INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(3), 307-322, June 2017.

A Rating-based Integrated Recommendation Framework with Improved Collaborative Filtering Approaches

S. Cheng, B. Zhang, G. Zou

Shulin Cheng
1. School of Computer Engineering and Science, Shanghai University
99 Shangda Road, BaoShan District, Shanghai, 200444, PR, China
chengshulin@shu.edu.cn
2. School of Computer and Information, Anqing Normal University
1318 Jixian North Road, Anqing, Anhui Province, 246133, PR, China
chengshL@aqnu.edu.cn

Bofeng Zhang*, Guobing Zou
School of Computer Engineering and Science, Shanghai University
99 Shangda Road, BaoShan District, Shanghai, 200444, PR, China
*Corresponding author: bfzhang@shu.edu.cn
guobingzou@gmail.com

Abstract: Collaborative filtering (CF) approach is successfully applied in the rating prediction of personal recommendation. But individual information source is leveraged in many of them, i.e., the information derived from single perspective is used in the user-item matrix for recommendation, such as user-based CF method mainly utilizing the information of user view, item-based CF method mainly exploiting the information of item view. In this paper, in order to take full advantage of multiple information sources embedded in user-item rating matrix, we proposed a rating-based integrated recommendation framework of CF approaches to improve the rating prediction accuracy. Firstly, as for the sparsity of the conventional item-based CF method, we improved it by fusing the inner similarity and outer similarity based on the local sparsity factor. Meanwhile, we also proposed the improved user-based CF method in line with the user-item-interest model (UIIM) by preliminary rating. Second, we put forward a background method called user-item-based improved CF (UIBCF-I), which utilizes the information source of both similar items and similar users, to smooth item-based and user-based CF methods. Lastly, we leveraged the three information sources and fused their corresponding ratings into an Integrated CF model (INTE-CF). Experiments demonstrate that the proposed rating-based INTE-CF indeed improves the prediction accuracy and has strong robustness and low sensitivity to sparsity of dataset by comparisons to other mainstream CF approaches.
Keywords: personalized recommendation, collaborative filtering, rating integration.

1 Introduction

The recommender system [21] has been studied by many researchers in the past decade, which are widely applied in many fields like information retrieval [5], item recommendation [22], E-commerce [13]. Recommender system obtained relatively promising results and facilitated users, in which collaborative Filtering (CF) recommendation methods are classic and useful ones, and do well in recommending items with ratings such as products, movies, music. CF recommendation approaches [21] [19] [23] can be divided into memory-based approaches and model-based approaches. Memory-based approaches are heuristic and comprises item-based and user-based approaches, while model-based approaches are built based on machine learning theory. Item-based and user-based approaches both leverage the idea of neighbors to generate recommendation by measuring the similarities between the target item and other items, or, between the target user and other users. And the similarities are viewed as weights between items or users in the process of rating prediction.

But many of item-based and user-based approaches predict unknown ratings from single perspective of either users or items, in which only partial information embedded in user-item matrix is utilized. Traditional approach [22] [19] with single view has relatively low performance due to the poor ability against the sparsity of user-item matrix except few ones like recommendation in Amazon [13]. Naturally, some researchers studied the imputation of missing data [6] [16] which produced relatively good performance. But they did not consider the information of multiple sources embedded in user-item rating matrix. Therefore, we study an integrated recommendation framework of CF approaches using the information of multiple sources from user-item matrix.

In this paper, we propose a rating-based integrated recommendation framework INTE-CF with improved CF methods. Our integrated framework is, to some extent, similar but different to hybrid recommendation approach which usually combined CF methods with content-based approaches applying the strategies of pre-fusion or/and post-fusion, or built linear combination of different CF methods. Our framework could directly obtain the values of optimal fusion parameters by one time of learning, whereas other methods like [16] [24] found out the suitable values of combination parameters by many times of learning and manual comparisons. In our framework, an objective optimization function for predicting unknown ratings is put forward by considering three varying information sources from different perspectives of improved traditional CF approaches. The framework can implement more accurate recommendations through learning the optimization parameters, whose advantages are fully leveraging the information embedded in user-item matrix from three different perspectives, reducing the dependence on missing data and balancing three CF methods by optimization parameters.

The remainder of the paper is organized as follows. We first summarized the related works in section 2. The rating-based integrated recommendation framework is presented in section 3. Section 4 presented the improved approaches of traditional item- and user-based methods. We designed a background rating prediction method based on both similar users and similar items in section 5. The details of integrated recommendation framework are demonstrated in section 6. The experimental results of the proposed scheme are discussed in section 7. Finally, we discussed the findings of our work along with the future work in the last section.

2 Related works

Since the recommender systems were generated, CF recommendation has been viewed as the most successful recommender method including memory-based heuristic approaches CF methods and model-based learning approaches [14]. There have been many CF recommender applications in academia and industry. To the best of our knowledge, Tapestry system [9] identifying like-minded user is the earliest real CF recommender system. And Amazon Web Site [13] is a famous application of CF approach.

In order to increase the recommendation accuracy, many scholars tried to improve CF approaches by varying similarity calculations between users or items. Breese [3] etc. compared the prediction accuracy of several similarity algorithms including correlation coefficient-based algorithm, vector-based algorithm and statistical Bayesian algorithm. Choi [4] etc. proposed a new similarity function for selecting neighbors for each target item. Others like Conditional Probability-Based Similarity algorithm [7] of item similarity and Genetic algorithm [2] of user similarity were also studied. Good similarity computation method as a kind of enhancement of CF methods indeed improves the recommendation accuracy to some extent. But it is sensitive to the data quality like the sparsity of dataset.

