A Diversity Adjusting Strategy with Personality for Music Recommendation

Feng Lu
Delft University of Technology
Delft, the Netherlands
F.Lu-1@student.tudelft.nl

Nava Tintarev
Delft University of Technology
Delft, the Netherlands
n.tintarev@tudelft.nl

ABSTRACT

Diversity-based recommender systems aim to select a wide range of relevant content for users, but diversity needs for users with different personalities are rarely studied. Similarly, research on personality-based recommender systems has primarily focused on the 'cold-start problem'; few previous works have investigated how personality influences users' diversity needs. This paper combines these two branches of research together: re-ranking for diversification, and improving accuracy using personality traits. Anchored in the music domain, we investigate how personality information can be used to adjust the diversity degrees for people with different personalities. We proposed a personality-based diversification algorithm to help enhance the diversity adjusting strategy according to people's personality information in music recommendations. Our offline and online evaluation results demonstrate that our proposed method is an effective solution to generate personalized recommendation lists that not only have relatively higher diversity as well as accuracy, but which also lead to increased user satisfaction.

CCS CONCEPTS

• **Information systems → Recommender systems**;

KEYWORDS

Recommender Systems, Diversity, Personality, Music Recommendation, Re-ranking

1 INTRODUCTION

As recommender systems have moved beyond accuracy for evaluation, metrics such as diversity and novelty have been proposed to evaluate the quality of recommender systems [20]. Such research [11, 18] also help to address the 'filter bubble' problem [21]. Research on the diversity metric have contributed to the emergence of diversity-based recommender systems, which endeavor to achieve an optimal balance between accuracy and diversity [25]. In addition, researchers suggest that there exists a connection between people's stable personality traits and their tastes and preferences [24]. The idea of personality-based recommender systems is thus proposed.

However, current research on personality and diversity based recommender systems are mostly separated [7]. In many diversity-based recommender systems [9, 31, 33], researchers usually set a fixed balance degree between accuracy and diversity for all users. Adjusting to diversity needs for users with different personalities are rarely studied. Similarly, research on personality-based recommender systems [14, 28] has utilized personality information to improve the calculation of user similarity in recommendations so as to mitigate the 'cold-start problem'. This paper combines these two branches of research together.

To address this gap, we investigate how personality information can be used to adjust the diversity degrees for people with different personalities in music recommender systems. Our research questions are therefore:

- *RQ1: Is there an underlying relationship between people's personality and their needs for recommendation diversity in Music domain?*
- *RQ2: What is the effect (on diversity and accuracy) of adjusting the diversity degrees in Music Recommender Systems based on users' personality information?*

First, to address RQ1 we conduct a pilot user study to investigate whether there exists a relationship between users' personality information and their diversity needs on music preference (Section 5.1). A relation model is built based on the pilot study results (Section 5.2). To address RQ2, we proposed a personality-based diversification algorithm referred to this relation model (Section 4). Our proposed diversification method adjusts the diversity degrees adaptively in music recommendations according to users' distinct personality information. Both offline (Section 5.3) and online studies (Section 5.5) evaluate the efficiency and effectiveness of our proposed algorithm. We conclude with limitations and suggestions for future work in Sections 7 and 8.

2 RELATED WORK

Existing research on personality-based and diversity-based Recommender Systems are mostly separated [7]. We discuss first research on diversity-based recommender systems, and then related research on personality-based recommender systems.

2.1 Diversity-based Recommender Systems

Many current diversity-oriented recommender systems [9, 31, 33] adopt a fixed strategy to adjust the diversity degree for all users, in which they usually pre-defined a score function balancing the diversity and accuracy with a parameter λ and re-ranked the generated recommendation list according to the calculated scores. For instance, Ziegler et al. [33] proposed the topic diversification approach towards balancing recommendation lists, which is a heuristic algorithm based on taxonomy similarity to increase the recommendation diversity. To balance the accuracy of suggestions and the user's extent of interest in specific topics, they defined a weighting parameter to control the impact of two ranking lists, one ranking the items that are similar to user's attribute-based preference and the other ranking the items in reverse. Vargas et al. [31] also proposed a similar re-ranking diversification method based on sub-profiles of users, in which they also used a parameter λ to

control the balance between the initial ranking score and diversity score. Building on this work, Di Noia et al adjusted the diversity function adaptively according to users' diversity inclination [9], which is calculated as Entropy from user preferences. The balancing parameter λ in their objective function is still fixed for all users, which means that although their diversity function might be adjusted properly, the recommendation balance between similarity and diversity is still the same for all users. All these work successfully increased the diversity in recommendations, while they rarely consider that different users with different personalities may have different diversity needs.

