Using Personality Information in Collaborative Filtering for New Users

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

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

ABSTRACT

Recommender systems help users more easily and quickly find products that they truly prefer amidst the enormous volume of information available to them. Collaborative filtering (CF) methods, making recommendations based on opinions from "most similar" users, have been widely adopted in various applications. In spite of the overall success of CF systems, they encounter one crucial issue remaining to be solved, namely the cold-start problem. In this paper, we propose a method that combines human personality characteristics into the traditional rating-based similarity computation in the framework of user-based collaborative filtering systems with the motivation to make good recommendations for new users who have rated few items. This technique can be especially useful for recommenders that are embedded in social networks where personality data can be more easily obtained. We first analyze our method in terms of the influence of the parameters such as the number of neighbors and the weight of rating-based similarity. We further compare our method with pure traditional ratings-based similarity in several experimental conditions. Our results show that applying personality information into traditional user-based collaborative filtering systems can efficiently address the new user problem.

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, Experimentation

Keywords

Recommender System, User-based Collaborative Filtering, Personality, User Similarity, New User

1. INTRODUCTION

The ubiquity of the web brings an explosive increase of accessible information. Recommender systems have emerged as an intelligent information filtering tool to help users effectively identify information items of interest from such an overwhelming set of choices and provide personalized services [25]. At the same time, recommender systems are considered to be a critical tool for boosting sales in e-commerce websites [2]. Therefore, variations of recommendation algorithms have been widely studied and incorporated in a wide range of online commercial websites [1].

Collaborative filtering (CF) is one of the most successful recommendation technologies. The basic idea behind this method is that it gathers the opinions of other users who share similar interests with a target user (referred to as the "*active user*") and assists this active user to identify items of interest based on these *neighbors*' opinions. These social information filtering approaches automate a process of "word-of-mouth" recommendations [29]. Compared to other recommendation technologies (e.g., content-based filtering), CF provides some prominent advantages to information filtering: (i) the capability to filter items whose content is not easily analyzed by automated processes; (ii) the ability to provide serendipitous recommendations; and (iii) support for social factors by taking into account the interests of like-minded users [10, 17]. Consequently, CF has been becoming popular in both academy and industry fields with great speed.

Despite the overall success of CF systems, they suffer one serious limitation, namely the *cold-start* problem [1]. It includes two major aspects: *new user* and *new item.* Before a recommender system can present a user with reliable recommendations, it should know about this user's preferences/interests, most likely from a sufficient number of behavior records, e.g., ratings or log-archives [11]. Therefore, a new user, having few records in a system, normally cannot get satisfied recommendations. Similar to the new user problem, new items which have not been rated could not be recommended, which is referred to as new item problem. In our study, we will focus on the new user problem. It is a key issue that determinants the initial success of e-retailers, since more accurate recommendations for new users could make these new users stay rather than pushing them to the competitors' sites.

To address the new user problem, most studies present hybrid recommender systems that combine both content information and ratings data [22, 27, 28] to circumvent the problem, where content-based similarity is used for new users or new items. In most currently used systems, demographic information is used as users' attributes to calculate similarity among users. For example, Pazzani [23] uses the gender, age, area code, education, and employment information of users in the restaurant recommendation application.

Recently, many studies have tried to incorporate human personality into recommender systems [6, 12, 13, 16]. Studies show that personalities influence human decision making process and interests [24]. Drawing on the inherent inter-related patterns among users' personalities and their interests/behaviors, personality-based recommenders were developed to provide personalized services. Empirical studies further revealed a significant user acceptance of such recommenders [12]. Additionally, research has suggested that human personality characteristics have the potential ability to lessen the cold start problem associated with commonly adopted collaborative filtering recommender systems [6]. However, few works have empirically verified this hypothesis.

As mentioned in [6], one apprehension of few researchers venturing into the personality-based recommender systems might be due to perceived difficulty in obtaining personality characteristics. However, it can be foreseen that social network sites have the capability of facilitating the personality acquisition processes. The primary functionality of Social Web is to help people socialize or interest with each other throughout the World Wide Web. During the forming of social community, personality is considered as one of the users' identities. On the other hand, personality profiles can be helpful to suggest more connections or enhance the relationships among friends, such as seeking friends with similar personality. Due to the particularity of the social web, people have a strong interest to do personality tests and share such information with their friends. One evidential observation is that there are a range of personality test applications at facebook.com and an amount of users are involved.

The main objective of our work is to investigate the performance of utilizing personality information in user-based CF systems to address the new user problem. We first propose a personality-based similarity measure and a general model which takes both the traditional rating-based similarity and the personality-based similarity into account to find the neighbors of an active user. We compared our method with the pure rating-based CF systems in different cold-start settings. The results positively support the advantage of the personality-based similarity in improving recommendation quality, at least in the case of sparse dataset, for new users.

