Enhancing Collaborative Filtering Systems with Personality Information

Rong Hu
Human Computer Interaction Group
Swiss Federal Institute of Technology in Lausanne
CH-1015, Lausanne, Switzerland
rong.hu@epfl.ch

Pearl Pu
Human Computer Interaction Group
Swiss Federal Institute of Technology in Lausanne
CH-1015, Lausanne, Switzerland
pearl.pu@epfl.ch

ABSTRACT
Collaborative filtering (CF), one of the most successful recommendation approaches, continues to attract interest in both academia and industry. However, one key issue limiting the success of collaborative filtering in certain application domains is the cold-start problem, a situation where historical data is too sparse (known as the sparsity problem), new users have not rated enough items (known as the new user problem), or both. In this paper, we aim at addressing the cold-start problem by incorporating human personality into the collaborative filtering framework. We propose three approaches: the first is a recommendation method based on users' personality information alone; the second is based on a linear combination of both personality and rating information; and the third uses a cascade mechanism to leverage both resources. To evaluate their effectiveness, we have conducted an experimental study comparing the proposed approaches with the traditional rating-based CF in two cold-start scenarios: sparse data sets and new users. Our results show that the proposed CF variations, which consider personality characteristics, can significantly improve the performance of the traditional rating-based CF in terms of the evaluation metrics MAE and ROC sensitivity.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering*; H.1.2 [**Models and Principles**]: User/Machine Systems – *human information processing*

General Terms
Algorithms, Performance

Keywords
Recommender System, Collaborative Filtering, Personality, User Similarity, Cold Start

1. INTRODUCTION
Recommender systems are being broadly adopted in various applications to suggest items of interest to users amidst the enormous volume of available information [1, 25]. Collaborative filtering (CF) is one of the most successful and widely implemented recommendation technologies [26]. It predicts the potential interests of a given user (called an active user) by taking into account the opinions of users with similar taste (i.e., social wisdom). Compared to other recommendation technologies (e.g., content-based filtering [1]), collaborative filtering technologies have the capability of working in domains where items' attribute contents are difficult to obtain or cannot easily be parsed by automatic processes. In addition, CF algorithms can provide serendipitous recommendations, which are not similar to the items in the active user's profile, but surprisingly interest him/her [18]. In other words, CF helps users discover new items.

Despite its widespread adoption, CF suffers from several major limitations including cold-start problems, system scalability, and synonymy [26]. In this study, we focus on the cold-start issue, which includes both data sparsity and new user problems. In the former, there is a severe lack of historical data. For example, in many real world applications, users' historical data, such as what they have viewed, purchased or rated, is sparse by nature because the website is in its initial operational stage. Therefore, it is highly probable that either the similarity between any two given users is nearly zero or the measures are so unreliable that they cannot be used [3]. A related case is the new user issue where systems cannot accurately identify recommendations for new users because of the limited number of ratings that they have provided. In either case, the problem is detrimental in effectively identifying similar users, which is considered the key module of collaborative filtering. The cold-start problem, therefore, hinders the implementation of CF methods in many practical applications.

In this paper, we present novel approaches that aim at overcoming the cold-start limitations. More specifically, our approaches attempt to make use of human personality characteristics as complementary information by incorporating them into the traditional rating-based CF methods. In the research realm of user modeling, human factors, such as personality and cognitive/learning style, have been demonstrated to play an important role in the personalization process [7, 16]. Prior studies have also shown that personality influences the human decision making process and reveals a person's long-term tastes [24]. Drawing on the inherent inter-related patterns between personalities and users' interests/behaviors, many researchers have recently investigated the incorporation of human personality into recommender systems and have achieved promising results [5, 12, 21]. In prior work, human personality characteristics have been postulated to have the potential ability to lessen the cold-start problem [5]. However, few works have thus far empirically verified this hypothesis.

RecSys'11, October 23–27, 2011, Chicago, Illinois, USA.

In this study, we are trying to fill the research gap by investigating *whether* personality can be applied to deal with the cold-start problem resulting from the lack of sufficient ratings, *how* personality can be used, and *when* the personality-based methods can work most effectively. The contributions of this paper are: 1) we present one pure personality-based CF and two integrative approaches, one combining personalities and ratings in a linear way and the other based on a cascading mechanism; 2) we conduct two experiments for both concerned cold-start issues to compare the performance of the proposed personality-based CF variations with that of the traditional rating-based CF, in terms of two evaluation metrics MAE and ROC sensitivity; 3) our results provide empirical evidences that the leverage of personality indeed can alleviate the cold-start problem existing in rating-based CF recommender systems.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of related research work. Section 3 describes the algorithm of traditional user-based CF systems. The following section presents our proposed personality-based approaches in detail. Section 5 describes our experimental study, including experiment dataset, methodology, evaluation metrics, results analysis and discussion, followed by a final section on conclusions and future work.

2. RELATED WORK

2.1 Cold-start Problem

In the literature, many researchers have proposed various approaches to deal with the cold-start problem. Some studies attempt to employ dimensionality reduction technologies to condense the dataset by removing unrepresentative or insignificant users or items. For example, one statistical method, Principle Component Analysis (PCA), is considered one of the most commonly used dimensionality reduction approaches [6].