2.2 Personality-based Recommender Systems

Recent works have explored the relationship between personality traits and user preferences in recommender systems [15, 28, 29]. Studies also show that personalities influence human decision making process and interests for music and movies [6, 13, 24], which implies that personality information should be considered if we want to deliver personalized recommendations. While, in another aspects, since users' attitudes towards new or diverse experiences vary considerably [26], personality can also be considered as a key aspect when incorporate novelty and diversity into recommendations, which means that the degree of diversity in presenting recommended items can also be personalized.

In contrast, most of the research work [13, 15] in personality-based recommender systems is designed to improve the user similarity calculation in recommendations to address the cold-start problem. Diversity degrees are usually the same for all users. They rarely consider that different users might also possess different attitudes towards the diversity of items, which means that personality information can also be useful when adjusting diversity degrees in Recommender Systems. As some recent studies have already shown that personality can affect people's needs for diversity degrees for items either in movie recommendations [6, 32] or book recommendations [26], people with different personalities may also need recommendations with different diversity degrees for music recommendations.

2.3 Addressed Research Gap

This paper addresses the gaps in the two research branches, by combining them. The main contributions of our research are threefold:

- We investigated the relation between users' personality factors and their diversity needs on music preference and found that there exist certain positive correlations between these two factors.
- We proposed a personality-based re-ranking diversification algorithm, which can adaptively set different diversity levels for user based on their personalities in music recommendations.
- We evaluated this strategy in both online and offline studies, which suggest that this approach is effective for improving both diversity, accuracy, and also user satisfaction.

To the best of our knowledge, in the music domain, we are the first to conduct such systematic user study on the correlation between personality and users' diversity needs. In the movie domain, Wu and Chen et al. [6, 32] conducted a similar research on movie recommendations. However, our research are dissimilar w.r.t. domain difference and the algorithms that were applied.

3 DIVERSITY AND PERSONALITY IN RECOMMENDATIONS

Before we proceed into our research steps, we first explain two key concepts as they are defined in our paper: diversity and personality. In this section, we mainly focus on the diversity metric, the personality model and the corresponding extraction methods we applied in our research.

3.1 Diversity

Diversity is usually considered as the inverse of similarity, which refers to recommending a diverse set of items that are different from each other to users [20]. This concept has been introduced into the field of recommender systems as one of the possible solutions to address the over-fitting problem and to increase users' satisfaction.

In this paper, we focus on Intra-List Diversity (ILD). Research that is focused on the definition and evaluation of the Intra-List Diversity starts with Bradley and Smyth [3], who define the diversity as the averaged pairwise distance (dissimilarity) between all items in the recommendation set, which can be calculated as follows:

$$ILD(R) = \frac{\sum_{i=1}^{n} \sum_{j=i}^{n} (1 - Similarity(c_i, c_j))}{n * (n-1)/2} \quad (1)$$

where $c_1..c_n$ are items in a set of recommendation list and R is the recommended list. Other metrics are also raised, such as Vargas et al.'s ILD metric [30] or Gini-coefficient measurement [10]. In this work we mainly refer to Equation 1 for the diversity metric.

3.2 Personality

Personality represents people's differences in their enduring emotional, interpersonal, experiential, attitudinal and motivational styles [17]. For the last few decades, a number of personality models and acquisition methods have been proposed. In this research, we mainly focus on the Big-Five Factor Model and the explicit acquisition methods (specifically, Ten-Item Personality Inventory).

Personality Model. We adopted one of the most commonly used personality model called the Big-Five Factor Model (FFM) [19], which defines personality as five factors: Openness to Experience (O), Conscientiousness (C), Extroversion (E), Agreeableness (A), and Neuroticism (N). Usability of this model in recommender systems can be found in Recommender Systems Handbook [27].