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

2. RELATED WORK

In this section, we briefly view some of the research literature related to our work.

2.1 New User Problem

A crucial concern of CF is the new user problem, which refers to difficulties encountered by recommender algorithms when a new user enters into a system. Various approaches have been proposed in previous studies. Most of them leverage hybrid recommendation approaches, which combine content meta-data and ratings to circumvent this problem [1, 7, 22]. Ahn [2] addresses this problem by proposing a new similarity measure called PIP (Proximity-Impact-Popularity) which utilizes domain specific interpretation of user ratings by considering proximity, impact and popularity factors when comparing two ratings. Different from other solutions, this method attempts to make better use of the limited amount of ratings data so as to address the new user problem and improve recommendation performance. However, it cannot address the difficulty of recommending items for completely new users with no ratings.

Some studies follow the idea of associating any new user with a stereotype among a set of predefined ones from learning processes [20]. For example, Pazzani [23] utilizes user's demographic information (e.g., age, gender, education) to identify stereotypes of users that like a certain object. When any new user comes, he/she is associated with one stereotype based on his/her demographic data and gets tailored recommendations. In this work, the demographic information is taken from new users' homepages automatically without extra user effort.

Alternatively, other researchers have suggested that further improvements dealing with cold start in CF systems can be achieved by leveraging user characteristics, specially detailed characteristics [15]. Even though few studies have been done on this topic, existed research has shown its promise [e.g., 15, 18, 20]. Particularly, personality is considered as a consistent behavior pattern and intrapersonal processes originating within the individual [24]. It is relatively stable and predictable. Therefore, it is reasonable to suppose that it is possible to address the cold start problem and possibly improve prediction in current CF systems by incorporating personality characteristics.

2.2 Personality-based Recommender Systems

Recently, personality characteristic information has been integrated into recommendation techniques. Lin and Mcleod [16] proposed a temperament-based filtering model incorporating human factors, especially human temperament, into information recommendation service. They empirically demonstrated that the accuracy and effectiveness of the temperament-based information filtering system surpassed those of pure content-based filtering. Nunes et al. [21] proposed one personality-based recommender by incorporating personality traits into user profiles for the social matching applications. Their system was designed under the scenario of the "Elections for President in France". The system recommends one president candidate to a user by matching the candidates' social reputation profiles (summarized personality profiles from all participates) with the personality profile of the user's ideal president. Their method obtains a high prediction accuracy.

Hu and Pu [13] developed a personality-based music recommender system based on the psychological findings of the correlations between human personality characteristics and musical preferences. For example, 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 (e.g., blues, jazz, classical and folk). Their results from an in-depth user study revealed the user acceptance of this personality-based system under the scenarios of making recommendations for active users and their friends. This work mainly emphasizes on user subjective perception of this emerging personality-based music recommender system.

In this paper, we focus on investigating how personality characteristics can be integrated in the commonly used CF framework and how its performances is compared to the traditional user-based CF systems, especially relative to the new user problem.

3. USER-BASED COLLABORATIVE FILTERING

User-based Collaborative Filtering has been studied in-depth during the last decades. It is one of the most successful and widely used recommendation technologies, owing to its compelling simplicity and excellent quality of recommendations. It assumes that if a group of users have similar interests in their previous behaviors, they will express similar interests on other more items in the future. Its basic idea is to find a group of users, who have a history of agreeing with an active user (i.e., they either gave similar ratings or purchased similar items). Once a neighborhood of users is formed, opinions from these neighbors are aggregated to produce recommendations for the active user.

Various algorithms for user-based CF can be grouped into two classes: *memory-based* (or heuristic-based) and *model-based* [1]. Memory-based algorithms essentially are heuristics that make rating predictions based on the entire collection of previously rated items. In contrast to memory-based methods, model-based algorithms use the collection of ratings to learn a *model*, which is then used to make rating predictions. In this paper, we concentrate on memory-based algorithms. In order to simply, the term CF mentioned in the following represents user/model-based CF, except particular specification.

3.1 Similarity Measure

The most important step in CF recommender systems is computing the similarity between users which is used to form a proximity-based neighborhood between a target user and a number of like-minded users. The neighborhood finding process is in fact the model-building or learning process for a recommender system algorithm. Various approaches have been used to compute the similarity $simr(u,v)$ between user u and user v [1, 2, 26]. The letter r means the similarity is calculated based on rating data, distinguishing from the personality-based similarity measure proposed later. The most commonly used similarity calculation method is *Pearson correlation coefficient* [2, 26]. More specifically, the proximity between user u and v is measured as,

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

where $r_{u,i}$ denotes the rating value which user u gave to item i. $\bar{r}_u$ is the average rating value of user u. I_u is the set of items that user u has rated.