Researchers have also used *hybrid* recommender systems, which combine content-based and collaborative techniques to alleviate the cold-start problem [1, 8]. Pazzani [23] proposed a hybrid recommendation approach in which demographic information was used for making predictions for similar users. The author extracted features from users' home pages to build their content-based profiles. Some studies augmented users' ratings vectors with additional "pseudo" ratings predicted by content-based methods. Nguyen et al. [20] exploited the readily available user data (e.g., demographic information) to predict unobserved ratings and favorite genres for a new user by a rule-based induction process to automatically associate a better initial profile for a new user. Lekakos and Giaglis [16] utilized users' lifestyle information to group users and predict "pseudo" ratings for unobserved items for users to densify the data set. Another branch in this category is to generate representative "pseudo" users and items. Park et al. [22] proposed to utilize *filterbots*, specialized content-analysis agents that act as an additional user or item in a collaborative filtering community. For example, an action-movie agent (*filterbot*) was considered as a particular user who only likes action movie. As a result, the users whose ratings agree with some of the *filterbots*' ratings would be able to receive better recommendations.

Some researchers have attempted to explore the transitive interactions between users and items to augment the user-item matrix and make it meaningfully "dense" for recommendation purposes. Huang et al. [13] employed an associative retrieval framework and related spreading activation algorithms to explore the transitive associations among users. Their method was demonstrated to have superior performance when users' historical data set is sparse.

2.2 Personality in Recommender Systems

Personality is defined as a "consistent behavior pattern and intrapersonal processes originating within the individual" [4]. It is relatively stable and predictable. Research has shown that personality is an enduring and primary factor which influences human behaviors [14] and that there is a significant connection between personality and people's tastes and interests [15, 24]. It infers that people with similar personality will have similar interests and similar behavioral patterns. In the literature, human personality has been widely studied in the field of user modeling [19, 21]. The most commonly used human psychological aspects in recommender systems include personality traits [5, 12], demographic information [23], emotion [7], temperament [17] and lifestyle [16].

Even though it is still an emerging topic, this concept has already attracted increasing attention from both academia and industry. Lin and Mcleod [17] proposed a temperament-based filtering model incorporating human factors, especially human temperament, into the processing of an information recommendation service. Their model categorized the information space into 32 temperament segments based on the Keirsey's theory. By combining concept learning and content-based filtering techniques, their model tries to infer the optimal information units which best match both users' temperaments and interests. They empirically demonstrated that the temperament-based information filtering method surpassed the content-based one in both accuracy and effectiveness.

In our previous study [12], we developed a personality-based music recommender system based on the relations between human personality characteristics and musical preferences revealed by prior psychological studies. For example, extravert people are likely to prefer the upbeat and conventional music; individuals who are inventive, have active imaginations, value aesthetic experiences, consider themselves to be intelligent, tolerant of others, and reject conservative ideals tend to enjoy listening to reflective and complex music. In this system, a user only needs to answer a short personality questionnaire (10 items), and then the system will be able to predict which kinds of music the user might like. The results of a large-scale user study show that the proposed system is likely to be preferred by novice users with little music knowledge.

Recently, some websites have also attempted to use personality characteristics to build users' interest profiles and recommend items, especially for the items like music and movies, which have a strong association with human personality. For example, Whattorent.com, a movie recommender system, recommends movies based on users' personality measured by 20 scene-oriented personality questions. The detailed introduction can be found in [11]. Yobo.com is a Chinese music recommender website, providing personality quizzes to infer users' "music DNA" or users' musical preferences. In addition, some online commerce websites, such as Gifts.com, are emerging to make gift suggestions based on recipients' personality measured by personality quizzes, with the aim of facilitating the gift selection process.

2.3 Personality Acquisition

Prior studies on the acquisition of user personalities support the feasibility of adopting user personality information into

recommender systems. Personality can be acquired in both explicit and implicit ways [5]. The former measures a user's personality by asking the user to answer a list of designed personality questions, and these personality evaluation inventories have been well established in the psychology field [9]. The implicit approaches acquire user information by observing users' behavioral patterns, which can further be divided into two primary dimensions: behavior-orientated and content-orientated. Behavior-orientated methods focus on analyzing users' interaction behavior with the acquisition interfaces (e.g., playing games), while content-orientated methods leverage on the behavioral contents users have created in the past (e.g., blogs, review comments).

The main challenge is to identify the most efficient and compelling methods that can be adopted in practical systems. Dunn et al. [5] compared three personality acquisition variants and demonstrated that the explicit personality acquisition interface was preferred by most of their experiment participants in terms of satisfaction and ease of use, and was considered to be the most compelling method. While one of the two implicit methods under investigation failed in the prediction task, the other obtained high satisfaction and accuracy. They pointed out that since the implicit methods require less effort from users, they could facilitate the acquisition processes if properly designed based on game theory or other behavior theories. In our study, we adopted the explicit way to measure users' personality, i.e., personality quizzes, considering it is the most simplest and compelling approaches as demonstrated in [5]. Minamikawa et al. [19] proposed estimating individual personality by analyzing blog texts posted by individuals. The estimation is performed using the Multinomial Naïve Bayes classifier with the feature words that are selected based on information gain.