Extraction Method. Current acquisition methods for personality can be classified into two groups: explicit methods (using questionnaires) and implicit methods (extract personality from social networks). We used the explicit method considering that it is more accurate than the implicit methods [27]. Specifically, we adopted a short personality test called Ten Item Personality Inventory (TIPI) [12] since it needs less time for users to finish. In TIPI, each personality factor of FFM is assessed by two questions. For instance, extraversion is assessed by 'Extraverted, enthusiastic' and 'Reserved, quiet'. Each question (ten in total) can be rated from 1 to 7, which can be then mapped into five personality factor scores. The scores

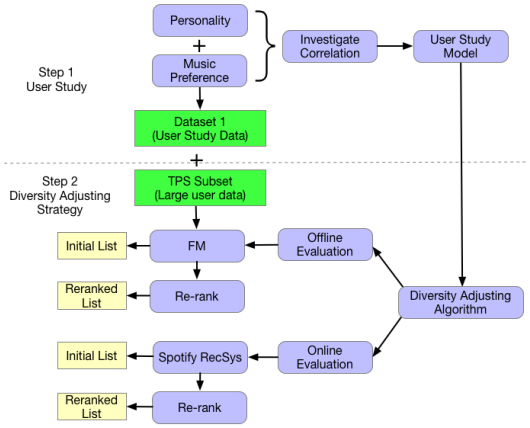

Figure 1: Research Steps

of each factor can be further mapped into four different personality levels: *Low*, *Medium Low*, *Medium High*, and *High* [12].

4 PERSONALITY-BASED RE-RANKING DIVERSIFICATION

In this section, we discuss the core of our work: the personality-based diversification algorithm, which consists of a) an objective function, and b) personality related parameters.

In a later section we will describe the pilot user study in which we identified the relationship between users' personality information and their diversity needs on music preference (Section 5.1). The results of that pilot study inform the parameters for both a) and b) above. Figure 1 outlines our overall research methodology, including offline and online studies to evaluate our algorithm. .

Normally, the recommendation process of a recommender system can be divided into two steps: first the system generates the predicted values for all unrated items for each user and secondly these items are sorted in descending order according to their predicted values. While in order to improve the diversity degrees of the recommendations, we use re-ranking as an improvement to the second step. We borrow the idea of the Topic Diversification method presented in Ziegler et al.'s work [33]. Specifically, greedy heuristics are used in our work, which have been demonstrated to be efficient and effective [9, 33]. The diversification algorithm is shown in Algorithm 1.

This greedy algorithm will iteratively select an item from the original list O (generated directly from a recommender system) and then puts it at the end of the current re-ranked list R until the size of R meets a size N (N=10 in our case) and the re-ranking process is complete. The core of the algorithm lies in the objective function (line 4, Algorithm 1) which controls the balance between similarity and diversity, so that at each re-ranking step, the algorithm can pick the next item that minimizes the objective function as the next item to be placed at the end of the current diversified re-ranked list. The target list is a re-ranked list with N top-ranked items (called Top-N items). In order to perform the re-ranking algorithm to make the re-ranked list diverse enough, the size of the input list should

Algorithm 1 The Diversification Algorithm to generate the re-ranked list R from the original list O

Input: (Original Recommendation List O (length: 5N), target list size N, personality-related parameters λ, θ_1, θ_2, ..., θ_n)
Output: Top-N re-ranked list R
1: $R(1) \Leftarrow O(1)$
2: **while** |R| < N: **do**
3: $Div_{overall}(c, R) = \sum_{i=1,2,..,n} \theta_i * Div_i(c, R)$
4: $c^* = \text{argmin}_{c \in O \setminus R} Obj(c, R) = Sim(c, P) * (1 - \lambda) + \lambda * Div_{overall}(c, R)$
5: $R = R \cup \{c^*\}$
6: $O = O \setminus \{c^*\}$
7: **end while**
8: **return** R

be much larger than the final re-ranked list (with N items). In our algorithm, we use 5N items for the input list.

The balancing parameter λ in the objective function (line 4, Algorithm 1) is controlled by personality factors in our algorithm. To adjust the diversity degrees more flexibly, we also introduce parameters θ_1, θ_2, .., θ_n to control the computation of the overall diversity. All of these two kinds of parameters (λ, θ_1, θ_2,..., θ_n) are affected by the personality factors.