In order to penalize similarity scores that are based on a small number of overlapping items which reflects a lack of confidence, a modified similarity score $simr'(u,v)$ is yielded as follows [22]:

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

where γ controls the required number of overlapping items between two users for similarity calculation. We adopt this modified correlation-based similarity score in our evaluation experiment and set $\gamma = 5$.

3.2 Rating Prediction

To predict the value of unknown rating $r_{u,i}$, CF systems make an aggregation of the ratings from some other (usually, the k most similar) users for item i. More specially, the predicted unknown rating $\tilde{r}_{u,i}$ can be calculated as,

$$\tilde{r}_{u,i} = \mathrm{aggr}_{v \in \Omega_u} r_{v,i}, \quad (3)$$

where Ω_u denotes the set of k users that are the most similar to user u and who have rated item i. Various aggregation strategies are designed and applied based on different applications, such as, averaging the ratings, or using similarities as weights while aggregating. In this paper, we adopt a more general aggregation function,

$$\tilde{r}_{u,i} = \overline{r_u} + \kappa \sum_{v \in \Omega_u} sim(u,v) \times (r_{v,i} - \overline{r_v}), \quad (4)$$

where multiple κ serves as a normalizing factor and is usually selected as $\kappa = 1/\sum_{v \in \Omega_u}|sim(u,v)|$, and $\overline{r_u}$ is the average rating of user u. This aggregation function takes into account the fact that different users may use the rating scale differently. For example, user u thinks a rating 3 in a 5-point rating scale means that this item is ok, while user v might thins a rating 3 represent a negative score. Therefore, the weighted sum uses the deviations from the average ratings instead.

4. PERSONALITY-BASED SIMILARITY

4.1 Personality-based Similarity Measure

As described above, traditional collaborative filtering systems find "neighbors" based on the ratings they gave in common. However, in practice, the item-user matrix R is always sparse. That is, the number of ratings already obtained is usually significantly small compared to the number of the ratings that need to be predicted. It is crucial to effectively predict ratings from a small number of examples. If one user has rated some items which few users have been rated, it will be not possible to find a sufficient number of neighbors and to make good recommendations. This situation especially happens when a new user enter into a system with few and no rating records. In order to address this problem, we proposed a new similarity measure method based on users' personality characteristics.

We treat users' personality characteristics as a vector. For user u, his/her personality descriptor $p_u = (p_u^1, p_u^2, \ldots, p_u^n)^T$ is a n-dimension vector, and each dimension stands for one characteristic consisting in his/her personality. For example, if users' personalities are measured by Big Five Factor model [8] which describes human personality along major five traits, p_u is a five-dimension vector and each dimension corresponds to one of the five personality traits (details see Section 4.2). Consequently, the personality similarity between two user u and v can be computed as the Pearson correlation coefficient of their personality descriptors.

$$simp(u,v) = \frac{\sum_k (p_u^k - \overline{p_u})(p_v^k - \overline{r_v})}{\sqrt{\sum_k (p_u^k - \overline{p_u})^2 \sum_k (p_v^k - \overline{p_v})^2}}. \quad (5)$$

To consider both rating-based and personality-based similarity measures at the same time, we combine them together and intuitively generate the following model,

$$sim(u,v) = \alpha * simr'(u,v) + (1-\alpha) * simp(u,v), \quad (6)$$

where $simr'(u,v)$ represents the item-based simialrity between user u and v, and $simp(u,v)$ represents the personality-based similarity. α is a weight parameter which controls the percentage of rating-based similarity contributed into the final similarity measurement.

4.2 Personality Model and Measurement

One of the most widely used and extensively researched personality models within psychology is known as Big Five Factor personality model [5, 8]. This model categorizes human personality traits into five bipolar dimensions:

- *Openness to Experience*: appreciation for art, emotion, adventure, unusual ideas, curiosity, and variety of experience.
- *Conscientiousness*: a tendency to show self-discipline, act dutifully, and aim for achievement; planned rather than spontaneous behavior.
- *Extroversion*: energy, positive emotions, urgency, and the tendency to seek stimulation in the company of others.
- *Agreeableness*: a tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others.
- *Neuroticism*: a tendency to experience unpleasant emotions easily, such as anger, anxiety, depression, or vulnerability.