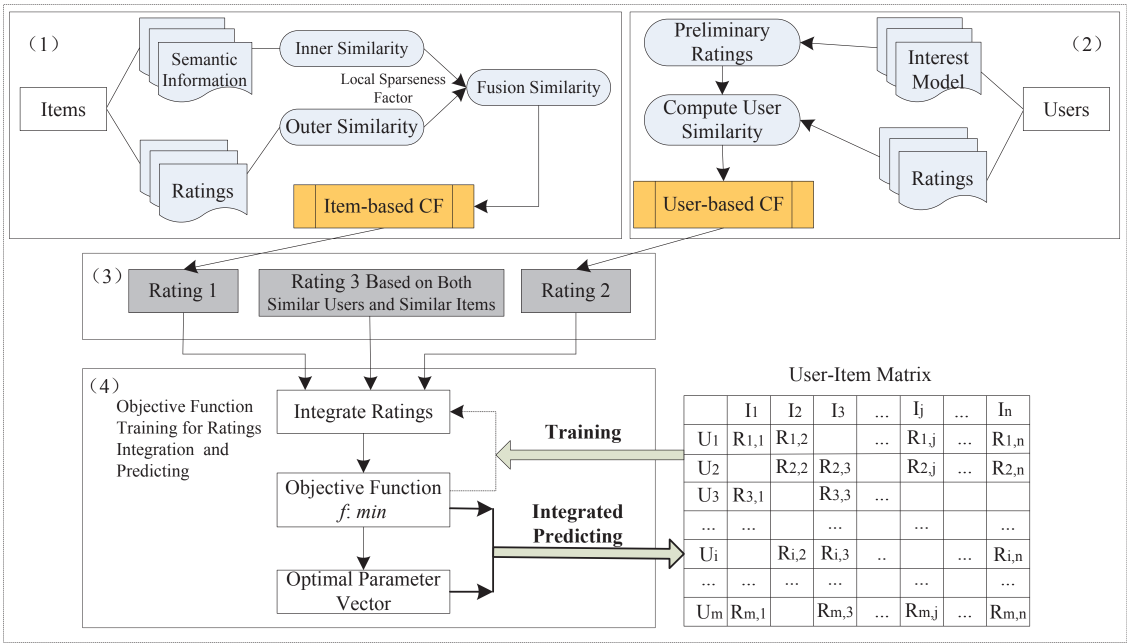
Despite the success of CF approaches, sparsity is still a major challenge and heavily affects recommendation accuracy. The fact is that a large volume of entries' value in user-item matrix is missing. Therefore, some solutions were proposed to address the issues of sparsity. The simplest ways [6] are using either the value of zero or the average rating of users or items. Obviously, these two ways are too coarse and imprecise. Later on, some relatively better methods were adopted, such as dimensionality reduction based on matrix factorization [20], imputation based on preliminary rating prediction of missing data [16] [8]. In recent years, some other information related to users or items was adopted to alleviate sparsity, like the trust and distrust relationships between users, which were studied in open Dataset of Epinions [1]. The information in social networks such as friend relations, social influence, is also researched to alleviate sparsity by some scholars [27]. Indeed, these information is useful, but they are not always obtained like in the MovieLens dataset, unless in social networks. It is straightforward that the information that could be utilized in recommendation depends on the questions of a certain specific domain. Therefore in this paper, we consider only the information embedded in user-item rating matrix and the semantic information of items to reduce the sparsity according to the limited available information.

In addition to taking full advantage of improvements of similarity computation, dimensionality reduction and other related information sources, there is another important improvement way called hybrid filtering which combines CF with other recommendation approaches. Lu [15] etc. proposed the CCF approach for the news topic recommendation in Bing, which combined CF approach and content-based filtering method. In E-commerce, Song [24] etc. leveraged both demographic recommendation techniques and CF algorithms to put forward a hybrid algorithm in order to improve recommendation accuracy. Ma [16] etc. proposed a linear combination of user- and item-based methods based on the missing value prediction by finding suitable combination parameters and obtained better performance. Moin [17] etc. suggested feature hybrid weighting schemes for improving the precision of neighborhood based CF algorithms, while it increases the complexity of computation. From the perspective of optimization, Nilashi [18] etc. proposed hybrid recommendation for CF method based on multi-criteria to improve prediction accuracy. Hybrid recommendations absorb the advantages of each recommender algorithm and do improve the precision of recommendation. They effectively alleviated sparsity and solved the problem of Cold Start to some extent, especially the combination of content-based method and CF approach. In this paper, we also leverage a similar but different idea of content-based method to improve the item-based CF approach, which utilize inner similarity of items and is discussed in subsection 4.1.

To sum all, CF is applied successfully into all kinds of recommendation fields and obtained a lot of improvements, which focuses on the similarity improvements either in user-based or in item-based methods and combinations with other types of recommendation algorithms. In this paper, we emphasis on the improvement of accuracy by CF-self integration based on three types of ratings deriving from three perspectives of users, items and both users and items. The integrated model is proved to be more accurate, effective and interpretive than some other mainstream CF methods demonstrated in our experiments. What's more, the model is easy to be paralleled to improve the running efficiency.

3 Integrated recommendation framework of CFs

Our proposed integrated CF recommendation framework, namely, INTE-CF is shown in Fig.1, which comprises four core parts. The first part is to generate the first type of rating (Rating 1) from the perspective of item by improving conventional item-based CF approach based on the fusion of two kinds of similarities of item. The second part is in charge of generating the second type of rating (Rating 2) from the perspective of user by improving the traditional user-based CF (UBCF) approach with UIIM extracted from user-item matrix. The third part demonstrates a combination model of generating the third type of prediction rating (Rating 3) based on both similar items and similar users from the two perspectives of item and user. The three types of ratings are integrated together to build an objective function f in the part 4, which is our proposed integrated optimal model. It is tuned by the optimization parameters which are learned by training sample data. The parts of 1, 2, and 3 serve the part 4. The details of generating each type of rating are demonstrated in the subsequent sections.



	I1	I2	I3	...	Ij	...	In
U1	R1,1	R1,2		...	R1,j	...	R1,n
U2		R2,2	R2,3		R2,j	...	R2,n
U3	R3,1		R3,3	...			
...	...		...		...		...
Ui		Ri,2	Ri,3	..		...	Ri,n
...	...	...	...		...		...
Um	Rm,1		Rm,3	...	Rm,j	...	Rm,n

Figure 1: The rating-based integrated recommendation framework of CF approaches

4 Improvements of traditional item-based and user-based CF

Item-based and user-based CF approaches have similar rationale to predict the unknown ratings in user-item matrix. Firstly, the neighbors of target user or item are obtained by similarities in the two CF approaches. Then the unknown rating of each entry related to the target user or item in user-item matrix is predicted by these neighbors whose similarities to the target user or item are viewed as weights in calculation. Lastly, the top-K recommendation list is generated in accordance with the predicted ratings. The details of them can refer to [7, 13, 19, 23].

