerce personalized recommendation analysis by deeply-learned clustering. Journal of Visual Communication and Image Representation. 2020 Aug 1;71:102735.
38. Selvi C, Sivasankar E. A novel Adaptive Genetic Neural Network (AGNN) model for recommender systems using modified k-means clustering approach. Multimedia Tools and Applications. 2019 Jun;78(11):14303-30.
39. Li M, Wen L, Chen F. A novel Collaborative Filtering recommendation approach based on Soft Co-Clustering. Physica A: Statistical Mechanics and its Applications. 2021 Jan;561:125140.
40. Nozari RB, Koohi H. A novel group recommender system based on members' influence and leader impact. Knowledge-Based Systems. 2020 Oct 12;205:106296.

41. Ioannidis VN, Zamzam AS, Giannakis GB, Sidiropoulos ND. Coupled graph and tensor factorization for recommender systems and community detection. IEEE Transactions on Knowledge and Data Engineering. 2019 Sep 16.
42. Wei J, He J, Chen K, Zhou Y, Tang Z. Collaborative filtering and deep learning based recommendation system for cold start items. Expert Systems with Applications. 2017 Mar 1;69:29-39.