In practice, we see that adopting various personality quizzes in social websites (e.g., Facebook.com) for entertainment is a popular trend. These plentful social resources could potentially be utilized to obtain individuals' personality profiles and enhance existing recommender systems.

3. RATING-BASED COLLABORATIVE FILTERING

Figure 1 illustrates the overall recommendation model described in this paper, including rating-based and personality-based two parts. In this section, we are going to briefly introduce the framework of rating-based (particularly user-based) collaborative filtering approaches (i.e., the left side of Figure 1). In this framework, the recommendation process can be broken into two major steps: neighborhood formation and rating prediction. In the first step, CF systems identify k most similar users to form "neighbors" for a target user. The key concern is how to accurately measure the similarities between users. Various similarity measures have been proposed in the literature [1, 2]. The Pearson correlation coefficient is one of the most commonly adopted similarity measure. Accordingly, the proximity between user u and user v can be formalized as,

$$simr(u,v) = \frac{\sum_{i \epsilon I_u \cap I_v} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \epsilon I_u \cap I_v} (r_{u,i} - \bar{r}_u)^2 \sum_{i \epsilon I_u \cap I_v} (r_{v,i} - \bar{r}_v)^2}}, \quad (1)$$

where $r_{u,i}$ and $r_{v,i}$ are the ratings on item i given by user u and user v respectively. $\bar{r}_u$ and $\bar{r}_v$ are their mean ratings. I_u is the set of items that user u has rated. Similarly, I_v is the set of items that user v has rated. The correlation is calculated on a rating set consisting of the items rated by both users (i.e., $I_u \cap I_v$). The more overlapped ratings are used, the more reliable the correlation value. Therefore, to penalize similarity scores calculated on a rating set of small size, a modified similarity $simr'(u,v)$ was yielded in [18],

$$simr'(u,v) = \frac{min(|I_u \cap I_v|, \gamma)}{\gamma} * simr(u,v), \quad (2)$$

where γ is a pre-defined constant to normalize the influence of the overlapping size. In our experiments, we used a value of 5 for γ.

In the second step, the final prediction is computed by aggregating neighbor's ratings on the predicted item. More specially, the predicted unknown rating $\tilde{r}_{u,i}$ on item i can be calculated as,

$$\tilde{r}_{u,i} = \bar{r}_u + \kappa \sum_{v \in \Omega_u} simr(u,v) \times (r_{v,i} - \bar{r}_v), \quad (3)$$

where multiple κ serves as a normalizing factor and is usually selected as $\kappa = 1/\sum_{v \in \Omega_u} |simr(u,v)|$. Ω_u is the set of user u's neighbors. This recommendation technology framework is used as a basis of the design of the personality-based approaches proposed later.

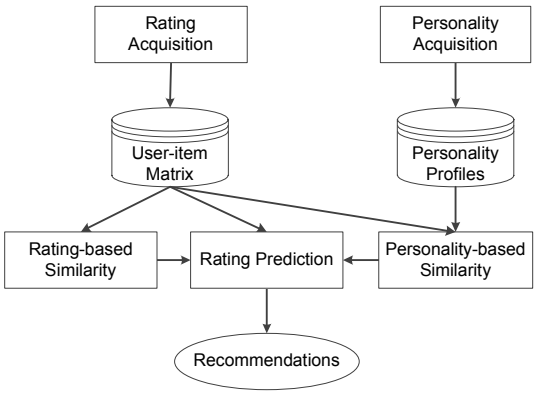


Figure 1. The overall proposed CF recommendation model.

4. PERSONALITY-BASED APPROACHES

In this section, we will first introduce a simple personality-based CF method, which merely replaces the ratings in Equation 1 with users' personality descriptors. Then, we propose two integration models with the aim of making more effective use of both personality and ratings to build user profiles. The overall framework is shown in Figure 1.

4.1 Personality-based CF

Instead of building the neighborhood for the target user based on sparse similarities due to few available ratings, we propose employing a personality-based neighborhood. People can be distinguished by their personalities, and people in the same personality segment are assumed to have similar behaviors or interests [4]. Therefore, it is feasible to consider the members in a personality-based neighborhood as reliable recommenders to each other.

We put a user's personality characteristics in a vector similar to the manner used in dealing with the rating data. More specifically, the personality descriptor of user u, $p_u = (p_u^1, p_u^2, \dots, p_u^n)^T$, is a n -dimension vector, and each dimension represents one characteristic in his/her personality profile, e.g., one of the personality traits. In our experiment, we adopt one of the most widely used and extensively researched personality models within psychology to build users' personality profiles. It is known as the Big Five Factor personality model [9]. It categorizes human personality traits into five bipolar dimensions: *Openness to Experience*, *Conscientiousness*, *Extroversion*, *Agreeableness*, and *Neuroticism* (also refers to as *emotion stability*). For details with regard to this model, you can refer to [9]. Along with the similarity computation in traditional CF methods, the personality similarity between two user u and v is computed using the Pearson correlation coefficient,

$$simp(u,v) = \frac{\sum_k (p_u^k - \overline{p_u})(p_v^k - \overline{p_v})}{\sqrt{\sum_k (p_u^k - \overline{p_u})^2 \sum_k (p_v^k - \overline{p_v})^2}}, \qquad (4)$$

where $\overline{p_u}$ and $\overline{p_v}$ refer to the average values of personality descriptor p_u and p_v respectively. In place of the rating-based similarity in Equation 3, we get a personality-based collaborative filtering approach, which merely depends on users' personality profiles, and can be used when the target user has not rated many items.