4.1 Improved item-based approach by similarity fusion

Similarities between items

The performance of recommender system partially depends on the computation of similarities between items. According to the principles related to dialectics, the relevance between things is determined by inner factors and outer factors. In conventional item-based CF approaches (IBCF), the similarities which are calculated in line with the ratings in user-item matrix are measured from outer factors, namely, the perspective of user evaluation. Actually, the similarities between items are also influenced, to a large extent, by the inner factors such as the properties of items, which embody item's inherent semantic information [26]. In other words, the similarities between items depend both on inner factors and outer factors. In this paper, inner factors denote the properties of item, which are utilized to characterize items and depend on the specific objects. For instance, if the object is movie, the properties could be genres etc.; if the object is product or commodity, the properties could be appearance, genres, color, function, price, quality etc. Therefore, it is necessary to take the two kinds of factors into account to measure the similarities between items. For convenience, the similarity produced by outer factors is called outer similarity and the similarity produced by inner factors is called inner similarity. The outer similarity is calculated by the ratings of items in user-item matrix showed in Eq. (1). The inner similarity is computed by the properties of items showed in the Eq.(2).

$$sim^{\mathrm{I}}_{\mathrm{out}}(i,j) = \frac{\vec{I_i}\cdot\vec{I_j}}{||\vec{I_i}||\times||\vec{I_j}||} \tag{1}$$

$$sim^{\mathrm{I}}_{\mathrm{in}}(i,j) = \sum_{k=1}\varphi(k)sim(\Theta(k),i,j) \tag{2}$$

where Θ denotes the property set of item i and item j , $sim(\Theta(k),i,j)$ represents the similarity of item i and item j on property k in Θ, $\varphi(k)$ is the weight of property k like the genre property of a movie.

Item local sparsity factor

User-item matrix is commonly heavily sparse. The existing sparsity called global sparsity degree is measured by the ratio that is equal to the number of unknown ratings over the number of total entries in user-item matrix. When calculating the item similarity, we defined a local sparsity factor, i.e., *item local sparsity factor* which is used to describe the sparsity of the set of co-ratings from the local perspective of item.

Definition 1. (Item local sparsity) Let U^{I}_i be the set of users who rated on item i and U^{I}_j be the set of users who rated on item j, then the item local sparsity is defined as:

$$SP^{\mathrm{I}}_{i,j} = \frac{2*|U^{\mathrm{I}}_i\cup U^{\mathrm{I}}_j| - (|U^{\mathrm{I}}_i + |U^{\mathrm{I}}_j|)}{2*|U^{\mathrm{I}}_i\cup U^{\mathrm{I}}_j|} \tag{3}$$

Fusion of inner similarity and outer similarity of item

According to the aforementioned analysis, it is reasonable to fuse the inner similarity and outer similarity of items. The item local sparsity factor can be used to balance the outer similarity and inner similarity. Therefore, we define the weight function between inner and outer similarities incorporating the item local sparsity by sigmoid function as follows:

$$f(SP^{\mathrm{I}}_{i,j}) = \begin{cases} \frac{1}{1+e^{-SP^{\mathrm{I}}}} & 0\le SP^{\mathrm{I}}_{i,j}<1 \\ 1 & SP^{\mathrm{I}}_{i,j}=1 \end{cases} \tag{4}$$

Clearly, $SP^{\mathrm{I}}_{i,j}$ is between 0 and 1, and $f(SP^{\mathrm{I}}_{i,j})$ belongs to 0.5 and 1, which guarantees inner similarity always be in the resulting item similarity, because inner similarity between two items is always useful and works. When $SP^{\mathrm{I}}_{i,j}$ equals 0, namely, two items have the complete common rating users and the set of co-rating users is full, the value of $f(SP^{\mathrm{I}}_{i,j})$ is 0.5, which means inner and outer similarities have same weights. When $SP^{\mathrm{I}}_{i,j}$ equals 1 that means item i and item j have no common rating users, $f(SP^{\mathrm{I}}_{i,j})$ is set to 1, i.e., the similarity between item i and item j only depends on the inner similarity. The resulting similarity after being fused based on $f(SP^{\mathrm{I}}_{i,j})$ is as follows:

$$sim^{\mathrm{I}}(i,j) = f(SP^{\mathrm{I}}_{i,j}) * sim^{\mathrm{I}}_{\mathrm{in}}(i,j) + (1 - f(SP^{\mathrm{I}}_{i,j})) * sim^{\mathrm{I}}_{\mathrm{out}}(i,j) \tag{5}$$

The resulting similarity of items embodies the two aspects of inner factors and outer factors, effectively alleviates the dependence on the sparsity of user-item matrix and overcomes the item Cold Start problem. $f(SP^{\mathrm{I}}_{i,j})$ balances the inner factors and outer factors. Therefore, IBCF can be improved by the item fusion similarity and called IBCF-I.

4.2 Improved user-based CF for rating prediction based on user item interest model

Although conventional UBCF method can predict rating with some extent accuracy, it still has the space of improvement. The key of conventional UBCF approach is to find quality neighbors of a target user. So the similarity calculation between users is important. But due to the heavy sparsity of initial user-item matrix, sometimes the similarity calculation like cosine similarity in conventional UBCF method has relatively low accuracy, even not correct occasionally. In order alleviate the sparsity, Deng [6] etc. proposed an approach of preliminary rating of unknown rating entries, which, while, still suffered from the sparsity, since it only used the existing known ratings. Different from Deng [6], we propose a preliminary rating model (PRM) based on user-item-interest to conquer the sparsity, which is similar to imputation. And the rating prediction method UBCF-I is put forward based on PRM.

Applying CF method, user-item matrix is the information source leveraged to make study and analysis. The UIIM model is built based on KNN cluster approach using inner similarities between items. Generally, User rates similar items with similar ratings. Therefore, items that user has rated can be clustered into k clusters in line with their inner similarities for building UIIM. Then the nearest cluster to the target item with unknown rating entry is selected. Lastly, it is utilized to make preliminary rating for the unknown rating entry. So there are more co-ratings between users, which are used to calculate the similarities between users. Obviously it can produce better accuracy of user similarity than traditional UBCF method. The detailed process is discussed as follows.