4.1 Objective Function

The core of the algorithm lies in the re-ranking objective function in line 4 (Algorithm 1), which is referred from the Maximal Marginal Relevance (MMR) [4]:

$$Obj(c, R) = Sim(c, P) * (1 - \lambda) + \lambda * Div_{overall}(c, R) \quad (2)$$

The left part of the function $Sim(c, P)$ considers the similarity aspect of the item c to users' initial interests P. In our work, we computed the similarity values as the rank of item c in the final list according to their predicted ratings sorted in the descending order. We did not use the predicted ratings directly considering that such predicted values may not be available for all recommender systems (e.g. Spotify). Thus, our $Sim(c, P)$ function becomes:

$$Sim(c, P) = Rank(c, O) \quad (3)$$

where $Rank(c, O)$ represents the rank of item c in the original recommendation list O generated by some recommendation algorithm.

The other part of the function $Div_{overall}(c, R)$ defines the overall diversity degree of the item c compared with the items so far selected in the re-ranked list R. Here, we define the overall diversity as the weighted combination of several diversity degrees for different attributes (e.g. track attributes like artists, genres in music recommendation). As shown in line 3 (Algorithm 1), the diversity function is defined as follows:

$$Div_{overall}(c, R) = \sum_{i=1,2,\ldots,n} \theta_i * Div_i(c, R) \quad (4)$$

where n represents the total number of attributes we used for computing the overall diversity $Div_{overall}(c, R)$, θ_i represents the weight for each attribute diversity degree. $Div_i(c, R)$ represents the different diversity degrees for different attributes, which is defined

Table 1: Mapping from Personality Factor Level to Personality Related Parameters

Personality Factor Level	Low	Medium Low	Medium High	High
$\lambda/\theta_1/\theta_2/.../\theta_n$	0.2	0.4	0.6	0.8

as ILD (equation 1). In our experiment, we used three attributes (n=3) that are closely correlated with the personality factors found in our pilot user study, which we will introduce later in Section 5.2.

The function of the control parameter λ will be explained in Section 4.2 and at the end of Section 5.2.

4.2 Personality Related Parameters

For now, we have defined our similarity function and diversity function. But we still have not incorporated the personality information. In our algorithm, the influence of the personality factors is exerted on the parameters (λ, θ_1, θ_2,..., θ_n) in our objective function.

Parameter λ affects the balance between similarity and diversity directly, thus it controls the degree of overall diversity needs. Parameters θ_1, θ_2,..., θ_n control the specific attribute diversity degrees accordingly. As mentioned in Section 3, each personality factor can be divided into four levels: **Low**, **Medium Low**, **Medium High**, and **High**. For each possible correlation between personality factors and overall/attribute diversity needs, we define their mapping function as follows in Table 1.

For θ_1, θ_2,..., θ_n, we take one more computation step: normalization. Thus, the final $\theta_1/\theta_2/.../\theta_n$ are computed as follows:

$$\theta_i = \frac{\theta_i}{\sum_{j=1,2,\ldots,n} \theta_j}, \quad i = 1, 2, .., n \tag{5}$$

Noted that, in order to conduct the mapping, we need to know the correlation between each personality factor and users' overall/attribute diversity needs beforehand. Parameter λ is decided by the personality factor that has a positive correlation with the overall diversity needs (e.g. in our case, it is Emotional Stability). While parameters θ_1, θ_2,..., θ_n are decided by the personality factors that are correlated with the attribute diversity needs. The specific corresponding personality factors for each parameter for the mappings will be shown in Section 5.2.

5 EXPERIMENT

Following our research steps in Figure 1, we first conducted a pilot study to explore the possible correlation between users' personality factors and the diversity needs on their music preferences. Our diversity adjusting strategy (in Section 4) is thus based on the findings in the pilot study. To evaluate the efficiency and effectiveness of our proposed personality-based diversification algorithm, we conducted both offline and online evaluation. For the page limitation, we will discuss our pilot study and offline evaluation briefly. Results for both pilot study and offline evaluation will be shown in this section. We will show the results for the online evaluation in the next section.

Table 2: Demographic profiles of the pilot study (numbers in the bracket stand for the total number of users).