4.2 Linear Hybrid CF

One intuitive way to combine personality with ratings in the framework of CF is to linearly integrate them into one similarity measure. More specifically, the similarity between user u and v can be calculated using the formula,

$$sim(u,v) = \alpha * simr(u,v) + (1-\alpha) * simp(u,v), \qquad (5)$$

where $simr(u,v)$ represents the item-based similarity between user u and v, and $simp(u,v)$ is their personality-based similarity. α is a weight parameter that controls the percentage of the contribution the rating-based similarity makes into the final similarity measure. In our experiment, we set $\alpha = 0.8 * |I_u \cap I_v| / (|I_u \cap I_v| + 5)$ to automatically adapt to the sparsity level of a dataset. That is, when rating data is reliable enough to make prediction, α is weighted highly, and vice versa. We slightly incline towards the personality-based similarity measure by introducing a constant multiplier 0.8 to reduce the relative weight of the rating-based similarity. The value is chosen based on pre-trials.

4.3 Cascade Hybrid CF

The second proposed integration approach follows a cascade mechanism which is inspired by [16]. It utilizes the personality-based approach to make initial predictions on the unobserved ratings with the aim of densifying the user-item matrix. The number of neighbors who are used to predict the "pseudo" rating is denoted by β. A new augmented rating vector, consisting of the original ratings provided by users and the ratings predicted by the CF approach merely relying on personality (described in section 4.1), is introduced for each user. That is, the rating $r'_{u,i}$ in the new rating vector is computed as,

$$r'_{u,i} = \begin{cases} r_{u,i} & \text{if rating for item } i \text{ has been provided by user } u \\ \tilde{r}_{u,i} & \text{otherwise} \end{cases}$$

Table 1. Statistical characteristics of experimental datasets.

		DM Dataset	Last.fm Dataset	DM +Last.fm Dataset
Personality	Has Personality Info.	Yes	No	Partly Yes
Data Set Size	No. of Users	111	119	230
	No. of Items	640	599	640
	No. of Ratings	2,485	5,657	8,142
Rating Sparsity	Sparsity Level (SL)	97.5%	92.2%	94.5%
	Average No. of Ratings per User	22.4	47.5	35.4
	Average No. of Ratings per Item	3.9	9.4	12.7

where $\tilde{r}_{u,i}$ is predicted by using the ratings from the users with similar personalities with the user u. Then, the traditional CF prediction process is applied on the "denser" user-item matrix to produces the final recommendations, refining the initial predictions.

5. EXPERIMENTAL STUDY

In order to investigate the performance of the personality-based CF methods, we conducted a series of experiments by comparing them with the traditional user-based CF method (introduced in Section 3) in both scenarios of sparse data and new users. Through our empirical study, we attempt to understand whether, when and how personality profiles can work effectively on making predictions when cold-start is the issue.

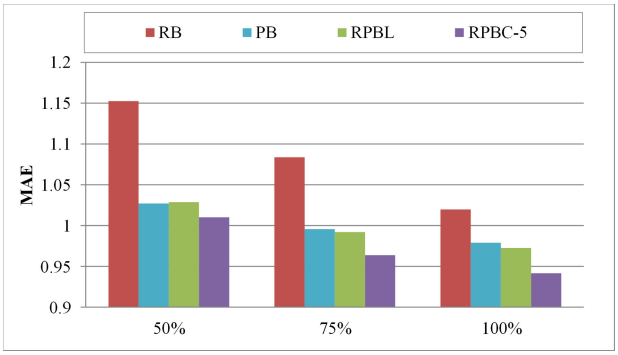
5.1 Experimental Data

Currently, most available test datasets only contain user ratings (e.g., the MovieLens dataset [1]), item attribute contents (e.g., IMDB[2]), or user demographic information. To the best of our knowledge, no dataset containing both users' personality information and ratings data is freely available. In this study, we utilized a music dataset collected in our previous work [12], which is referred here as the DiscoverMusic (DM) dataset. We filtered out users who rated less than 20 songs. The reduced data set consists of 111 users with their personality profiles measured by the Big Five Model, 640 songs that were rated by at least one of the users, and a total of 2,485 ratings.

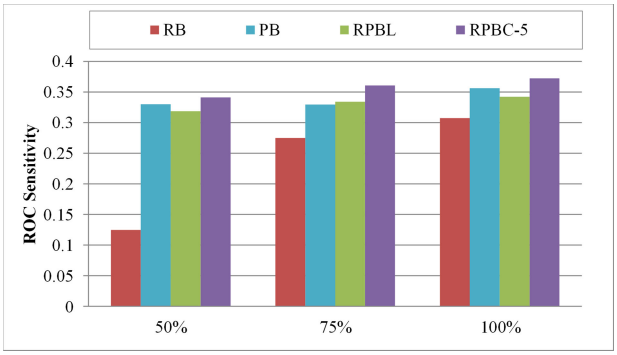
To enlarge our DM dataset, we adopted part of data distributed by Last.fm. This dataset represents the listening habits for nearly 1,000 users until May 5, 2009, and contains 19,150,868 listening records for 992 users. We made a mapping from listening frequencies of songs for each user to ratings on a 5-point rating scale consistent with the DM dataset. More specifically, if a user listened to a song twice to 4 times, we assign it a rating of 3 (neutral). A song gets a rating of 4 (like), if it was listened to 5-7 times by this user. For more than 7 times, it is reasonable to say that the user truly likes this song and to assign it a rating of 5 (like very much). We did not consider the songs that were only listened to once in a user's log, since it is fairly likely that the user was just exploring in this case. In addition, we removed the records whose songs are not in the DM dataset and the users with rating counts