In literature, several rating instruments have been developed to measure the Big-Five dimensions, from well-established multi-item instruments (e.g., NEO Personality Inventory, or Revised NEO-PI-R [5]) to extremely brief ones (e.g., 5-Item Personality Inventory (FIPI) or 10-Item personality Inventory (TIPI) [8]). Even though the comprehensive instruments have the superior capability of assessing finer facets within each personality dimension, they require more user effort to accomplish it than the brief ones. In the context of online system, users would be unlikely to dwell at the website for a long time to complete a multi-item questionnaire [21]. In our implementation, therefore, TIPI developed by Gosling et al. [8], is utilized. In each Big Five dimension, there are two items which attributes in the two poles of this dimension. Each item consists of two descriptors, separated by a comma, for example, “Extraverted, enthusiastic” in the positive direction of extraversion dimension. Each item was rated on a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree). The TIPI personality acquisition process takes about 2-3 minutes to complete.

5. EMPIRICAL ANALYSIS

In this section, we first describe the experiment setup, including experiment design, evaluation metrics and used dataset. Then, we present experimental results and discussions. Our main goal is to explore the possibilities of combining different similarity measures to formulate an efficient recommendation algorithm.

5.1 Experiment Design

In this experiment, we try to compare the predictive accuracy of the three similarity measures on a sparse dataset. The dataset is split into a *training set* and a *test set*. We adopt the all-but-one protocol considering the high sparsity of our dataset [4]. That is, we randomly select one of the rating entries of each user to be the tested item, and use the remaining entries for training. Consequently, as much data as possible from each test user can be used to train the recommenders. It is meaningful when we evaluate the effects of the parameters in our general model.

5.2 Evaluation Metrics

We adopt two kinds of evaluation metrics in our evaluations. One takes into account the ability of accurately predict ratings in the case of individual item-by-item recommendations. The other considers the recommendation filtering task as a two-class classification, like or dislike. In this case, evaluation metrics measures how effectively positive and negative items are classified correctly.

5.2.1 Predictive Accuracy Metric

First, we try to measure how close the predicted ratings generated by various algorithms are to user real ratings. Mean absolute error (MAE) is the most prominent and broadly adopted predictive accuracy metric in information retrieval and recommender community [9]. It is calculated as the average of the differences between predicated ratings $\tilde{r}_i$, and users' real ratings r_i. More specially, MAE is formulated as:

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

where n is the number of tested items.

Even though it is important to measure the ability of accurately predicting ratings for a recommendation algorithm, we can observe that the majority of users only care about whether a recommended item is interesting or not, rather than exact rating values. In these cases, filtering is considered as a binary classification and the predicted rating errors might not really reflect the filtering performance. For example, if a rating of 3.5 is considered as the cut-off between good and bad, a one-star error that predicts a 4 as 5 and one that predicts a 3 as 4 are totally different. The former makes no difference to users, while the latter erroneously classifies a bad item as a good item. Therefore, we utilize decision-support accuracy metrics as well in our evaluation.

5.2.2 Decision-Support Accuracy Metrics

Decision-support accuracy metrics (also called classification accuracy metric [9]) evaluate a recommender system based on whether it makes correct or incorrect decision outcome, i.e., whether it correctly predicts that an item might be interesting for an active user [7]. To evaluate recommendations generated by different similarity methods, we use two metrics widely used in the information retrieval (IR) community namely, *recall (also called sensitivity)* and *specificity*. However, we slightly modify the definition of them to fit in our all-but-one experiment design. We first defined the rating 3.5 as our cut-off threshold on a 5-point rating scale from 1 to 5. That is, all ratings which are greater than 3.5 are considered as “good” (or “relevant”), others as “bad” (or “irrelevant”).

We define the collection of all tested items for user u in l (l = 20, in our experiment) runs of our experiments as *check set* (in order to distinguish from the *test set* used in the evaluation of top-N recommendation list [19, 26]) T_u, which contains all items to be predicted for this user. Furthermore, T_u can be divided into two sets: relevant set REL_u containing all relevant items and irrelevant set $IRREL_u$ containing all irrelevant items. All *hits*, items in the set of REL_u are predicted to be good as well, are include in the *hit set* HIT_u. All *avoids*, items in the set of $IRREL_u$ are predicted to be bad as well, are include in the *avoid set* AV_u.

We define *recall* as the ratio of hit set size to the relevant set size. More specifically, considering the average of all n tested users, it is formularize as:

$$recall = \frac{\sum_u \frac{|HIT_u|}{|REL_u|}}{n},$$

where n is the number of users tested. A recall value of 1.0 indicates that the commendation algorithm was able to retrieval all relevant items, whereas a recall vale of 0.0 indicates that the recommendation algorithm was not able to recommend any of the relevant items.