Let I_p be the known rating item set of user u_p, I_q be the known rating item set of user u_q, $I^{\cup}_{p,q}$ be the union set of I_p and I_q, and $I^{\cap}_{p,q}$ be the intersect set of I_p and I_q with co-ratings namely, $I^{\cup}_{p,q} = I_p \cup I_q$, $I^{\cap}_{p,q} = I_p \cap I_q$. Then the unknown rating item sets N_p and N_q of user u_p and user u_q are $N_p = I^{\cup}_{p,q} - I_p$ and $N_q = I^{\cup}_{p,q} - I_q$, respectively.

The process of preliminary ratings of N_p and N_q are similar, here we take an example for N_p. Assume item $I_j \in N_p$, firstly, compute the semantic distances of item I_j to the k clusters in UIIM of the user u_p, and sort them by ascend. The cluster which is the nearest to the item I_j is selected as the neighbors called I_n. Then calculate the preliminary rating $R'_{p,j}$ of unknown rating entry of user u_p on item I_j according to the neighbors I_n, as follows:

$$R'_{p,j} = \frac{\sum_{l \in I_n} sim^{\mathrm{I}}(j,l) * R_{p,l}}{\sum_{l \in I_n} sim^{\mathrm{I}}(j,l)} \tag{6}$$

So far, each entry in the union set $I^{\cup}_{p,q}$ has co-ratings of user u_p and user u_q either known rating or preliminary rating. The resulting similarity between user u_p and user u_q is quality. Therefore, in user space the similarities between user u_p and other users can be calculated effectively. The nearest neighbor user set NU_p used to calculate the similarities to the target user u_p is formed in lines with the rule of top-K. Finally, UBCF-I is applied to predict the unknown ratings of user

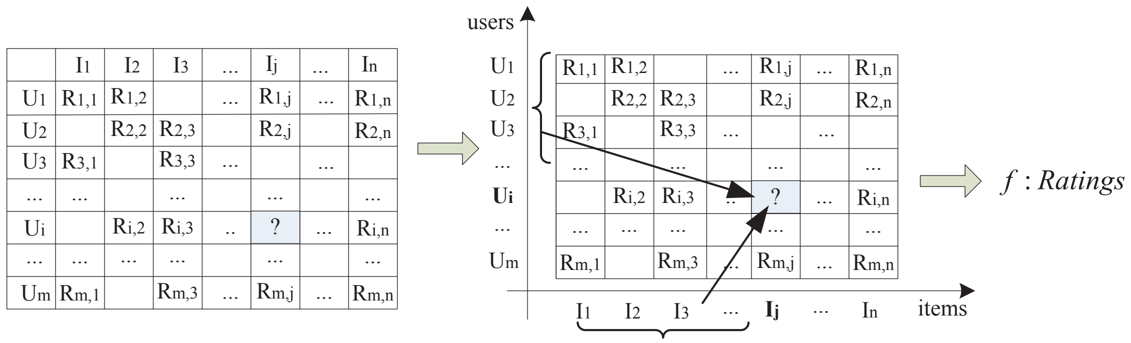
u_p.

UBCF-I method has some advantages (1) searching similar items for preliminary ratings is in a small item scope rather than in the whole item space, which derived from the most related items in UIIM; (2) avoiding the sparsity of computing the item similarity for making preliminary rating, especially for the case two users have many ratings but few common ones, since all the missing values between two users are imputed in our method when computing their similarity.

Although UBCF-I has some strong points, we don't intend to deeply research the serious user Cold Start problem in which the number of user ratings is close or equal to 0. To well solve the problem needs some other information like user social information [12, 27], trust relationships [1] and etc., which should be deeply studied but not always be achieved in traditional dataset such as MovieLens, unless in social network. Therefore, as for the very serious user Cold Start problem in which clustering is ineffective, considering averaging user's average rating and the item's average rating as the preliminary rating is a good selection [8].

5 Rating based on both similar users and similar items

IBCF-I and UBCF-I predict ratings from the perspectives of similar items and similar users, respectively. But only depending on one of them is undesirable [22, 23]. It is necessary to think about that taking both similar users and similar items into account, which correspond to the rows and columns in user-item matrix, respectively, can provide more effective information sources for predicting ratings. That means similar users making similar item ratings provides an extra and useful information source for prediction. But how to make full use of the information that derives from both similar users and similar items? Firstly, reorder user-item matrix according to the similarities of users and similarities of items. Second, generate the predictive ratings by fusing the two similarities towards the target user and the target item related to the entries with unknown ratings in user-item matrix. Therefore, we proposed one kind of CF method using the compound similarity based on the two-dimension coordinates to address the problem. For the convenience of expression, we call this method UIBCF-I. The rationale of UIBCF-I is shown in Fig.2 as follows.



(a) Origin User-Item Matrix

(b) Expression of Re-Ranked Matrix in Coordinate

Figure 2: Principle of predicting rating based on UIBCF-I

The part (a) denotes the original user-item matrix and part (b) represents the rebuilding and mapping of user-item matrix in two-dimension coordinates. Horizontal axis denotes item and vertical axis represents user. The entries of user-item matrix correspond to the points in coordinates. All the users and items are ordered in descend by the magnitudes of similarities to the target user and the target item. The question mark denotes the entry of unknown rating related to the target user U_i and the target item I_j. The top-K most similar users U_{ss} and top-M most similar items I_{ss} to the user U_i and the item I_j are selected, respectively. The predictive rating can be calculated by the compound similarity of similar users and similar items in Eq.(7).

$$R_{\mathrm{ss}}^{p}(i,j)=\frac{\sum\limits_{k\in U_{\mathrm{ss}}}\sum\limits_{m\in I_{\mathrm{ss}}} sim^{\mathrm{SS}}(i,j,k,m)*R(k,m)}{\sum\limits_{k\in U_{\mathrm{ss}}}\sum\limits_{m\in I_{\mathrm{ss}}} sim^{\mathrm{SS}}(i,j,k,m)} \tag{7}$$

where sim^{SS} represents the compound similarity of similar users and similar items, which is computed in Eq. (8) as follows:

$$sim^{\mathrm{SS}}(i,j,k,m)=\lambda_1 sim^{\mathrm{U}}(i,k)+\lambda_2 sim^{\mathrm{I}}(j,m) \tag{8}$$

λ_1 and λ_2 are tuning parameters, whose values are commonly denoted by 0.5 respectively.