[1] http://www.movielens.org

[2] http://www.imdb.com



(a) MAE



(b) ROC sensitivity

Figure 2. Prediction performances in the scenario of sparsity datasets. (RB: rating based CF, PB: personality based CF, RPBL: rating-personality based linear hybrid approach, RPBC-5: rating-personality based cascade hybrid approach with β = 5.)

lower than 15. The statistical characteristics of these datasets are summarized in Table 1. The sparsity level (SL) is computed as [26]. That is, $\text{SL} = 1 - \frac{\text{\#nonzero entries}}{\text{\#total entries}}$.

5.2 Methodology

To evaluate the recommendation performance, we employed the leave-one-out (LOO) cross-validation method. The dataset is split to construct a *training set* and a *test set*. The Last.fm dataset was only used for training. The cross-validation method replicated the error estimation process by treating each user in the DM dataset as the test user, and the others combined with the users in the Last.fm dataset are used for training.

We then used each of the rating entries for each user as the tested item, and the remaining entries for training. How to sample the training data depends on the experiment designs. In the experiment on the sparsity problem, we randomly sampled 50%, 75% and sampled 100% of the ratings provided by each user in the training set to simulate the scenarios with varying sparsity levels. In the experiment on the new user issue, we employed the Given2, Given5 and Given10 experimental protocols. That is, we randomly selected 2, 5 or 10 ratings from the test user's rating data except the tested one to form the training set.

5.3 Evaluation Metrics

We used two major categories of metrics for evaluating the prediction accuracy in our study. The first one is statistical accuracy metrics, which evaluates the accuracy of a predictor by comparing predicted values with user-provided values. To evaluate this, we used *Mean Absolute Error* (*MAE*) which is one of the most prominent and broadly adopted predictive accuracy metric in the information retrieval and recommender community [10]. MAE measures the average absolute deviation between a recommender system's predicted rating $\tilde{r}_i$ and a true rating r_i for item i, and is computed as,

$$\text{MAE} = \frac{\sum_{i=1}^{n} |\tilde{r}_i - r_i|}{n},$$

where n is the number of tested items. A lower *MAE* value means better prediction performance.

The other is decision-support accuracy metrics which measure how well predictions help users in selecting high quality items (i.e., ones of be interest), also referred to as classification accuracy metrics. We used *Receiver Operating Characteristic (ROC) sensitivity* in our experiments, similar to [8]. *ROC* is the extent to which an information filtering system can distinguish between good and bad items. ROC sensitivity measures the probability with which a system accepts a relevant item (defined as the item liked by the tested user). It can be formulated as,

$$\text{ROC sensitivity} = \frac{\sum_{\text{u}} \frac{|HIT_u|}{|REL_u|}}{n},$$

where the relevant set REL_u contains all relevant (i.e., liked) items in user u's test set. HIT_u is a hit set including the accepted relevant items (i.e., the items in REL_u are correctly predicted to be relevant by the system). n is the size of the tested users. $|\cdot|$ denotes the size of item sets. A ROC sensitivity value of 1.0 indicates that the recommendation algorithm is able to predict all relevant items correctly, whereas a value of 0.0 indicates that it predicts any of the relevant items as bad. To measure this metric, we defined the rating 3.5 as the cut-off threshold on a 5-point rating scale from 1 to 5 in our experiment. That is, all ratings which are greater than 3.5 were considered relevant, all others being irrelevant (i.e., disliked).

5.4 Results Analysis

In this section, we will show the empirical results on the recommendation performance of the CF variations incorporating personality information, the personality-based CF (PB), the linear hybrid (RPBL) and the cascade hybrid (RPBC), in the simulated scenarios of sparse datasets and new users. The rating-based CF approach (RB) was used as the baseline.

5.4.1 Optimal Values of Parameters

To investigate the influence of the neighbor size when predicting "pseudo" ratings as the input of the traditional CF approach, we compared the recommendation performances in three settings: 5 neighbors (N5, β = 5), 10 neighbors (N10, β = 10) and 15 neighbors (N15, β = 15). Our analysis uses ANOVA with the Bonferroni procedure for multiple comparison statistics. In this paper, a difference is considered as statistically significant if it reaches the 95% confidence level (i.e., p-value < 0.05). The same criterion is also used in the analysis of the experimental results below.

The results show that the performance measured by MAE does not have significant differences among these three settings under

Table 2. Overall Recommendation Quality Comparison.