Age	≤20 (5); 21-30 (83); 31-40 (32); 41-50 (18); 51-60 (5); ≥ 60 (5)
Gender	Male (96); Female (47); Not tell (5)
Nationality	Asia (53); Europe (38); South America (42); North America (12); Africa (3)
Education Level	Graduate School (83); College (45); High School (20); Others(2)

5.1 Pilot study

To address our first research question, we conducted a pilot study, in which we collected users' personality information and their music preferences (preferred songs). We designed a website [1] for the user survey. The survey contains four main parts:

- **User's basic information**: Collecting users' demographic information such as their age range and gender.
- **Personality test**: The personality test in our pilot study is conducted via the TIPI, in which users need to answer ten self-assessment questions. Each question should be rated from 1 to 7, from 'Disagree strongly' to 'Agree strongly' (e.g., I see myself as extraverted, enthusiastic).
- **Music preference collection**: Users' music preference is collected by means of Spotify Web API, with which users are asked to provide at least 20 preferred songs that they normally listen to and can best describe their music taste. Users are also asked to rate their selected songs from 1 to 5 (least preferred to most preferred).
- **User comments**: A free-text comment section is included.

5.2 Pilot Study Results

We spread the survey via two channels: Crowdsourcing platforms and students at several universities (e.g., TU Delft, Netherlands; EPFL, Switzerland; and Lanzhou University, China). The majority (around 80%) of the participants are recruited from Crowdflower (now called Figure Eight) [2]. To ensure the quality of the data collected, we also inserted some test questions into the survey to help us filter suspicious responses. On the Crowdflower platform, workers also need to submit their contributor ids and verification codes which are displayed at the end of the survey. These verification methods helped us remove a number of irresponsible participants, especially from the Crowdsourcing platform. Results for the user survey are shown below.

Participants. 148 participants were recruited to participate in the survey, the demographic properties of these participants are shown in Table 2.

Relation between Personality Factors and Single Attribute Diversity of Music Preference. When studying the correlation between personality factors and each attribute's diversity degrees, we first calculated the personality scores for each user from the TIPI question scores. Then, we computed the diversity scores for each attribute within the list of tracks a user has selected using the ILD (Equation 1) metric. For each track, we have chosen six attributes to compute specific diversity degrees: *Release Times*, *Artists*, *Number of*

[1] Available at: https://music-rs-personality.herokuapp.com
[2] Crowdflower: https://www.figure-eight.com

Table 3: Spearman Correlation coefficient between personality factors/demographic values and diversity degrees w.r.t. single attribute (*p-value<0.05 and **p-value<0.01). E: Extraversion, A: Agreeableness, C: Conscientiousness, ES: Emotional Stability, O: Openness.

	E	A	C	ES	O	Gender	Age
Div(Release times)	-0.03	-0.12	0.01	0.11	-0.15	0.00	**0.28****
Div(Artists)	0.10	0.09	0.11	**0.22****	-0.04	-0.03	-0.16
Div(Artists number)	0.00	**0.25****	0.13	0.15	0.07	0.06	-0.14
Div(Genres)	0.07	0.00	-0.01	**0.25****	0.06	0.06	0.03
Div(Tempo)	0.11	0.09	0.11	**0.24****	0.08	-0.17*	-0.02
Div(Key)	**0.21****	0.05	0.06	0.17*	0.08	-0.13	-0.10

Artists, *Genres*, and two audio features (*Tempo* and *Key*). Spearman's rank correlation coefficient was used to calculate the correlation between the five personality factors and the diversity scores for each attribute. In addition, considering that some demographic values might also have some impact on the diversity needs for users when delivering recommendations, we also included two demographic values (age and gender) in the correlation comparison. Results are shown in Table 3.

Relation between Personality Factors and Overall Diversity. Besides studying the correlation between the personality factors and diversity scores for single attribute, we also computed the correlation between the overall diversity and user's personality values. Considering that different users usually place different weights on attributes (e.g. some user may consider that the diversification of Artists is the most important), we assigned three different sets of weights to the six attributes (*Release Times, Artists, Number of Artists, Genres, Tempo, Key*) in reference to [16]: **Overall_Div1**: 'Equal weights method' (1/6, 1/6, 1/6, 1/6, 1/6, 1/6); **Overall_Div2**: 'Rank-order centroid (ROC) weights' (0.41, 0.24, 0.16, 0.10, 0.06, 0.03); **Overall_Div3**: 'Rank-sum (RS) weights' (0.29, 0.24, 0.19, 0.14, 0.09, 0.05).

