Revisiting Tendency based Collaborative Filtering for Personalized Recommendations

Rama Syamala Sreepada
National Institute of Technology Rourkela
Rourkela, Odisha 769008
515cs1002@nitrkl.ac.in

Bidyut Kr. Patra
National Institute of Technology Rourkela
Rourkela, Odisha 769008
patrabk@nitrkl.ac.in

Avijit Chakrabarty
National Institute of Technology Rourkela
Rourkela, Odisha 769008
216cs4185@nitrkl.ac.in

Satyadev Chandak
National Institute of Technology Rourkela
Rourkela, Odisha 769008
satyadev17chandak@gmail.com

ABSTRACT

Recommender systems alleviates the problem of information overload by providing personalized suggestions to the users. In this context, recently introduced tendency based recommendation technique is proven to be more simple, intuitive and accurate than the traditional collaborative filtering (CF) techniques. This approach computes two important statistics namely, *user tendency* and *item tendency* from the rating dataset in order to predict the final rating of an unrated item. Tendency of an item is computed using all the ratings received by it. However, these ratings might include the ratings provided by ambiguous (unstable) users. Another prominent drawback of the tendency based approach is that the tendency of an item is generic and remains unchanged across the users leading to a non-personalized recommendations. In this paper, we propose to compute item tendency in two different aspects. In the first aspect, we use an information theoretic approach to discover the most unambiguous users and utilize their ratings to compute the item tendency. In the second aspect, we compute the item tendency with respect to the active user, thereby making the tendency personalized to the users. Finally, we propose to obtain stable neighbor sets for each active user, thus making the recommendations more appropriate and accurate. Real-world datasets (Yahoo! Music, Netflix and MovieLens) are used to evaluate our approach. Experimental results show that the proposed techniques outperform the tendency based approach and traditional CF approaches across standard performance metrics.

CCS CONCEPTS

•**Information Systems** → *Information Retrieval;* •**Retrieval Tasks and Goals** → **Recommender Systems;**

KEYWORDS

Preference learning, Collaborative filtering, Tendency based CF, Personalized recommendations, Stable users

1 INTRODUCTION

In this digital era where information is abundant, making right choice or decision is highly challenging to the users. This problem is often referred as the "information overload" problem. Recommender Systems (RS) has been successful in helping people overcome the information overload problem by providing personalized suggestions to the users. The primary task of a recommender system is to provide personalized suggestions for products or items to individual user by filtering through large product or item space. These suggestions are provided by exploiting users' past rating and review history, browsing and purchase history, profile (age, gender, occupation, *etc*). In this context, RS has become an important part of e-commerce industry over the last few decades.

The algorithms that perform the task of recommendation are broadly classified into *Collaborative Filtering (CF)* and *Content-based Filtering* techniques. Content based filtering techniques provide recommendations by analyzing the active user's content, profile of the item and profiles of items s/he preferred in past [9], [13]. On the other hand, CF techniques are popular for being deployed to provide recommendations in commercial domains. CF based techniques are categorized into Model-based techniques and Neighborhood-based techniques. Model-based techniques learn the user's features and item's features through building a model using machine learning techniques [8]. On the other hand, Neighbourhood based CF computes the set of neighbors (items/ users) to the target item (or active user) based on the ratings obtained in the past. Neighbourhood based CF is broadly categorized as user-based CF and item-based CF. In [14], Resnick *et al.* proposed user-based CF, which computes the similar users (neighbors) of an active user by comparing the ratings of the active user and other users who rated the same set of items. Similarly, Sarwar *et al.* proposed item-based CF technique which predicts the ratings by computing similar items [15]. Similarity measures like pearson correlation, cosine similarity, adjusted-cosine similarity, *etc* are used while computing the neighbors (users/ items) [14], [15]. In [16], Wang *et al.* proposed a probabilistic approach in unifying the item-based and user-based predicted ratings. Likewise, in [12] *et al.* proposed an approach to modify the similarity measures while using a weighted approach in unifying user-based and item-based techniques [12]. To cope up with the sparse ratings by new users and for cold-start users, Ahn *et al.* proposed a heuristic similarity measure (PIP) which considers the proximity, impact and popularity of the co-rated items [1]. The PIP measure captures three important aspects (factors) namely, proximity, impact and popularity between a pair of ratings on the same item. The proximity factor is the simple arithmetic difference between two ratings on an item with an option of imposing penalty if they disagree in ratings. The agreement (or disagreement) is decided with respect to an absolute reference, *i.e.,* median of the rating scale. The impact factor shows how strongly an item is preferred or disliked by the users. It imposes penalty if ratings are not on the same side of the median. Popularity factor gives important to a rating which is far away from the item's average rating. This factor captures global information of the concerned item. Jesus *et al.* proposed a similarity measure named as Mean Jaccard Difference [4]. In this approach, varied similarity measures such as Mean Squared Difference, Jaccard coefficient, *etc.* are used on a weighted sum to predict the final rating. Liu's approach in [11] is in the similar lines of [1] where the similarity measure, NHSM is proposed as a function of significance, proximity and singularity. This paper overcomes the drawback of PIP approach which could not produce normalized similarity values. The term proximity is similar to the proximity measure in the PIP approach. Significance and Singularity are similar to impact and popularity of PIP approach respectively. However, in NHSM, all similarity values lie in the range (0,1). Further, each user's preferences and Jaccard similarity are also considered while computing the similarity between the users.