Along with recall, *specificity* is defined as the proportion of irrelevant items which are correctly identified. More specifically, it is formalized as:

$$specificity = \frac{\sum_u \frac{|AV_u|}{|IRREL_u|}}{n}.$$

A specificity value of 1.0 indicates that the all irrelevant items are successfully eliminated from recommended items, whereas a specificity vale of 0.0 indicates that all irrelevant items are incorrectly classified as good ones and might be recommended to users.

5.3 Dataset

Currently, most available test datasets only contain user rating records (e.g., the MovieLens dataset[1] and the EachMovie dataset[2]) or contents of items (e.g., IMDB[3]). To the best of our knowledge, none of the available datasets contains both users' personality information and their ratings. Therefore, we can only conduct our experiment on a music data set that we have accumulated in our previous study [13]. In that study, users were asked to answer a personality questionnaire based on the Big Five Model and rate the recommended songs. This dataset includes 1,581 songs (1956 – 2009) covering 14 genres in four musical preferences. We only considered users who rated 20 or more songs. Therefore, the reduced data set include 113 users and 646 songs that were rated by at least one of the users. Each user has a 5-dimensional personality descriptor along five traits as well. We compute the *sparsity level* of the dataset as [26],

$$sparsity\ level = \frac{1-\#\text{non entries}}{\#\text{total entries}}.$$

The statistical characteristics of this data set are shown in Table 1.

Table 1. Statistical characteristics of our ratings data set.

Size	#users	113
	#items	646
	#ratings	2479
Sparsity	sparsity level	96.604%
	Ave. #ratings per user	21
	Ave. #ratings per item	3
Rating Distribution	#ratings on the value of 1	254
	#ratings on the value of 2	470
	#ratings on the value of 3	747
	#ratings on the value of 4	585
	#ratings on the value of 5	423

5.4 Experiment Results

In presenting our evaluation experimental results, we firstly investigate the influences of various parameters in the general model. Then, we compare the performances of rating-based similarity (RBS), personality-based similarity (PBS) and their hybrid (RPBS) in different start-up settings.

5.4.1 Influence of Model Size

The traditional user-based recommendations are computed using a model that utilizes the ratings from k most similar users to predict the unknown items for the active user. To evaluate the sensitivity of the different algorithms on the value of k, we performed an experiment in which we let k take the values from 5 to 110 in increments of 5. Note that these results were obtained using the value of parameter $\alpha = 0.5$ for the hybrid similarity measure. The results are shown in Figure 1.

Figure 1(a) shows the MAE results regarding different neighbor sizes. As we can see, MAE results for rating-based CF reaches the minimal value when $k = 40$ and keeps it while the number of neighbors increases. It happens in traditional user-based CF when the item-user matrix R is sparse. Even thought the threshold of the number of neighbor is designed to be high, the actual neighbors can be found in the dataset is few due to the small overlap of ratings. However, the personality-based similarity has not this restraint. The similarity can be calculated only if the personality characteristics vectors of two users are known. Therefore, the MAE values for the other two algorithms can keep decreasing in the tested range until the number of neighbors reaches 90 where the improvement become gently.

Figure 1(b) and (c) respectively shows the recall and specificity results under different neighbor sizes. Along with the results on MAE, the quality for the rating-based CF method increases in the beginning, but remains on one value when the number of neighbors reaches 50 and 40 respectively, while other two methods still increase their recall and specificity until the size increases to the point around 90. Therefore, in our later experiments, we assign the number of neighbors to be 90.