From Table 3 and 4, we concluded four important correlations. For single attribute diversity, we find:

- **C1.** Personality factor *Extraversion* has a positive correlation with the diversity degree of *Key*.
- **C2.** Personality factor *Agreeableness* has a positive correlation with the diversity degree of *Artists Number*.
- **C3.** Personality factor *Emotional Stability* has a positive correlation with the diversity degrees of *Artist*, *Genre* and *Tempo*.

We also find that: **C4.** Personality factor *Emotional Stability* has a positive correlation with the *overall diversity degree*.

These correlations can then be used to map the parameters in our diversification algorithm (c.f., Section 4.2). Specifically, λ is adjusted according user's **Emotional Stability** level. We used three attribute diversity in the later experiment (see Section 5.5): *Genre*, *Artists Number*, and *Key*. Thus, θ_1, θ_2, and θ_3 are adjusted according to user's **Emotional Stability**, **Agreeableness** and **Extraversion** respectively.

Table 4: Spearman Correlation coefficient between personality factors/demographic values and overall diversity (p-value<0.01). E: Extraversion, A: Agreeableness, C: Conscientiousness, ES: Emotional Stability, O: Openness.**

	E	A	C	ES	O	Gender	Age
Overall_Div1	0.11	0.09	0.08	**0.31****	0.03	0.01	-0.05
Overall_Div2	0.11	0.08	0.08	**0.28****	0.01	0.00	-0.10
Overall_Div3	0.12	0.06	0.07	**0.29****	0.02	0.00	-0.09

5.3 Offline Evaluation

Since our diversification algorithm is built upon a re-ranking algorithm (a diversification method by re-ordering the recommendation list), its final diversity degree is affected by some re-ranking related parameters such as the size of the final top-N re-ranked list (N). The personality related parameters (λ, θ_1, θ_2, θ_3, see Section 4.2) will also greatly influence the final diversity degrees of the recommendation lists. Thus, we have conducted a series of offline evaluations to test the influence of different parameters. The parameters we tested are:

- The size of the final top-N re-ranked list (N).
- The size of the input list (LS).
- The size of the unrated items used for recommendation (K).
- Personality related parameters λ.

In order to generate initial recommendations with high quality, we used a state-of-the-art recommendation algorithm called Factorization Machine (specifically, fastFM [1]) [23]. To train the FM sufficiently, we combined our pilot study dataset (148 users' data in Section 5.2) with a complementary dataset with much larger user data: The Echo Nest Taste Profile Subset (TPS) [3] [2]. We made a few data selection beforehand. We first ruled out those tracks that have only been listened to once. Then we ruled out those users who listened to fewer than 100 tracks in total. The TPS dataset only contains track play counts. We further mapped the play counts into the integer ratings (1-5) using the rating mapping algorithm mentioned in [5].

We then first split our pilot study dataset into two subsets: training Set M_1 and testing set T. T contains the top-5 rated tracks (ratings all ≥ 4) for each user, which we will consider as the relevant items to each user. The remaining user data of the pilot study dataset (M_1) is combined with the TPS subset (M_2) to form our whole training set M. After training the FM, we used this FM to generate recommendations for users in the testing set T.

Hit Rate. The first metric we used in our offline evaluation was an accuracy measure. Hit rate was chosen due to the large item count (number of distinct tracks) and the small number of listening history per user [8]. Instead of using all unseen items (all items not used for training for each user) for prediction and counting the number of 'hits' (relevant items) in the top-N list, in our testing method, each relevant item (known top-5 rated relevant items for each user) in the Testing Set is evaluated separately by combining it with K (we used K=100) other items that this user has not rated.

[3]The Echo Nest Taste profile subset: http://labrosa.ee.columbia.edu/millionsong/tasteprofile, extracted in July, 2018

Table 5: Comparison of the two lists on the accuracy (Hit rate) and diversity (ILD) for N=10, LS=50, K=100.

	Initial List	Re-ranked List
Hit rate@10	0.043	0.141
ILD@10	0.390	0.483

We assume that these unrated items will not be of interest to user u, representing the irrelevant items. The task of the FM is then to rank these K+1 items for each user. For each user, we generate the two recommendation lists: initial recommendation list (top-N items from the initial list generated by FM) and our re-ranked list. We then check whether this item is in the two lists. If in, we consider it as hit, if not, we consider it as miss. This process is repeated for each item in the Testing Set. The final hit rate is computed as: $H(N) = \#hit/|T|$.