Apart from the traditional techniques mentioned above, Lemire *et al.* proposed Slope One predictor algorithm which works on the concept called 'popularity differential' between items [10]. This paper proposed three approaches namely, Slope One, Weighted Slope One and Bipolar Slope One predictors. The Slope One approach recommends items by computing the rating deviation between two items rated by the users. The Weighted Slope One approach also computes the item rating differential by considering the number of users who rated the items. The Bipolar Slope One predictor creates sets of liked and disliked items of each user. The rating differential of liked set and disliked set is computed separately. These differentials are combined using a weighted approach while predicting the rating of the target item. In [5], Cacheda *et al.* proposed a tendency-based technique to further improve the recommendation accuracy and computational cost. This approach computes the tendency of a user as the deviation between the ratings of the user and each item's mean rating. Similarly, item tendency is also computed. However, the tendency of the item remains unchanged across all the users, which makes recommendations non-personalized to the active user[10].

In this paper, we propose to address the drawbacks of tendency approach in two aspects. In the first aspect, we identify the stable users using an information theoretic approach. In the second aspect, we exploit neighborhood based approach to personalize the item tendency. Finally, we propose a hybrid approach to overcome the shortcomings of the tendency based approach. The main contributions of this paper are summarized below:

- An information theoretic approach is proposed to discover the stable and ambiguous users in the system. The ambiguous users' ratings are discarded while computing the item tendency.
- To personalize the item tendency, the nearest neighbors of each active user are computed. The neighbors' ratings are utilized to compute the personalized item tendency. Finally, we propose a hybrid approach to obtain the neighbors of each active user from the stable user set and utilize their ratings to compute the personalized item tendency for each active user.
- The experimental results on real-world datasets (Yahoo! Music, Netflix and Movielens) show that the proposed techniques outperform the original tendency based approach and existing CF based techniques.

Rest of the paper is structured as follows. Section 2 provides an overview of tendency based recommender system and information theory based approach. In Section 3, we explain the proposed approach. Section 4 gives the experimental details and evaluation results. Finally, we conclude our work in Section 5.

2 BACKGROUND

In this section, we discuss two important recommendation techniques, tendency based approach and information theoretic approach which are explored in this paper.

2.1 Tendency based Approach

Tendency based approach was introduced in 2011 and it has started gaining popularity due to its simplicity, intuitiveness and effectiveness [5]. This approach computes two important statistics: *user tendency* and *item tendency.* The tendency of a user is calculated as the aggregate deviation from the user's ratings to each rated item's average rating. The overall tendency of a user u is calculated as shown in Equation 1.

$$\tau_u = \frac{\sum_{i \in I}(r_{ui} - \bar{r}_i)}{|I_u|} \tag{1}$$

where r_{ui} is the rating of user u on item i, $\bar{r}_i$ is the average rating of the item i and I is the set of items rated by the user. Likewise, the tendency of an item is computed as the aggregate deviation between the item's ratings and each user's average rating. The overall tendency of an item i is computed as shown in Equation 2.

$$\tau_i = \frac{\sum_{u \in U_i}(r_{ui} - \bar{r}_u)}{|U_i|} \tag{2}$$

where, $\bar{r}_u$ is the average rating of the user u and U_i is the set of users who rated item i.

Based on the tendencies of the users (and items) and the average rating of the active user (and target item), four ways of predicting the final rating are are proposed in [5]. For an active user u and a target item i, the predicted rating $\hat{r}_{ui}$ is computed from one of the following cases.

(1) User tendency is positive and item tendency is positive: This case arises when a user has a tendency to rate the item above the item's average rating and the item has the tendency to be rated above the user's average rating. Therefore, the predicted rating might also lie above the average rating of item and user.
$\hat{r}_{ui} = max(\bar{r}_u + \tau_i, \bar{r}_i + \tau_u)$
(2) User tendency is negative and item tendency is negative: User has a tendency to rate an item below the item's average rating and the item has a tendency to be rated below the user's average rating. The rating is predicted below the user's and item's average ratings.
$\hat{r}_{ui} = min(\bar{r}_u + \tau_i, \bar{r}_i + \tau_u)$
(3) User tendency is negative and item tendency is positive: In this scenario, the predicted rating lies between average rating of the user and average rating of the item.
$\hat{r}_{ui} = min(max(\bar{r}_u, ((\bar{r}_i + \tau_u) \times \beta) + (\bar{r}_u + \tau_i)(1 - \beta)), \bar{r}_i)$, where β controls the contribution of average rating of the user and average rating of the item. The value of β ranges from 0 to 1.
(4) If the user and item tendencies do not fall in either of the conditions mentioned above, the final rating is predicted as shown below.
$\hat{r}_{ui} = \beta \times \bar{r}_u + (1 - \beta) \times \bar{r}_i$

This intuitive approach is more efficient in comparison to the traditional CF techniques. As this approach needs lesser computational time, it is capable of providing online recommendations as well. However, it can be observed from Equation 2 that the item tendency is computed from the ratings of all the users who rated the target item. Therefore, the item tendency remains same irrespective of the active user, thereby leading to "non-personalization" recommendations to the users [10]. This paper does not discuss any guideline to obtain the best value of contribution factor (β). Also, it should be noted that the consistency in users' ratings also plays a significant role in providing predictions to the users. Inconsistent users' ratings could effect the recommendations provided to the users. The following section explains the techniques to discover the consistent users in the system.

2.2 Information Theory based Approach

In this section, we discuss two information theoretic approaches that are proposed in the area of recommender systems.