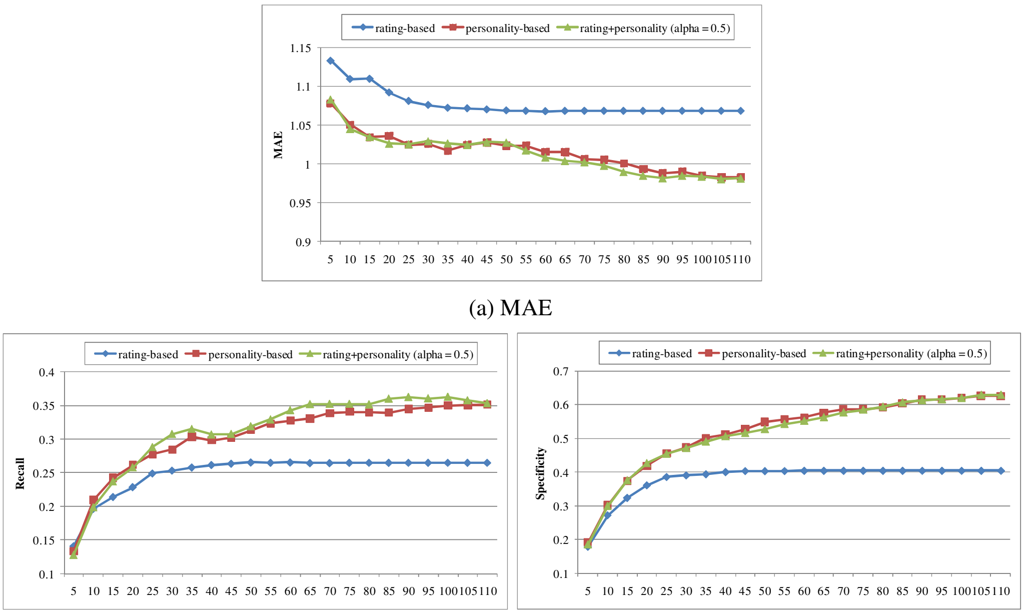
5.4.2 Influence of Weight

One parameter in the combined similarity model, α, is used to control the extent to which the rating-based similarity measurement can contribute. To study the sensitivity of the recommendation algorithm on this parameter, we performed a sequence experiments in which we varied α from 0.0 (pure personality-based CF) to 1.0 (pure rating-based CF) in increments of 0.1. Figure 2 shows the results of MAE, recall and specificity for different values of α. The higher recall and specificity values and lower MAE value indicate better performance. Note that these results were obtained by assigning the number of neighbor $k = 90$.

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

[2] http://www.research.compaq.com/SRC/eachmovie

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



(a) MAE

(b) Recall

(c) Specificity

Figure 1. The influence of the number of neighbors.

The results in Figure 2 show that the hybrid similarity can achieve the best recommendation quality on MAE, recall and specificity when the value of α is around 0.5. However, as we can see, the performance is not highly sensitive to the changes of α in this setting, as long as α is less than 0.9. It might because we choose a big size of neighbors for displaying best performance for each method. However, the number of actual neighbors in rating-based method cannot reach this value due to the sparsity of the dataset. In this case, personality-based similarity will play a dominate role no matter which value α is chosen to be except the value "1". In our later experiments, we assign the parameter α to 0.5 to investigate the recommendation performance on different start-up settings.

5.4.3 Influence of Training Size

The main objective of our work is to evaluate the performance of personality-based similarity in addressing new user problem. In this section, we will compare the three different similarity measures in different start-up settings with training sets of varying degrees of sparsity. As our previous experiment setup, we randomly choose one item for each one user as the test item, and others are used as training data. Differently, we further randomly select parts of the ratings from the original set with the percentage of μ to form our training sets of different degree of sparsity. We increase μ from 10% to 100% in increments of 10%, so that we have 10 start-up settings for each μ. When sampling, we keep all new training sets have the same distribution in each rating scale (from 1 to 5) as the original set, as did in [3]. We perform 20 runs for each test item. The average results are shown in Figure 3.

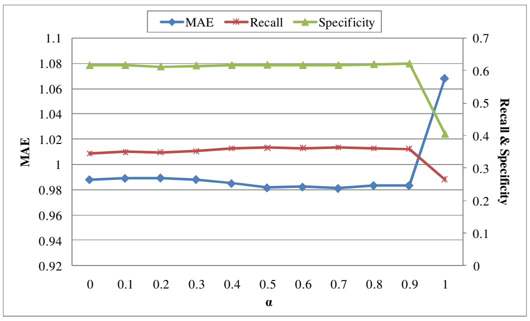
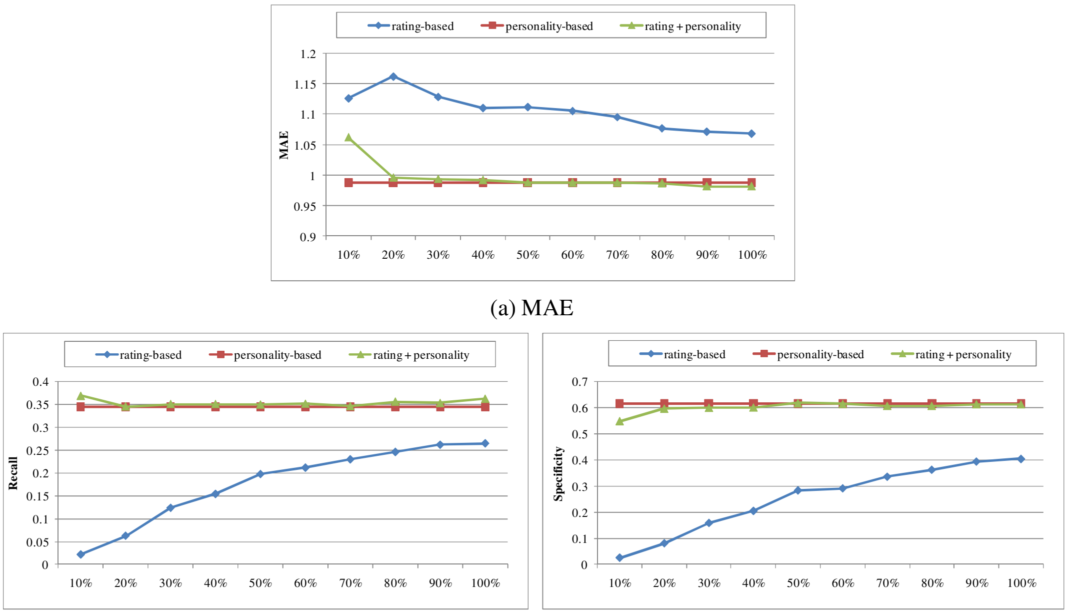