Scenarios	MAE				ROC Sensitivity			
	RB	PB	RPBL	RPBC-5	RB	PB	RPBL	RPBC-5
50%	1.153	**1.027 (10.9%)**	**1.029 (10.7%)**	**1.010 (12.4%)**	0.125	**0.330 (164.9%)**	**0.319 (155.7%)**	**0.341 (173.7%)**
75%	1.084	**0.996 (8.1%)**	**0.992 (8.4%)**	**0.964 (11.1%)**	0.275	**0.330 (19.8%)**	**0.340 (21.4%)**	**0.361 (31.1%)**
100%	1.020	**0.979 (4.0%)**	**0.973 (4.6%)**	**0.942 (7.7%)**	0.307	0.356 (15.9%)	0.342 (11.3%)	**0.372 (21.2%)**
Given 2	1.144	**1.109 (3.1%)**	**1.112 (2.8%)**	**0.942 (17.7%)**	0.045	**0.377 (738.4%)**	**0.374 (730.5%)**	**0.372 (727.4%)**
Given 5	1.129	**1.025 (9.2%)**	**1.025 (9.2%)**	**0.942 (16.6%)**	0.205	**0.363 (76.7%)**	**0.353 (72.0%)**	**0.372 (81.5%)**
Given 10	1.074	**0.994 (7.4%)**	**0.989 (7.8%)**	**0.942 (12.3%)**	0.300	**0.342 (14.0%)**	**0.345 (14.9%)**	**0.372 (24.0%)**

Note: The values inside the parentheses are the improvement percentages compared to the baseline RB approach in each scenario. The significant differences are presented in boldface (p-value < 0.05).

the CF neighbor size k ranging from 5 to 100 with the increment of 5. Concerning the measure ROC sensitivity, the only significant difference is observed between N5 and N15 ($p = 0.012 < 0.05$) when $k = 5$ (i.e., small CF neighbor size). That is, there is no statistically significant difference from the influence of the neighbor size in predicting "pseudo" ratings when k is not small. Therefore, a value of 5 is a reasonable setting for parameter β and was adopted in our study. The corresponding cascade integration approach is denoted as RPBC-5. Consequently, we ran our following comparison experiments under the settings of CF neighbor size k ranging from 5 to 100 with the increment of 5 for each scenario, and the best results are chosen as the representatives for that scenario setting.

5.4.2 Performance in Sparse Datasets

For evaluating the recommendation quality in sparse datasets, we simulated three training datasets with varying sparsity degrees by randomly sampling 50%, 75% and 100% of the ratings from each user in the training set as training data. We performed 5 runs for the scenarios of 50% and 75%. The average results are shown in Figure 2. A pairwise t-test was used for comparison statistics. The comparison results are summarized in Table 2. To save space, a performance result in **boldface** is significantly different from that given by the baseline RB approach in the same scenario and its improvement percentage is listed below the absolute value.

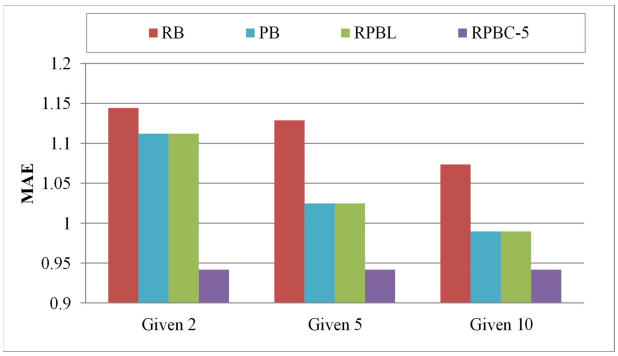
The results clearly indicate that the cascade hybrid approach (RPBC-5) outperforms other collaborative filtering approaches under study on both accuracy metrics. On accuracy metric MAE, it achieves 1.01, 0.96 and 0.94 on average in the sparsity settings of 50%, 75% and 100% respectively. On ROC sensitivity, the cascade hybrid approach RPBC-5 obtains 0.34, 0.36 and 0.37 on average in the sparsity settings for 50%, 75% and 100% respectively. The differences between RPBC-5 and other CF methods (RB, PB and RPBL) are significant on both MAE and ROC sensitivity metrics (p-value < 0.05). In particular, its maximal improvements are 12.4% and 174% on MAE and ROC sensitivity respectively when comparing to the scenario of sampling 50% of the training data. This provides strong evidence that the cascade hybrid approach can effectively alleviate the sparsity problem in collaborative filtering systems.

The results also show that the personality-based CF (PB) and the linear hybrid CF (RPBL) achieve similar performance in all settings. Their performances fell between those of RPBC-5 and RB, but are much closer to the RPBC-5. We suspect that the similar performance was because of the sparsity of our dataset. The ratings cannot contribute effectively on prediction. In contrast to RB, PB and RPBL obtain significant improvements on MAE in all settings. On ROC sensitivity, the significant differences can only be found in the settings using 50% and 75% of the training data. On the other hand, the rating-based CF performs poorly in the sparse datasets.