Clarity based approach: In [2], Alejandro *et al.* proposed an information theoretic approach to discover the users who significantly impact the predictions and performance recommender systems. This approach introduced a metric named 'clarity'. The clarity of a user is computed in terms of ratings provided by the user in the past. The higher the clarity, the lesser the user is ambiguous. Clarity of a user u is defined as shown in Equation 3.

$$clarity(u) = \sum_{x \in X} p(x|u) \log_2 \frac{p(x|u)}{p(x)} \qquad (3)$$

where X is the vocabulary space, $p(x|u)$ is the user model and $p(x)$ is the background model. The vocabulary space could be whole rating set, whole item set or items rated by the user u. For instance, if whole rating set is considered as the vocabulary space, the clarity of the user, u is computed as shown in Equation 4.

$$clarity(u) = \sum_{r \in R} p(r|u) \log_2 \frac{p(r|u)}{p(r)} \qquad (4)$$

where R is the set of possible rating values. The correlation between user's clarity and Normalized Discounted Cummulative Gain (nDCG) is computed to understand the impact of consistent users. The nDCG is a widely used measure in Information Retrieval [7]. The nDCG measures the ranking quality of the recommendations provided to the user. If the correlation between user's clarity and nDCG is high, then it reveals that the users having high clarity play a significant role while building the recommender system.

Entropy-based approach: In [3], Alejandro *et al.* proposed another information theoretic approach to select the neighbors for each active user. This paper proposed to compute entropy of a user as shown in Equation 5. This entropy determines the uncertainty of user's preferences over the items rated by him/ her.

$$entropy(u) = -\sum_{i \in I_u} p(i|u) \log_2 p(i|u) \qquad (5)$$

where I_u is the set of items rated by user u and $p(i|u)$ is the probability of item i being rated by user u.

In our paper, we propose an approach to utilize the highly unambiguous users and create a stable neighborhood for each active user and recommend appropriate items for them.

3 PROPOSED APPROACH

In this section, we propose two techniques to extend the tendency based approach. All the steps of the proposed techniques are discussed in the following subsections.

3.1 Stable User Approach

As discussed in section 2, clarity and entropy of a user highly correlate with the accuracy of the recommendations provided to the user. Entropy of a user can be used to find the stability of the user's ratings across all the items rated by him/her. Lower the entropy of the user, higher the consistency in users' preferences over the items. However, the tendency based approach computes the item tendency using all the users who rated the item (1). This user set could have potentially ambiguous users, which subsequently hampers the tendency of the item and therefore provides inaccurate recommendations to the users. In this section, we propose to identify stable users and utilize their ratings to compute item tendency.

Let U be the set of users, I be the set of items and R be the set of all possible rating values in the system. In this scenario, the vocabulary space could be items/ users/ ratings. Therefore, we emphasize to compute stability of users' preferences over all the items rated by him/her. Therefore, in this paper the set of items rated by the user is considered as the vocabulary space.

Entropy of a user u is computed in terms of the probabilistic distribution of the items rated by the user (shown in Equation 6.

$$entropy(u) = -\sum_{i \in I_u} p(i|u) \log_2 p(i|u) \tag{6}$$

where, I_u is the set of items rated by the user u and $p(i|u)$ is the user conditioned probability over the rated items. The probability of item i being rated by a user u is computed as mentioned in Equation 7.

$$p(i|u) = \sum_{r \in R} p(i|r)p(r|u) \tag{7}$$

where, R is the set of all possible rating values, $p(i|r)$ is the probability of item i being rated with a rating value of r. This rating-conditioned probability of the item i is computed as shown in Equation 8.

$$p(i|r) = \frac{|\{u \in U | r(u,i) = r\}|}{|\{r_{ui} \in U \times I | r_{ui} = r\}|} \tag{8}$$

The second term, $p(r|u)$ (Equation 7) is the user-conditioned probability of the rating. The $p(r|u)$ is defined as the probability of a rating value r being rated by given a user u. This is computed as mentioned in Equation 9.

$$p(r|u) = \frac{|\{i \in I_u, r(u,i) = r\}|}{|I_u|} \tag{9}$$

After computing the entropy of all the users, the users whose entropy is lesser than a threshold, θ are considered to be stable users. Let the stable users set be U^s ($\{u \in U | entropy(u) < \theta\}$). This stable user set is used to compute the item tendency (τ_i^s) as shown in Equation 10.

$$\tau_i^s = \frac{\sum_{u \in U_i^s} (r_{ui}^s - \bar{r}_u^s)}{|U_i^s|} \tag{10}$$

where U_i^s is the set of stable users who rated target item i. Subsequently, mean of the target item is computed using the stable user set who rated target item i (as shown in Equation 11).

$$\bar{r}_i^s = \frac{\sum_{u \in U_i^s} r_{ui}^s}{|U_i^s|} \tag{11}$$

It can be noted that the user tendency is unchanged as shown in equation 1. After computing the tendency of the active user and the tendency of the target item, the rating can be predicted as described in original tendency based approach (Section 2) to predict the final rating. For instance, if the tendency of an active user is positive and personalized tendency of a target item is positive, the final rating (of the target item i for an active user u) is predicted as shown in Equation 12.

$$\hat{r}_{ui} = max(\bar{r}_u + \tau_i^s, \bar{r}_i^s + \tau_u) \tag{12}$$

In a scenario where the tendency of an active user is negative and personalized tendency of a target item is negative, the final rating is predicted as shown in Equation 13.

$$\hat{r}_{ui} = min(\bar{r}_u + \tau_i^s, \bar{r}_i^s + \tau_u) \tag{13}$$

The final rating is predicted in other scenarios as well.