Figure 2. The recommendation quality on different values of parameter α from 0.0 (personality-based) to 1.0 (rating-based).

The result shows superior performance of the personality-based similarity and the hybrid similarity measures over the traditional rating-based one in the situation of sparse user-item matrix, especially when the sizes of training sets are small. To specifically show the improvements, we present the results under the settings of $\mu = 50\%$ and $\mu = 100\%$, considering RBS as the baseline. When the training sets include 50% of all the ratings, there is an improvement of 11% for both PBS and RPBS on MAE. The increases of recall are 74% and 77% for PBS and RPBS respectively, and the increases of specificity are 117% and 111%. When training sets include 100% of the overall ratings (In this setting, there are 20 training ratings on average for each user), we get a 7% decrease on MAE for PBS and an 8% decrease for RPBS. At the same time, there are improvements of 30% and 37% on recall for PBS and RPBS respectively and 52% on specificity for both PBS and RPBS.

Along with the results shown in the previous sections, the performance of personality-based similarity is somewhat similar to that of the hybrid method. It is because we choose a large size of neighbors and our testing dataset is sparse.



(a) MAE

(b)Recall

(c) Specificity

Figure 3. The comparison of recommendation quality in different start-up settings. X-axis represents of percentage of training set sizes.

5.5 Limitations

Firstly, the experimental data set is relatively small. It only contains 113 users, 646 songs and 2479 ratings. The statistical results might be sensitive to the distribution of the used dataset. More experiments using larger data sets are needed to verify the findings of this work. Secondly, we didn't evaluate the performance of each similarity measure using a dataset whose user-item matrix has a higher density. In that situation, users have more co-rated items so that the rating-based similarity measure could work more effectively. It is interesting and valuable to compare these three similarity measures in such situations so as to find out how personality-based similarity could work there. Finally, we only used a music dataset in our current evaluation experiment. It is still unclear how results would become if these similarity measures are used in other domains. Music is a special product domain, since the relationship between musical preferences and human personalities has been revealed. It is still an open issue whether the personality-based similarity measure can perform as well in other domain as it does in music domain. More works are needed to deal with these issues in the future.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed to incorporate personality-based similarity into user-based CF recommender systems to address the new user problem. We experimentally evaluated the recommendation quality in terms of mean absolute error (MAE), recall and specificity by comparing the personality-based similarity measurement, the traditional rating-based one and their hybrid. Our results showed that both personality-based similarity and the hybrid scheme lead CF recommender systems to generate more accurate recommendations than the traditional rating-based one in a sparse music dataset. In our test setting with 100% training data (on average, 20 ratings per user), in contrast to RBS, PBS has 7% improvement on MAE, 30% improvement on recall and 52% improvement on specificity. With respect to RPBS, there are at least 8% improvement on MAE, 37% on recall and 52% on specificity. In addition, our results from different start-up settings positively support that the personality-based similarity measure can effectively address the new user problem adhere to the traditional user-based CF recommender systems.

However, since the dataset used in our current experiment is relatively small, more evaluation studies in bigger datasets are needed to verify our findings. Additionally, it is meaningful to investigate the performance of the personality-based similarity measure in a dense dataset where users have many co-rated items and to explore the possibility of generalizing to other item domains.

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. Knowledge and Data Eng*., 17, 6(2005), 734-749.

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

[3] Basu, C., Hirsh, H., and Cohen, W. 1998. Recommendation as Classification: Using Social and Content-Based Information in Recommendation. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)*. C. Rich, and J. Mostow, Eds. AAAI Press 1998.