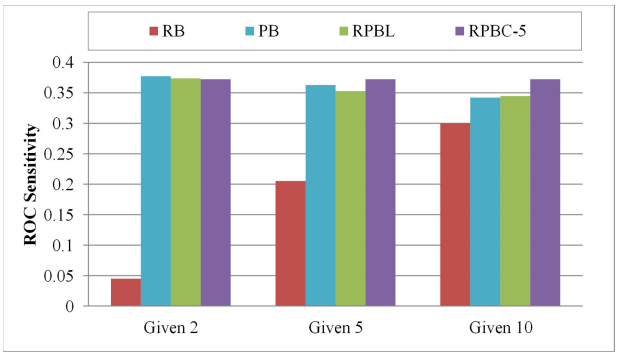
5.4.3 Performance for New Users

To evaluate the performance of various methods on the new user problem, we simulated a group of scenarios by randomly sampling 2 (Given 2), 5 (Given 5) and 10 (Given 10) ratings from the target user's profile for training. We performed 5 runs for each setting. The results are shown in Figure 3. For each approach, only the optimal results are displayed in each experimental setting. Similarly, we used the pairwise t-test to validate the significance of improvements in contract to the baseline RB approach. The comparison details are listed in Table 2.

Along with the results above, the cascade hybrid approach (RPBC-5) is illustrated to outperform other collaborative filtering approaches on both accuracy metrics. On average, its best performance achieves 0.94 and 0.37 on MAE and ROC sensitivity respectively in all settings. The differences between RPBC-5 and other CF methods (RB, PB and RPBL) are significant on MAE in all settings. With respect to the metric ROC sensitivity, the significant differences between RPBC-5 and RB can be found in all settings, but those between RPBC-5 and PB/RPBL are only found in the setting of Given 10. In particular, compared with the baseline rating-based CF, RPBC-5 has 17.7%, 16.6% and 12.3% improvements on MAE in the setting of Given 2, Given 5 and Given 10 respectively. Regarding ROC Sensitivity, there are 727%, 81.5% and 24.0% improvements respectively. This shows



(a) MAE



(b) ROC sensitivity

Figure 3. Prediction performances in the scenario of new user. (RB: rating based CF, PB: personality based CF, RPBL: rating-personality based linear hybrid approach, RPBC-5: rating-personality based cascade hybrid approach with β = 5.)

that the cascade hybrid approach can effectively work on the user cold-start problem in collaborative filtering systems.

Similarly, the personality-based CF (PB) and the linear hybrid CF (RPBL) achieve similar performance in this experiment, and both obtain significant improvements on MAE and ROC sensitivity in all settings, compared to the baseline rating-based CF approach (RB), which fails to make predictions for new users.

5.5 Discussion

We assessed the performance of the proposed personality-based CF methods on both situations of sparse dataset and new user issues. The results are promising and positively support that incorporating personality information into the collaborative filtering framework indeed effectively addresses the cold-start problem. In particular, the cascade hybrid approach outperforms the other approaches under study in both scenarios in terms of prediction accuracy (MAE) and classification accuracy (ROC sensitivity). Their performance improvements are statistically significant.

The linear hybrid approach and personality-based approach lag slightly behind. Considering MAE, they both significantly outperform the traditional rating-based CF in both scenarios. In terms of ROC sensitivity, the performance differences are also significant, except in the setting of 100% in the scenario of sparsity. Thus, these two methods perform well in the cold-start setting as well. Furthermore, we can infer that the pure rating-based method can perform better as the user-item matrix becomes denser. In our experiment, the linear hybrid approach and personality-based approach always perform similarly. This might be because our experimental settings simulated the cold-start scenarios where the predication capability of ratings is limited, so personality somehow dominates the prediction process. However, more studies are needed to verify this premise.

With respect to the computational efficiency issue, the average running time needed to generate recommendations using the cascade hybrid approach is 2.1 times more than what was needed with the linear hybrid approach, 2.4 times more than the rating-based approach and 5.3 times more than the pure personality-based approach. The cascade hybrid approach requires more time due to its extra computation for the "pseudo" ratings.

Considering the tradeoff between accuracy and efficiency, we can derive the following suggestions when designing personality-merged CF recommender systems. Even though the cascade approach is not the most efficient method compared to others, we still recommend employing it in a CF recommender system when its user historical data is severely insufficient, since it has the best performance on addressing the cold-start issue. We further suggest running the "pseudo" rating calculation offline. As the available rating information increases, the system can adopt the linear hybrid method due to its capability on addressing the cold-start problem and computation efficiency. The personality-based approach is more suitable for the situation where the target user has not offered any ratings or left any available behavioral traces.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed incorporating users' personality information into the collaborative filtering framework to address the cold-start problem. We presented three methods incorporating human personality into user modeling: the first is based on users' personalities alone; the second based on a linear combination of both personality and rating information; and the third using a cascade mechanism. To evaluate their effectiveness, we conducted an experimental study comparing the proposed approaches with the traditional rating-based CF in both cold-start scenarios: sparse data sets and new users. Our results indicate that the proposed CF variations significantly outperform the traditional rating-based CF as measured by MAE and ROC sensitivity, especially the cascade hybrid approach.

Our future work includes testing our methods with more datasets (both scale and types) and product domains to be able to generalize the findings in this paper. We would like to understand further the performance difference between the personality-based method, and the linear hybrid methods. We would also like to understand users' privacy concerns when disclosing personality information and whether the benefits of receiving more useful recommendations outweigh the risk of disclosing this information.

7. ACKNOWLEDGMENTS

We thank the EPFL and the ministry of education of the People's Republic of China for supporting the reported research work.