Likewise, clarity of a user can be used to find the most stable users. In this paper, the set of items rated by the user is considered as the vocabulary space. Clarity of a user u is computed as shown in Equation 14.

$$clarity(u) = \sum_{i \in I_u} p(i|u) \log_2 \frac{p(i|u)}{p(i)} \tag{14}$$

where probability of an item i is computed as the number of ratings received by an item over all the rated items in the system (as shown in Equation 15).

$$p(i) = \frac{|\{u \in U | r(u,i) \neq \phi\}|}{|\{r_{uj} \in U \times I | r_{uj} \neq \phi\}|} \tag{15}$$

Upon computing the clarity of all the users, the users whose clarity value is greater than a certain threshold, θ_1 are considered to be stable. As explained in this section, stable user set is computed to find the item tendency as shown in Equation 10 and the mean of the target item is computed as shown in Equation 11.

However, the stable user approach does not guarantee personalized tendency computation *w.r.t* the active user and therefore does not ensure personalized recommendations. In the following subsection, we propose an efficient approach to compute personalized item tendency.

3.2 Personalized Tendency Approach

Tendency based approach captures the tendency of the users based on the ratings provided by him/her. It also captures the tendency of items based on the rating received. The item tendency is computed using all the users who rated the item. Therefore, item tendency remains unchanged across all the users in the system. This results in non-personalized suggestions to the users. In this paper, we argue that the tendency of an item varies from user to user and it has to be personalized *w.r.t.* the active user. We propose to compute the personalized item tendency for an active user as follows.

Let u and i be the active user and target item, respectively. Let U_i be the set of users who rated the target item i. To ensure the tendency of an item is personalized for the active user, we select a subset of U_i who are neighbors of the active user. Let the resultant subset be U_k ($U_k \subseteq U_i$). In this paper, we compute the neighbors using similarity measrues like Pearson Correlation, PIP [1] and NHSM [11]. Subsequently, we compute the personalized item tendency (τ_i^p) for the active user (u) using the ratings provided by the neighbors (U_k) on the target item (i) as shown in Equation 16.

$$\tau_i^p = \frac{\sum_{\acute{u} \in U_k} (r_{\acute{u}i} - \bar{r}_{\acute{u}})}{|U_k|} \tag{16}$$

where $r_{\acute{u}i}$ is the rating provided by a neighbor ($\acute{u}$) on the target item (i) and $\bar{r}_{\acute{u}}$ is the mean rating of the neighbor $\acute{u}$.

Existing tendency based approach computed the item's mean rating as an aggregation of all the ratings received by the item. However, as discussed earlier, this leads to a non-personalization problem. In our approach, we consider only the neighbors' ratings while computing the mean of the item rating. The target item's mean rating ($\bar{r}_i^p$) is computed as shown in the Equation 17.

$$\bar{r}_i^p = \frac{\sum_{\acute{u} \in U_k} r_{\acute{u}i}}{|U_k|} \tag{17}$$

After obtaining the user tendency τ_u and the personalized item tendency τ_i^p, mean rating of the active user $\bar{r}_u$, personalized mean rating of the target item $\bar{r}_i^p$, we incorporate these statistics and modify the different scenarios (mentioned in Section 2) to predict the final rating of a target item i for an active user u. For instance, if the tendency of an active user is negative and personalized tendency of a target item is positive, the final rating (of the target item i for an active user u) is predicted as shown in Equation 18.

$$\hat{r}_{ui} = min(max(\bar{r}_u, (((\bar{r}_i^p + \tau_u) \times \beta) + (\bar{r}_u + \tau_i^p)(1-\beta)), \bar{r}_i^p) \tag{18}$$

In a scenario where the tendency of an active user is negative and personalized tendency of a target item is negative, the final rating is predicted as shown in Equation 19.

$$\hat{r}_{ui} = min(\bar{r}_u + \tau_i^p, \bar{r}_i^p + \tau_u) \tag{19}$$

Likewise, the final rating is predicted in other scenarios as well.

Finally, we propose a combined approach to utilize the ratings received from stable neighbors while computing the personalized item tendency *w.r.t* an active user. This approach is discussed in the subsequent section.

3.3 Stable Neighborhood based Approach

In this section, we propose a hybrid method to reap the advantages of the Personalized Tendency Approach (PTA) and Stable User Approach (SUA) described in the previous sections. The PTA approach computes the neighbors of the active user and computes personalized item tendency. However, while computing the personalized item tendency, ratings of ambiguous/ unstable neighbors (users) creep into the system. This issue subsequently hampers the prediction of ratings and recommendations to the users. On the other hand, the SUA approach utilizes the ratings of the stable users to compute the item tendency. However, computing the item tendency using SUA is not personalized as this method uses the ratings of all the stable users irrespective of the active user, which makes the item tendency generic. To overcome the shortcomings of both the approaches, we propose to fuse both the approaches and compute the neighbors of the active user from the stable user set. Therefore item tendency for the item i is computed only from the stable neighbors of the active user u_a. The personalized item tendency in this hybrid approach is computed as shown in Equation 20.

$$\tau_i^{sp} = \frac{\sum_{u \in U_i^{sp}} (r_{ui} - \bar{r}_u)}{|U_i^{sp}|} \tag{20}$$

where $U_i^{sp} = (U_k) \cap (U_i^s)$, U_i^s is the set of stable users who rated target item (i) and U_k is the set of neighbors to the active user (u_a). Mean rating of the target item is computed from the ratings of stable neighbors of the target item as shown in Equation 21.

$$\bar{r}_i^{sp} = \frac{\sum_{u \in U_i^{sp}} r_{ui}}{|U_i^{sp}|} \tag{21}$$

This method ensures more accurate and personalized predictions as the computations are solely performed using the stable users. After obtaining the item tendency and item mean, the rating of an active user on a target item is predicted as mentioned in the Section 2.