6 Integrated CF recommendation model by ratings fusion

6.1 Overview

The core task of a recommendation algorithm is to predict which items a user relatively most likes based on his/her observed feedback which denotes ratings on items here. So far, we have obtained three types of ratings, namely, ratings of 1, 2 and 3 according to the aforementioned content. They are obtained by three different methods from three varying information sources. Each of them has its own strength and weakness. How to combine the three types of ratings represents a novel challenge. We proposed the optimal integration framework INTE-CF based on the three types of ratings by learning relevant parameters. In INTE-CF model, the ratings predicted by IBCF-I, UBCF-I and UIBCF-I from three varying perspectives and using different information sources which complement each other. And UIBCF-I could also be viewed as a background method of IBCF-I and UBCF-I and smooth the rating predictions generated by IBCF-I and UBCF-I. Therefore, the integration of the three types of ratings not only leverages the three varying information sources but also can reduce the dependence on data sparsity.

Let $U=\{u_1,u_2,...,u_m\}$ be the set of m users and $I=\{i_1,i_2,...,i_n\}$ be the set of n items. r_{ui},$\hat{r}_{ui}$ denote the rating and predicted rating of user u on item i, respectively. $\hat{r}_{ui}^{(1)},\hat{r}_{ui}^{(2)},\hat{r}_{ui}^{(3)}$ represent user $u\prime s$ predicted ratings by IBCF-I, UBCF-I and UIBCF-I on item i, respectively. $\vec{\hat{r}}_{ui}$ is a predicted rating vector composed of $\hat{r}_{ui}^{(1)},\hat{r}_{ui}^{(2)},\hat{r}_{ui}^{(3)}$. We also use $R\in\mathbb{R}^{m\times n}$ to represent the matrix of observed ratings and $\vec{w}=(w_1,w_2,w_2)$ to denote a parameter vector. For convenience, we use $S_u^*\subseteq U\times I$ to denote the set of user-item pairs of user u, for which the observed ratings are available.

6.2 Integration

We have obtained user $u\prime s$ three predicted ratings of $\hat{r}_{ui}^{(1)},\hat{r}_{ui}^{(2)},\hat{r}_{ui}^{(3)}$ on item i, which are leveraged to predict user $u\prime s$ preference on item i from three different perspectives. In order to achieve more accurate user $u\prime s$ predictive rating on item i, we proposed an algorithm of combining them with an integrated model to implement the aforementioned framework INTE-CF as follows:

$$\hat{r}_{ui}=\hat{r}_{ui}^{(1)}*w_1+\hat{r}_{ui}^{(2)}*w_2+\hat{r}_{ui}^{(3)}*w_3=\vec{\hat{r}}_{ui}\vec{w}^{\mathrm{T}} \tag{9}$$

Actually, it is an optimization problem of the following general form:

$$\min_{\vec{w}}(\ell(r_{ui},\hat{r}_{ui})+\mathbb{R}(\vec{w})) \tag{10}$$

Here $\ell(r_{ui}, \hat{r}_{ui})$ is a loss function measuring the discrepancy between the observed rating and the predicted rating of user $u\prime s$ on item i. The regularization function $\mathbb{R}(\vec{w})$ overly penalizes the model to suppress overfitting.

The goal of the model is to make the predicted rating $\hat{r}_{ui}$ as close to the observed rating r_{ui} as possible. The common and good selection for the loss function is to use squared error loss form:

$$\ell(r_{ui}, \hat{r}_{ui}) = \frac{1}{2} \sum_{(u,i) \in S_u^*} (r_{ui} - \hat{r}_{ui})^2 \tag{11}$$

Certainly, there are several other forms of loss function such as [14]. Here we used squared error form for loss function due to its simplicity and easiness of implementation.

We used the Frobenius norm of parameters to build the regularization function $\mathbb{R}(\vec{w})$, which was adopted by Koren [10] et al. due to its smooth differentiable property.

$$\mathbb{R}(\vec{w}) = \frac{1}{2}\lambda ||\vec{w}||_F^2 \tag{12}$$

where the parameter $\lambda \geq 0$ is used to control the strength of regularization and helps to balance between training error and model complexity. So the training model can be rebuilt as follows according to the equations of (10), (11) and (12):

$$f(\vec{\hat{r}}_{ui}, \vec{w}) = \min_{\vec{w}}(\ell(r_{ui}, \vec{r}_{ui}) + \mathbb{R}(\vec{w})) = \frac{1}{2}\min_{\vec{w}}\Big(\sum_{(u,i) \in S_u^*} \left(r_{ui} - \vec{\hat{r}}_{ui}\vec{w}^T\right)^2 + \lambda ||\vec{w}||_F^2\Big) \tag{13}$$

For convenience, we convert Eq.(13) into Eq. (14):

$$\begin{cases} f_{\min} = \frac{1}{2}\Big(\sum\limits_{(u,i) \in S_u^*} \left(r_{ui} - \vec{\hat{r}}_{ui}\vec{w}^T\right)^2 + \lambda ||\vec{w}||_F^2\Big) \\ s.t. \quad ||\vec{w}||_F^{\frac{1}{F}} = 1, \quad w_j \geq 0, \quad j \in \{1, 2, 3\} \end{cases} \tag{14}$$

Obviously, this is an optimization problem and could be solved by dynamic program with constraints. We adopted the Stochastic Gradient Descent (SGD) [11] method to learn the parameters in order to accelerate the optimization process. The optimization procedure is shown in Algorithm 1. The algorithm takes as input the matrix R of observed ratings, error ε and a group of vectors $\vec{\hat{r}}_{ui}$s which derive from the three types of predicted ratings.

7 Experiments

In order to verify our proposed integrated model INTE-CF, we experimented on the classic dataset of MovieLens[1] and EachMovie[2]. Due to the high similar results on the two datasets, we only report the experiment results of MovieLens (out of space consideration). The MovieLens dataset is comprised with 943 users, 1682 movies (items) and 100,000 ratings (1-5 scales) with the global sparsity of 0.93695, where each user has rated at least 20 items.