[4] Breese, J. S., Heckerman, D., and Kadie, C.1998. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*. G. F. Cooper, and S. Moral, Eds. Morgan-Kaufmann, San Francisco, Calif., 43–52.

[5] Costa, P.T. and McCrae, R.R. 1992. NEO PI-R Professional Manual. In: Psychological Assessment Resources, Odessa, FL.

[6] Dunn, G., Wiersema, J., Ham, J., and Aroyo, L.2009. Evaluating Interface Variants on Personality Acquisition for Recommender Systems. In: *Houben, G.J., McCalla, G., Pianesi, F., Zancanaro, M. (eds.) User Modeling, Adaptation, and Personalization*. LNCS, vol. 5535, pp. 259--270. Springer, Heidelberg.

[7] 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 *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*, J. Hendler, and D. Subramanian, Eds. AAAI Press, Menlo Park, Calif., 439–446.

[8] Gosling, S. D., Rentfrow, P. J., and Swann, Jr. W.B.2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37, pp. 504--528.

[9] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl J. T.2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Trans. Information Systems*, vol. 22, no. 1, pp. 5-53.

[10] Herlocker, J. L., Konstan, J. A., Borchers, A., and Riedl, J.1999. An algorithmic framework for performing collaborative filtering. *In Proc. of SIGIR*.

[11] Hofmann, T.2004. Latent semantic models for collaborative filtering. *ACM Trans. Info. Syst.*, vol 22(1):89-115.

[12] 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*, pp. 367—372.

[13] Hu, R. and Pu, P. 2010. A Study on User Perception of Personality-Based Recommender Systems. In: *P. De Bra, A. Kobsa, and D. Chin (Eds.): UMAP 2010*, LNCS 6075, pp. 291–302.

[14] John, O.P. and Srivastava, S.1999. The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In: Pervin, L., John, O.P. (eds.) Handbook of Personality: Theory and Research, 2nd edn., pp. 102–138. Guilford, New York.

[15] Lam, X.N., Vu, T., Le, T.D. and Duong, A.D. 2008. Addressing Cold-Start Problem in Recommendation Systems. In: ICUIMC 2008, pp. 208–211. ACM Press, New York.

[16] Lin, C. and McLeod, D.2002. Exploiting and Learning Human Temperaments for Customized Information Recommendations. *IMSA*, 218-233.

[17] Linden, G., Smith, B., and York, J.2003. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, Jan/Feb.:76-80.

[18] Lekakos, G. and Giaglis, G.M. 2006. Improving the Prediction Accuracy of Recommendation Algorithms: Approaches Anchored on Human Factors. Int. with Comp. 18, 410–431.

[19] Karypis, G.2001. Evaluation of Item-Based Top-N Recommendation Algorithms, In *Proceedings of the tenth international conference on Information and knowledge management*, October 05-10, Atlanta, Georgia, USA.

[20] Nguyen, A., Denos, N. and Berrut, C. 2007. Improving New User Recommendations with Rule-Based Induction on Cold User Data. In: RecSys 2007, pp. 121–128. ACM Press, New York.

[21] Nunes, M. A. S. N., Cerri, S. A., and Blanc, N. 2008. Improving recommendations by using Personality Traits. In: *International Conference on Knowledge Management. I-KNOWS08*, Graz-Austria.

[22] Park, S.-T., Pennock, D. M., Madani, O., Good, N., and DeCoste, D.2006. Naive filterbots for robust cold-start recommendations, in: *Proceedings of KDD'06*, ACM, Philadelphia, PA, USA

[23] Pazzani, M.1999. A Framework for Collaborative, Content-Based, and Demographic Filtering. *Artificial Intelligence Rev.*, pp. 393-408, Dec.

[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. *Journal of Personality and Social Psychology*, 84, 1236—1256.

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

[26] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J.2000. Analysis of recommendation algorithms for E-commerce. In *Proceedings of the 2nd ACM Conference on Electronic Commerce (EC'00).* ACM, New York. 285–295.

[27] Salter, J. and Antonopoulos, N. 2006. CinemaScreen recommender agent: combining collaborative and content-based filtering, *IEEE Intelligent Systems* 21, 35 – 41.

[28] Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. 2002. Methods and metrics for cold-start recommendations, in: *Proceedings of SIGIR'02*, ACM, Tampere, Finland, pp. 253 – 260.

[29] Shardanand, U. and Maes, P.1995. Social information filtering: Algorithms for automating "word of mouth", in *Proceedings of ACM Conference on Human Factors and Computing Systems*, pp. 210–217, Association of Computing Machinery, New York.