4 EXPERIMENTAL RESULTS

This section explains the datasets used, evaluation metrics and experimental results. In this paper, Yahoo! Music, Netflix and MovieLens (1M and 10M) datasets are used to evaluate our approach. To test our approach on various levels of sparsity, we made subsets of the MovieLens 1M, 10M and Netflix and Yahoo! Music datasets. The datasets description is provided in Table 1.

Evaluation Metrics: Mean Absolute Error (MAE), Root Means Squared Error, Precision are the most popular metrics to evaluate a recommender system [6]. In this paper, MAE and precision are used to evaluate and compare existing and proposed techniques. MAE computes the absolute difference between the user's actual rating and predicted rating across all the ratings to be predicted [6]. MAE is calculated as mentioned in Equation 22.

$$MAE = \frac{\sum_{i=1}^{R_T} |r_{ui} - \hat{r}_{ui}|}{R_T} \tag{22}$$

where, R_T is the number of rating instances in the test set, r_{ui} is the actual rating value and $\hat{r}_{ui}$ is the predicted rating value. Precision is defined as the ratio of the number of relevant items that are recommended to the total number of recommended items [6]. It is computed as shown in Equation 23.

$$precision = \frac{N_{RecRel}}{N_{Rec}} \tag{23}$$

where N_{RecRel} is the the number of items that are both relevant and recommended and N_{Rec} is the number of recommended items. On the other hand, *Recall* is defined as the ratio of the number relevant items recommended to the total number of relevant items in the system. However, in the context of recommender systems, the value of recall is subjective and significantly depends on the total number of relevant items rated by each user [6]. Therefore, in this paper, we chose to compare and discuss the results of MAE and precision on the existing techniques and proposed techniques.

Table 1: Details of the Datasets Used

Dataset	# Users $\lvert U\rvert$	# Items $\lvert I\rvert$	# Ratings R_I	Sparsity (%) 1- (R_I*100/($\lvert U\rvert * \lvert I\rvert$))
Yahoo! Music	5,000	1,000	170,159	96.59
Netflix	8,141	9,318	196,656	99.74
MovieLens 10M	6,801	9,545	935,147	98.55
MovieLens 1M	950	3,304	146,639	95.32

Experimental Results and Comparison: Here, we discuss the experimental setup, results and comparison with the existing traditional user-based CF approach and tendency-based approach. The datasets are split into 75%, 10% and 15% of the rating instances as training set, validation set and test set respectively. The contribution factor (β), user entropy threshold (θ) and user clarity threshold (θ_1) are obtained using validation set. From the experiments on the validation set, we understand that there is no β value at which the system achieves the least MAE and highest precision. Therefore, we considered the harmonic mean of the β values as the contribution factor for further computation. Also, the best θ and θ_1 values are observed when the number of stable users is approximately 90% of the total users in the system. In this paper, we used Pearson Correlation (PC), PIP and NHSM similarity measures for obtaining similar users to each active user [1], [11]. For the sake of readability, the implemented techniques and the respective method names are mentioned in Table 2.

MAE results on Yahoo! Music dataset: We implemented the proposed approaches, tendency based approach, the traditional CF approaches and tested on Yahoo! Music dataset. The MAE results are plotted in Figure 1. The plot shows the values of MAE with varying the user neighborhood size from 10 to 150. From the results, it can be observed that the least error is incurred by the Stable Neighborhood - Clarity with PIP approach (1.04). The tendency based approach incurs an error of 1.22 and the MAE values for the existing CF techniques lie in the range in 1.27 to 1.19. The personalized tendency approach (with PC/ PIP/ NHSM) incur lesser error than the tendency based approach and the existing CF techniques ranging between 1.18 to 1.26. Also, from the plot it can be observed that Stable Neighborhood approaches with different similarity measures incur the least MAE (1.04) when compared to all other approaches.

MAE results on Netflix dataset: The proposed approaches are compared with the existing techniques and the MAE results are plotted in 2. From the results, it can be observed that the proposed approaches outperform the tendency based techniques and existing CF techniques. The MAE values of the proposed techniques lie in the range 0.801 to 0.836 where the least MAE is incurred by Stable Neighborhood - Clarity with NHSM approach and the highest error is incurred by Stable User - Clarity approach. Tendency based technique incur an MAE of 0.825 whereas the Personalized Tendency approach with PC/ PIP/ NHSM incur lesser MAE (in the range 0.808 to 0.816). It can be understood that the hybrid (stability and personalization) techniques outperform the stable user techniques on Netflix dataset. The stable user with entropy and clarity techniques incur an error of 0.866 and 0.836 respectively. On the other hand, Stable Neighborhood (clarity and entropy) with varied similarity measures incur the least error range of 0.801 to 0.819. The personalized tendency techniques (with PC/ PIP/ NHSM) incur error in the range of 0.808 to 0.817. The MAE values of the existing techniques lie in the range 0.825 to 0.889 where the least MAE is incurred by the tendency based approach and the highest MAE is incurred by UBCF with PC similarity measure. Overall, the proposed approaches performed significantly better than the existing approaches on Netflix dataset.

MAE results on MovieLens 10M dataset: The MAE results on MovieLens 10M dataset using varied techniques are reported in Figure 3. The tendency based approach incurs an error of 0.710. The personalized tendency with NHSM incurs the least error (0.687) among all the techniques. However, it can also be observed that the existing CF techniques (using PIP and NHSM) incurred approximately the same error as personalized tendency approach. The error values across all the techniques lie in the range 0.687 to 0.735 where the highest error is incurred by UBCF with PC similarity measure. It can be noted that the stable neighborhood approaches could not perform on-par with the personalized tendency approach in terms of MAE on MovieLens 10M dataset.