To better validate our proposed INTE-CF model, we conducted 4 groups of experiments corresponding to 200, 400, 600 and 800 users and relevant data extracted from the dataset at random. The purpose of dividing the dataset into 4 groups is to find out the difference of optimal parameters which, we thought, depend on the information of specific dataset, such as the size, sparsity. The experiments were finished in line with 10-fold cross-validation. We have two goals to conduct the experiments. One is to validate the higher prediction accuracy and more effectiveness of INTE-CF model. The other is to find out the component variation law of optimal parameter vector with varying scale dataset.

[1] http://www.grouplens.org/
[2] http://www.research.digital.com/SRC/EachMovie/

Algorithm 1 The optimization of INTE-CF model

Input: The rating matrix R, error ε and a group of vectors $\vec{\hat{r}}_{ui}$s of three types of predicted ratings.
Output: Model parameter vector $\vec{w} = (w_1, w_2, w_3)$.
Begin
Initialize the vector $\vec{w} = (0.333, 0.333, 0.334)$, $k \leftarrow 1$ and $f^{(0)} = 0$;
Calculate $f^{(1)}$;
while $|f^{(k)} - f^{(k-1)}| > \varepsilon$ **do**
    $k \rightarrow k + 1$;
    $s_1^{(k)} \leftarrow -\frac{\nabla f(w_1)}{||\nabla f(w_1)||}$, $s_2^{(k)} \leftarrow -\frac{\nabla f(w_2)}{||\nabla f(w_2)||}$;
    $w_1 \leftarrow w_1 + \alpha_k s_1^{(k)}$, $w_2 \leftarrow w_2 + \alpha_k s_2^{(k)}$;//$\alpha_k$ is learning parameter
    $w_3 \leftarrow 1 - w_1 - w_2$;
    Calculate $f^{(k)}$;
end while
Return vector $\vec{w} = (w_1, w_2, w_3)$
End

7.1 Preliminary

Here we conducted the experiments on the dataset of MovieLens. In order to calculate the inner similarities between items (movies) discussed in the sub-section of 4.1, we need to quantify the information of item's genre properties which characterize the movie's inherent features. The genre of each movie in the dataset is multi-valued and has 20 possible values such as drama, action and comedy. In general, more than one of these genres present with different degree in a movie. Some of them with high presence are called major or dominating genres for that movie. For example for the movie "*Copycat*" as presented in the MovieLens dataset [28] has "*Crime/Mystery/Thrill/Drama*", *Crime* is the most dominating genre value; Mystery is the second one, etc. The rest in 20 possible genres do not present. Therefore, in order to quantify the presence degree, we utilize a Gaussian-like function [28] to compute it.

$$\mu(g_i, I_j) = r_i / 2^{\sqrt{\alpha * N_j * (r_i - 1)}} \tag{15}$$

Where g_i denotes the genre i of item I_j, N_j represents the total number of the present genres, r_j denotes the rank position that indicates the magnitude of presence of g_i and $1 \leq r_i \leq N_j$. The rank positions of those genres with no presence equal 0. And $\alpha > 1$ is a constant threshold which controls the difference in presence degree of the g_i in the item I_j. Here we set $\alpha = 1.2$ which makes the calculation perfect [28]. Due to no information of rank positions of genres in the dataset, we complemented them by crawling the information from the online Movie Database(http://www.imdb.com/).

7.2 Metrics

As for assessing the accuracy of a recommender system with prediction ratings, one of the most popular evaluation metrics is Mean Absolute Error (MAE) [21, 27], which measures the average absolute deviation between the real rating assigned by the user and the predicted rating calculated by a certain recommendation algorithm. Therefore, we use MAE to measure the prediction quality of our proposed integrated framework INTE-CF with other mainstream CF methods, which is defined as follows.

$$\mathrm{MAE} = \frac{\sum_{(u,i)\in R_{\text{test}}} |r_{u,i} - \hat{r}_{u,i}|}{|R_{\text{test}}|} \tag{16}$$

where R_{test} is the set of all user-item pairs (u, i) in the test set. The smaller MAE value means a better performance.

7.3 Preliminary experiments of verifying UBCF-I and IBCF-I methods

We first conducted a preliminary experiment to verify the effectiveness of our proposed improved CF methods IBCF-I and UBCF-I, which are compared to the conventional methods of UBCF and IBCF so that we could proceed to do the next further experiments of INTE-CF model. We randomly selected half of data of the MovieLens dataset to conduct the preliminary experiment. The data was split into two parts, namely, training set 80% and prediction set 20%. The related users and items are 444 and 1605, respectively. The global sparsity is 0.9298. The experimental results are showed in Fig.2.

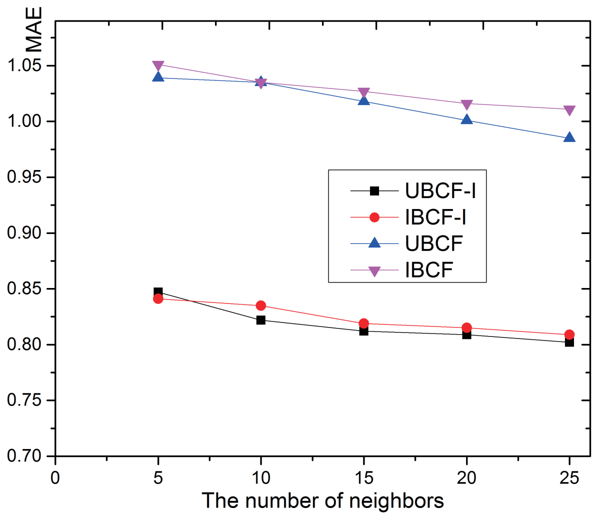


Figure 3: Preliminary experiment about comparisons of MAE between improved methods and conventional methods

Obviously, UBCF-I and IBCF-I methods are more accurate than the conventional methods of UBCF and IBCF from overall view. UBCF-I obtains maximum 19.57% and average 17.25% increases than UBCF, respectively. Similarly, IBCF-I obtains maximum 18.93% and average 18.12% increases than IBCF method, respectively. They all get lower MAE value with the increase of neighbors since UBCF-I approach benefits from UIIM and IBCF-I method benefits from fusion similarity composed of inner and outer similarities. We achieved the significant improvements of UBCF-I and IBCF-I approaches on MAE and not planned to further analysis and conduct the preliminary experiments. We emphasized on the subsequent experiments of our integrated framework model INTE-CF.