8. REFERENCES

[1] Adomavicius, G. and Tuzhilin, A. 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. on Knowl. and Data Eng. 17*, 6, 734-749.

[2] Ahn, H.J. 2008. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. *Inf. Sci. 178*, 1, 37-51.

[3] Billsus, D. and Pazzani, M.J. 1998. Learning Collaborative Information Filters. In *Proceedings of the Fifteenth International Conference on Machine Learning* (1998). Morgan Kaufmann Publishers Inc., 657311, 46-54.

[4] Burger, J.M. 2010. *Personality*. Wadsworth Publishing, Belmont, CA.

[5] Dunn, G., Wiersema, J., Ham, J. and Aroyo, L. 2009. Evaluating Interface Variants on Personality Acquisition for Recommender Systems.*User Modeling, Adaptation, and Personalization*, Houben, G.-J., McCalla, G., Pianesi, F. and Zancanaro, M., eds. Lecture Notes in Computer Science 5535, Springer Berlin / Heidelberg, 259-270.

[6] Goldberg, K., Roeder, T., Gupta, D. and Perkins, C. 2001. Eigentaste: A Constant Time Collaborative Filtering Algorithm. *Inf. Retr. 4*, 2, 133-151.

[7] Gonzalez, G., Rosa, J.L.d.l., Montaner, M. and Delfin, S. 2007. Embedding Emotional Context in Recommender Systems. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop* (2007). IEEE Computer Society, 1547669, 845-852.

[8] Good, N., Schafer, J.B., Konstan, J.A., Borchers, A., Sarwar, B., Herlocker, J. and Riedl, J. 1999. Combining collaborative filtering with personal agents for better recommendations. In *Proc. Conf. Am. Assoc. Artificial Intelligence (AAAI-99)* (Orlando, Florida, United States, 1999). AAAI, 315352, 439-446.

[9] Gosling, S., Rentfrow, P. and Swann, W. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*. *37*, 6, 504-528.

[10] Herlocker, J.L., Konstan, J.A., Terveen, L.G. and Riedl, J.T. 2004. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst. 22*, 1, 5-53.

[11] Hu, R. and Pu, P. 2009. A comparative user study on rating vs. personality quiz based preference elicitation methods. In *Proceedings of the 13th international conference on Intelligent user interfaces* (Sanibel Island, Florida, USA, 2009). ACM, 1502702, 367-372.

[12] Hu, R. and Pu, P. 2010. A Study on User Perception of Personality-Based Recommender Systems.*User Modeling, Adaptation, and Personalization*, De, B., Kobsa, A. and Chin, D., eds. 6075, Springer Berlin / Heidelberg, 291-302.

[13] Huang, Z., Chen, H. and Zeng, D. 2004. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Trans. Inf. Syst. 22*, 1, 116-142.

[14] Jung, C.G. 1971. *Psychological Types*. Princeton University Press, Princeton, N.J.

[15] Kemp, A.E. 1996. *The Music Temperament: Psychology and Personality of Musicians*. Oxford University Press, New York.

[16] Lekakos, G. and Giaglis, G.M. 2006. Improving the prediction accuracy of recommendation algorithms: Approaches anchored on human factors. *Interact. Comput. 18*, 3, 410-431.

[17] Lin, C.-H. and McLeod, D. 2002. Exploiting and Learning Human Temperaments for Customized Information Recommendation. In *Internet and Multimedia Systems and Applications (IMSA 2002)* (Kaua'i, Hawaii, USA, 2002). IASTED/ACTA Press, 218-223.

[18] Linden, G., Smith, B. and York, J. 2003. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE*. *7*, 1, 76-80.

[19] Minamikawa, A. and Yokoyama, H. 2011. Blog tells what kind of personality you have: egogram estimation from Japanese weblog. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (Hangzhou, China, 2011). ACM, 1958856, 217-220.

[20] Nguyen, A.-T., Denos, N. and Berrut, C. 2007. Improving new user recommendations with rule-based induction on cold user data. In *Proceedings of the 2007 ACM conference on Recommender systems* (Minneapolis, MN, USA, 2007). ACM, 1297251, 121-128.

[21] Nunes, M.A.S.N. 2009. *Recommender Systems based on Personality Traits: Could human psychological aspects influence the computer decision-making process?* VDM Verlag, Berlin.

[22] Park, S.-T., Pennock, D., Madani, O., Good, N. and DeCoste, D. 2006. Naïve filterbots for robust cold-start recommendations. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (Philadelphia, PA, USA, 2006). ACM, 1150490, 699-705.

[23] Pazzani, M.J. 1999. A Framework for Collaborative, Content-Based and Demographic Filtering. *Artif. Intell. Rev. 13*, 5-6, 393-408.

[24] Rentfrow, P.J. and Gosling, S.D. 2003. The do re mi's of everyday life: the structure and personality correlates of music preferences. *J Pers Soc Psychol*. *84*, 6 (Jun), 1236-1256.

[25] Resnick, P. and Varian, H.R. 1997. Recommender systems. *Commun. ACM*. *40*, 3, 56-58.

[26] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. 2000. Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM conference on Electronic commerce* (Minneapolis, Minnesota, United States, 2000). ACM, 352887, 158-167.