MAE results on MovieLens 1M dataset: To test the proposed approaches on varied number of users, we created a subset of 950 users and retained all the items in the MovieLens 1M dataset. The MAE results of the proposed approach and the existing approaches are shown in Figure 4. The tendency based approach is found to be worst performer in terms of MAE on MovieLens 1M dataset. Traditional approaches such as UBCF incur an approximate MAE of 0.776 whereas the tendency based approach encountered an error of 0.783. The MAE values of the traditional UBCF techniques lie in the range 0.752 to 0.777. The stable user -clarity and entropy approaches incurred errors of 0.764 and 0.763 respectively. The stable neighbhorhood approaches incurred and approximate error of 0.754 using PC/ PIP/ NHSM. The personalized tendency approach with PC/ PIP / NHSM measures lie in the range 0.747 to 0.758. The overall results show that the information theoretic and hybrid approaches performed better than the tendency and existing CF techniques. Subsequently, personalized tendency approach performed better than all the approaches.

Table 2: Implemented Techniques

	Implemented Technique	Acronym Used
Existing Techniques	Traditional User Based Collaborative Filtering with PC/ PIP/ NHSM	UBCF with PC/ PIP/ NHSM
	Tendency based approach	Tendency based approach
Proposed Techniques	Personalized Tendency Approach with PC/ PIP/ NHSM	Personalized Tendency with PC/ PIP/ NHSM
	Stable User Approach with Clarity	Stable User - Clarity based
	Stable User Approach with Entropy	Stable User - Entropy based
	Stable Neighborhood Approach using Clarity and PC/ PIP/ NHSM	Stable Neighborhood Approach - Clarity with PC/ PIP/ NHSM
	Stable Neighborhood Approach using Entropy and PC/ PIP/ NHSM	Stable Neighborhood Approach - Entropy with PC/ PIP/ NHSM