7.4 Experiments for predictive accuracy

We conducted 4 experiments in which users are divided into 4 groups of 200, 400, 600 and 800. For convenience, we call them G2, G4, G6 and G8, respectively. We compared our integrated model INTE-CF to two individual predictors, namely, UBCF-I and IBCF-I, and other two combination predictors, namely, our proposed UIBCF-I method and a linear combination method [16],

Table 1: Comparison to other CF methods: A smaller value means a better performance

Groups	G2	G4	G6	G8
INTE-CF	**0.792**	**0.744**	**0.731**	**0.711**
UIBCF-I	0.887	0.827	0.774	0.748
UBCF-I	0.845	0.775	0.763	0.759
IBCF-I	0.873	0.778	0.786	0.764
UI-Linear	0.861	0.819	0.762	0.728

Table 2: The value of average optimal vector in 4 groups

W	G2	G4	G6	G8
W3	0.625	0.697	0.711	0.725
W2	0.204	0.192	0.174	0.166
W1	0.171	0.111	0.115	0.109

which, for convenience, is called UI-Linear showed at the last row in Table 1. The optimal number of neighbors was 35 selected by many tests. Table 1 summarizes the results, showing the how INTE-CF approach outperforms the other methods in all 4 groups of experiments.

INTE-CF approach is the best recommendation method in Table 1. UBCF-I and IBCF-I have relatively low performance compared to INTE-CF, although they are improved based on standard CF method. UIBCF-I similar to UI-Linear is less accurate than UBCF-I and IBCF-I in G2 and G4 because of the less data when both considering similar users and similar items. But their accuracy increases fast with more users and items. If there are enough users and items, they will outperform UBCF-I and IBCF-I just like in G6 and G8, since they benefit from the strengths of combination. INTE-CF has the best performance which fuses three information sources deriving from UBCF, IBCF and UIBCF-I, and absorbs the advantages of them. We also found that the performances of all the methods have been improved with more users in dataset. It is evident that more users and items produce more ratings on whole, which provide more accurate prediction when applying enhanced CF methods.

7.5 Discussion about optimal parameter vector

Each group of experiments generates 10 optimal parameter vectors in 10-fold cross-validation experiments. In order to demonstrate the overall changes of vectors and proportions of UBCF-I, IBCF-I and UIBCF-I, we select the average optimal parameter vector of each group of experiments for the comparisons. Table 2 and Fig.3 show the changes of value derived from all the 4 groups of average optimal parameter vectors.

Each vector contains three components which are W1, W2 and W3 corresponding to the weights ω_1, ω_2 and ω_3 of IBCF-I, UIBCF-I and UIBCF-I in parameter vector, respectively. In each group of experiment, W1, W2 and W3 have the similar situations. W3 is dominant value and plays an important role in predicting ratings, especially in more rating data. W1 and W2 decreased with the increase of rating data, maybe since both more similar users to the target user and more similar items to the target item result in great influence on UIBCF-I. The optimal value of W3 is in vibration around 0.7 in most cases. Combining Table 1 and Fig.3, our proposed model of INTE-CF makes full use of the three kinds of information sources of IBCF-I, UBCF-I and UIBCF-I from varying views and obtains the best performance.

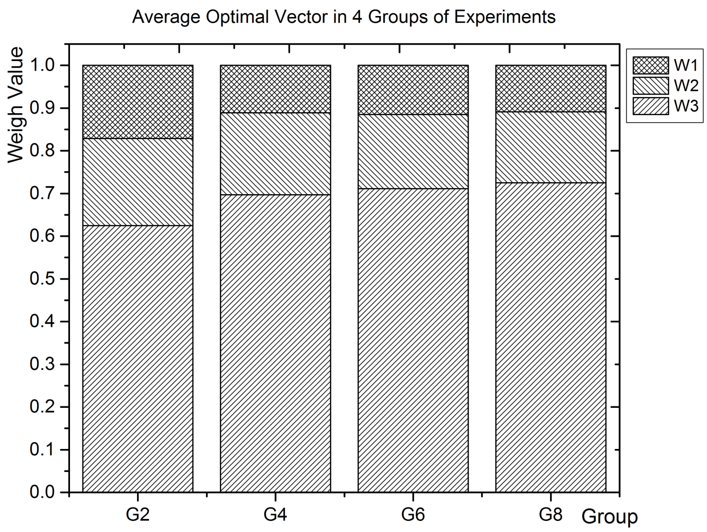


Figure 4: Three components of average optimal parameter vector in 4 groups

Table 3: Statistics information of 4 groups

Groups	G2	G4	G6	G8
Users	200	400	600	800
Items	1409	1484	1596	1655
Existing data count	22378	41826	64384	85697
Theoretical data count	281800	553600	957600	1324000
Global sparsity	0.9206	0.9296	0.9328	0.9353

Table 3 gives the statistics information of 4 groups of experiments. The global sparsity in each group is adjacent regardless of the increase of the rating data, and is close to the global sparsity 0.93695 of the whole dataset. The three components of the optimal parameter vector changes a little in 4 groups. The background method of UIBCF-I plays a great role in 4 groups, especially in G8 whose global sparsity is relatively large. The optimal parameter vector is low sensitive to the data size. The component weights of UIBCF-I in 4 groups are all high since compound similarity between items does work and UIBCF-I makes full use of both similar users and similar items.

8 Conclusions and future work

As for the shortage of individual predictors of conventional item-based and user-based CF recommendation approaches utilizing single information source, we proposed a rating-based integrated framework to combine three CF recommendation methods of IBCF-I, UBCF-I and UIBCF-I. UIBCF-I is considered as a background method to smooth the rating predictions of UBCF-I and IBCF-I. Meanwhile, we improved traditional item-based CF by inner similarity and outer similarity, and user-based CF by preliminary ratings based on UIIM. Furthermore, we built an optimal learning model INTE-CF of the framework by dynamic program with constraints to find out the optimal parameter vector in rating predictions. The experiments showed that our new integration framework of CFs is effective in improving the prediction accuracy of CF recommendation approaches. That is to say our integrated model INTE-CF leveraging the three kinds of information sources achieves the best performance. But INTE-CF pays the price of a little more running time which is not avoided but worthy. Fortunately, some calculations which accelerate the overall execution process are off line or incremental, including inner similarity of between items, clustering of UIIM, and the predictions of three ratings (rating 1, 2 and 3) can be parallel processing.