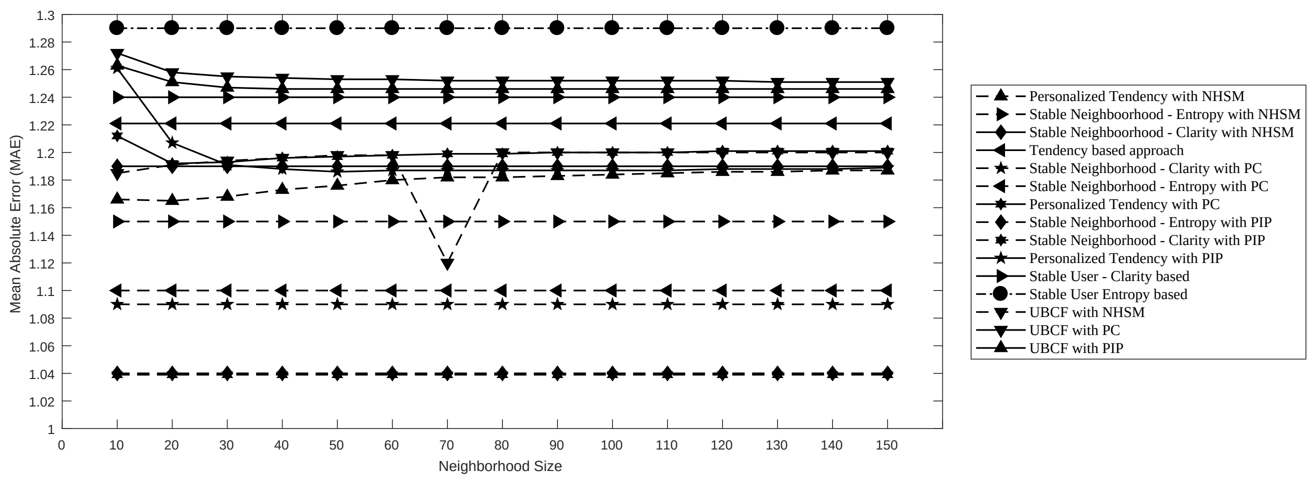

Figure 1: MAE results on Yahoo! Music dataset on different neighborhood sizes

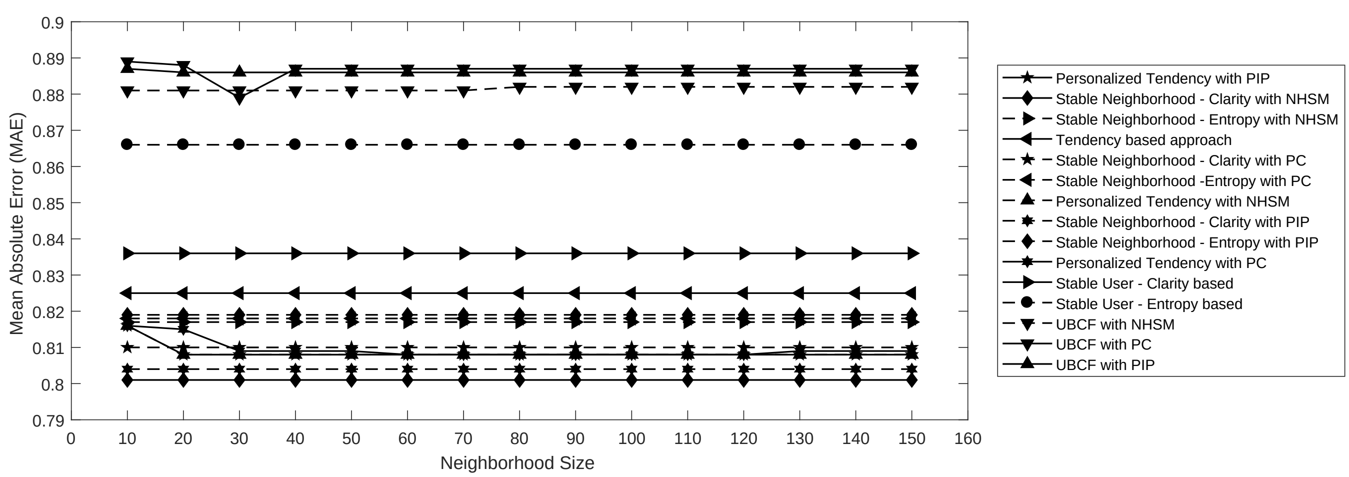

Figure 2: MAE results on Netflix dataset on different neighborhood sizes

Precision results: We computed precision of all the approaches discussed earlier and the results are reported in Table 3. From the experimental results, it can be noted that the personalized tendency techniques performed significantly better than the tendency based approach and the existing CF techniques across all the datasets. On Yahoo! Music dataset, the Stable Neighbhorhood - Clarity with NHSM achieves the highest precision (0.257). The precision values of the existing techniques lie in the range 0.219 to 0.237 where the UBCF with PC achieves the least precision and UBCF with NHSM achieves the highest (among the existing techniques). It can be observed that the Personalized Tendency with PIP/ NHSM measures achieved better precision (0.248/0.251) when compared to the tendency based approach

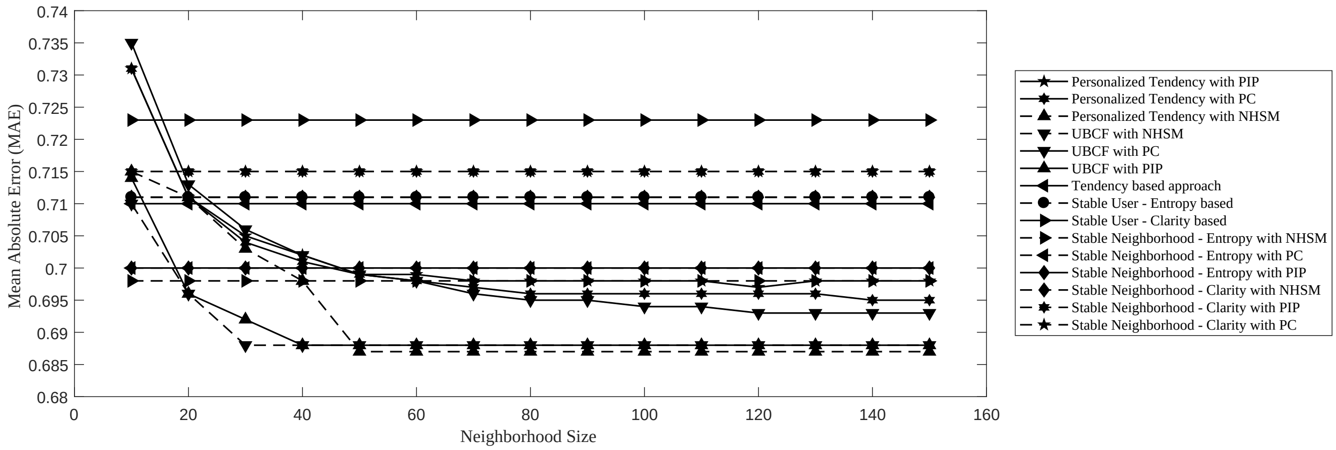

Figure 3: MAE results on MovieLens 10M dataset on different neighborhood sizes

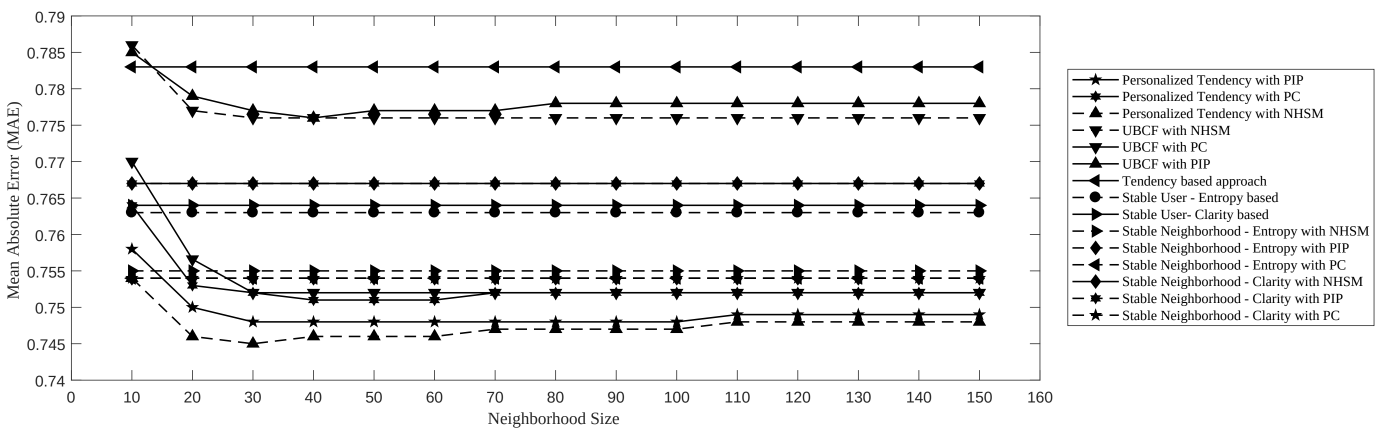

Figure 4: MAE results on Movielens 1M dataset on different neighborhood sizes

(0.225). On Netflix dataset, Stable Neighborhood - Entropy with NHSM achieves the highest precision of 0.815. The tendency based approach achieves a precision of 0.775 and the personalized tendency approach (with PC/ PIP/ NHSM), Stable Neighborhood approaches outperform the tendency based approach and existing CF approaches in terms of precision on Netflix dataset. Also, Stable Neighborhood - Entropy with NHSM achieves the highest precision (0.826) on MovieLens (ML) 10M dataset. Precision values for the existing techniques lie in the range 0.791 to 0.819 where UBCF with PIP achieves the least precision and tendency based approach achieves the highest precision. The precision values of the proposed techniques lie in the range 0.741 to 0.826 where the least precision is achieved by Stable User - Clarity based approach and the highest precision is achieved by Stable Neighborhood - Entropy with NHSM. It can be observed that the proposed techniques outperform the existing techniques in terms of precision. In terms of MAE, the proposed techniques are on-par with the compared techniques. Similar trend is observed on MovieLens 1M dataset where the highest precision is achieves by Stable Neighborhood - Entropy with PC (0.842). The tendency based approach achieves a precision of 0.819 and the UBCF techniques achieved a precision of 0.776, 0.786 and 0.798 with PC, PIP and NHSM similarity measures respectively. The personalized tendency technique and the Stable Neighborhood - Entropy significantly outperform the tendency based approach and existing UBCF techniques in terms of precision on MovieLens 1M dataset. To summarize, the proposed techniques outperform the existing approaches on all the datasets.

5 CONCLUSION AND FUTURE WORK

In this paper we proposed a personalized tendency approach to overcome the drawback of tendency based technique i.e. generalization of item tendency. For better predictions, we proposed a fusion of an information theoretic approach and

Table 3: Precision results on Yahoo! Music, Netflix, MovieLens 10M and MovieLens 1M datasets

	Technique	Yahoo! Music	Netflix	MovieLens 10M	MovieLens 1M
Existing Techniques	UBCF with PC	0.219	0.700	0.817	0.776
	UBCF with PIP	0.229	0.710	0.791	0.783
	UBCF with NHSM	0.237	0.710	0.799	0.798
	Tendency based approach	0.225	0.775	0.819	0.819
Proposed Techniques	**Stable User - Clarity based**	0.225	0.755	0.741	0.834
	Stable User - Entropy based	0.223	0.756	0.820	0.839
	Personalized Tendency with PC	0.219	0.778	0.818	0.841
	Personalized Tendency with PIP	0.248	0.779	0.816	0.830
	Personalized Tendency with NHSM	0.251	0.780	0.818	0.832
	Stable Neighborhood – Entropy with PC	0.238	0.811	0.822	**0.842**
	Stable Neighborhood – Entropy with PIP	0.240	0.811	0.812	0.832
	Stable Neighborhood – Entropy with NHSM	0.252	**0.815**	**0.826**	0.837
	Stable Neighborhood – Clarity with PC	0.237	0.796	0.817	0.812
	Stable Neighborhood – Clarity with PIP	0.233	0.798	0.816	0.806
	Stable Neighborhood – Clarity with NHSM	**0.257**	0.798	0.818	0.812

personalized tendency approach using entropy and clarity. The evaluation results on four real–world datasets show that the proposed approaches clearly outperform the tendency based approach and the existing CF techniques as well. This work can be further explored in streaming data scenarios. Also, a better guideline for choosing the contribution factor β can be proposed in future.

6 ACKNOWLEDGMENT

This work is partially supported and funded by Visvesvaraya Ph.D Scheme, MediaLab Asia, Government of India.

REFERENCES

[1] Hyung Jun Ahn. 2008. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. *Information Sciences* 178, 1 (2008), 37–51.

[2] Alejandro Bellogín, Pablo Castells, and Iván Cantador. 2011. Predicting the performance of recommender systems: an information theoretic approach. *Advances in Information Retrieval Theory* (2011), 27–39.

[3] Alejandro Bellogín, Pablo Castells, and Iván Cantador. 2014. Neighbor selection and weighting in user-based collaborative filtering: a performance prediction approach. *ACM Transactions on the Web (TWEB)* 8, 2 (2014), 12.

[4] JesúS Bobadilla, Fernando Ortega, Antonio Hernando, and JesúS Bernal. 2012. A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based Systems* 26 (2012), 225–238.

[5] Fidel Cacheda, Víctor Carneiro, Diego Fernández, and Vreixo Formoso. 2011. Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Transactions on the Web (TWEB)* 5, 1 (2011), 2.

[6] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 5–53.

[7] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[8] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009).

[9] Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the 12th international conference on machine learning*, Vol. 10. 331–339.

[10] Daniel Lemire and Anna Maclachlan. 2005. Slope one predictors for online rating-based collaborative filtering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*. SIAM, 471–475.

[11] Haifeng Liu, Zheng Hu, Ahmad Mian, Hui Tian, and Xuzhen Zhu. 2014. A new user similarity model to improve the accuracy of collaborative filtering. *Knowledge-Based Systems* 56 (2014), 156–166.

[12] Hao Ma, Irwin King, and Michael R Lyu. 2007. Effective missing data prediction for collaborative filtering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research*

and development in information retrieval. ACM, 39–46.

[13] Michael Pazzani and Daniel Billsus. 1997. Learning and revising user profiles: The identification of interesting web sites. *Machine learning* 27, 3 (1997), 313–331.

[14] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. ACM, 175–186.

[15] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. ACM, 285–295.

[16] Jun Wang, Arjen P De Vries, and Marcel JT Reinders. 2006. Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 501–508.