In the future work, we will compare INTE-CF model to more other methods such as SF [25] and evaluate it on more metrics. The parallelization of INTE-CF is also interesting when encountered big data. And we will continue to optimize INTE-CF by considering the rating biases of users and items and the influence of time.

Acknowledgment

We would like to thank all of the anonymous reviewers for their insightful comments and useful suggestions that must lead to a much higher quality of our manuscript.This work was partially supported by the National Science Natural Foundation of China (nos. 61303096).

Bibliography

[1] Anand D., Bharadwaj K.K. (2013); Pruning trust-distrust network via reliability and risk estimates for quality recommendations, *Social Network Analysis and Mining*, 3(1), 65-84, 2013.

[2] Bobadilla J., Ortega F., Hernando A.; Alcal J. (2011); Improving collaborative filtering recommender system results and performance using genetic algorithms, *Knowledge-based systems*, 24(8), 1310-1316, 2011.

[3] Breese J.S., Heckerman D., Kadie C. (1998); Empirical analysis of predictive algorithms for collaborative filtering, *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, 43-52, 1998.

[4] Choi K., Suh Y. (2013); A new similarity function for selecting neighbors for each target item in collaborative filtering, *Knowledge-Based Systems*, 37, 146-153, 2013.

[5] Das A. S., Datar M., Garg A., Rajaram S. (2007); Google news personalization: scalable online collaborative filtering, *Proceedings of the 16th international conference on World Wide Web*, 271-280, 2007.

[6] Deng A.L., Zhu Y.Y., Shi B. (2003); A collaborative filtering recommendation algorithm based on item rating prediction, *Journal of Software (Chinese)*, 14(9), 1621-1628, 2003.

[7] Deshpande M., Karypis G. (2004); Item-based top-n recommendation algorithms, *ACM Transactions on Information Systems (TOIS)*, 22(1), 143-177, 2004.

[8] Ghazanfar M.A., Pršgel-Bennett A. (2013); The Advantage of Careful Imputation Sources in Sparse Data-Environment of Recommender Systems: Generating Improved SVD-based Recommendations, *Informatica (Slovenia)*, 37(1), 61-92, 2013.

[9] Goldberg D., Nichols D., Oki B.M., Terry D. (1992); Using collaborative filtering to weave an information tapestry, *Communications of the ACM*, 35(12), 61-70, 1992.

[10] Koren Y. (2010); Collaborative filtering with temporal dynamics, *Communications of the ACM*, 53(4), 89-97, 2010.

[11] Li Q., Sato I., Murakami Y. (2007); Efficient stochastic gradient search for automatic image registration, *International Journal of Simulation Modelling (IJSIMM)*, 6(2), 114-123, 2007.

[12] Li W., Ye Z., Xin M., Jin Q. (2015); Social recommendation based on trust and influence in SNS environments, *Multimedia Tools and Applications*, 1-18, 2015.

[13] Linden G., Smith B., York J. (2003); Amazon.com recommendations: Item-to-item collaborative filtering, *IEEE Internet computing*, 7(1), 76-80, 2003.

[14] Liu N.N., Zhao M., Yang Q. (2009); Probabilistic latent preference analysis for collaborative filtering, *Proceedings of the 18th ACM conference on Information and knowledge management*, 759-766, 2009.

[15] Lu Z., Dou Z., Lian J., Xie X., Yang Q. (2015); Content-Based Collaborative Filtering for News Topic Recommendation, *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 217-223, 2015.

[16] Ma H., King I., Lyu M.R. (2007); Effective missing data prediction for collaborative filtering, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 39-46, 2007.

[17] Moin A., Ignat C.L. (2014); *Hybrid weighting schemes for collaborative filtering (Doctoral dissertation, INRIA Nancy)*, France, 2014.

[18] Nilashi M., bin Ibrahim O., Ithnin N. (2014); Hybrid recommendation approaches for multi-criteria collaborative filtering, *Expert Systems with Applications*, 41(8), 3879-3900, 2014.

[19] Park D. H., Kim H. K., Choi I.Y., Kim J.K. (2012); A literature review and classification of recommender systems research, *Expert Systems with Applications*, 39(11), 10059-10072, 2012.

[20] Paterek A. (2007); Improving regularized singular value decomposition for collaborative filtering, *Proceedings of KDD cup and workshop*, 5-8, 2007.

[21] Ricci F., Rokach L., Shapira B. (2011); *Introduction to recommender systems handbook*, Springer, 2011.

[22] Sarwar B., Karypis G., Konstan J., Riedl J. (2001); Item-based collaborative filtering recommendation algorithms, *Proceedings of the 10th international conference on World Wide Web*, 285-295, 2001.

[23] Shi Y., Larson M., Hanjalic A. (2014); Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges, *ACM Computing Surveys (CSUR)*, 47(1), 3-45, 2014.

[24] Song R.P., Wang B., Huang G.M., Liu Q.D., Hu R.J., Zhang R.S. (2014); A hybrid recommender algorithm based on an improved similarity method, *Applied Mechanics and Materials*, 475, 978-982, 2014.

[25] Wang J., De Vries A.P., Reinders M.J. (2006); Unifying user-based and item-based collaborative filtering approaches by similarity fusion, *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 501-508, 2006.

[26] Xu S. Y., Raahemi B. (2016); A Semantic-based service discovery framework for collaborative environments, *International Journal of Simulation Modelling (IJSIMM)*, 15(1), 83-96, 2016.

[27] Yang X., Guo Y., Liu Y., Steck H. (2014); A survey of collaborative filtering based social recommender systems, *Computer Communications*, 41, 1-10, 2014.

[28] Zenebe A., Zhou L., Norcio, A.F. (2010); User preferences discovery using fuzzy models, *Fuzzy Sets and Systems*, 161(23), 3044-3063, 2010.