ILD. We also compare the diversity degrees for both recommendation lists using intra-list diversity.

5.4 Offline Evaluation Results

Our offline evaluation results show that, for both N and LS (K is fixed to 100), the hit rate for both lists will increase when N and LS increases (hit rate for our re-ranked list is always higher than the initial list). Diversity degrees also increase when we increase the two parameters (ILD for our re-ranked list is always higher). For parameter K, results show that both hit rate and ILD drop when we increase K. For the personality related parameter λ, we find that both the hit rate and ILD values will increase when we keep increasing λ.

After separately evaluating the influence of these parameters, we then made a final comparison on the two lists. Results are shown in Table 5. We see that our re-ranked list outperforms the initial list both in hit rate and ILD.

5.5 Online Evaluation

Considering that offline evaluation metrics cannot always reflect the actual user satisfaction for recommendations in real life. To further evaluate whether our personality-based diversification algorithm can really enhance user satisfaction and users' perception of list diversity, we therefore conducted the following online evaluation.

Similar to our pilot study (in Section 5.1), we also constructed a website [4] for the evaluation.

5.5.1 ***Materials.*** Two materials are needed from the users beforehand: the Personality Profile and the User Interests.

Personality Profile. We still adopted the Big-Five Factor Model as the basic personality model in our system. Ten Items Personality Inventory (TIPI) is also used to extract these five personality factors from users.

User Interests & Recommendation. To generate the initial recommendation list, we request users to offer their music interests in advance. In our online evaluation, we used Spotify Recommendation System based on their open Web APIs [5] in order to provide real-time recommendations. User interests are represented as 'seed information' in Spotify Recommendation. Three kinds of seed information are used: artists, tracks, and genres. Spotify has a restriction on the total number of input seeds, which is maximally 5. To ensure that the originally generated recommendation list (which has 100 tracks) is already diverse enough, we use at least 1 artist seed, 1 track seed, and 1 genre seed for every recommendation.

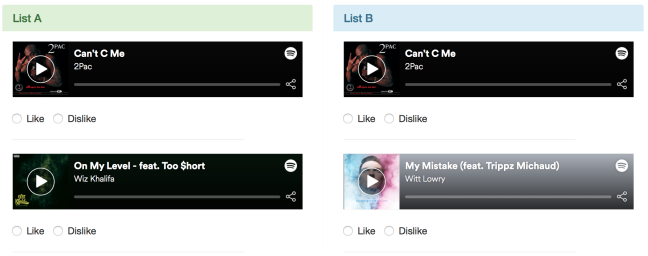

Figure 2: Example of the two recommendation lists shown to users. One is the initial list and the other is the re-ranked list (in random order). Users can click on the button to have a preview for each track. Users also need to choose whether they like the track or not. The first two tracks are shown in this figure. In total, there are ten tracks in each list.

5.5.2 ***Independent Variables.*** After we obtain the two materials from users, we then generate the recommendations for them. In the evaluation, similar to offline evaluation, we generate two recommendation lists (initial list and re-ranked list) for each user, each list contains 10 tracks. We adopted a within-subjects experimental design where the two recommendation lists are displayed to the users at the same time (see Figure 2). Thus, the independent variables here are the two recommendation lists. The order of presentation was balanced between participants.

5.5.3 ***Dependent Variables.***

Precision@10. In order to directly measure the precision of the recommendations, we ask the users to rate each track as 'Like' or 'Dislike'. Tracks rated as 'Like' are considered as relevant items. The Precision@10 for each list is computed as proportion of relevant items in the whole list.

Diversity. For both lists, we also used ILD (Equation 1) to compute the diversity degrees.

User Feedback. In addition to calculating the precision and ILD for each recommendation list, we also ask user for some feedback on the two lists via a post-task questionnaire. Each user needs to express their opinions on both lists in terms of the following three main aspects:

- Recommendation Quality (Q1 & Q2): *"The items in List A/B recommended to me matched my interests."*
- Recommendation Diversity (Q4 & Q5): *"The items in List A/B recommended to me are diverse."*

[4] Available at https://music-rs-personality-online.herokuapp.com

[5] Spotify Recommendation: https://developer.spotify.com/documentation/web-api/reference/browse/get-recommendations/

- User Satisfaction (Q7 & Q8): *"Overall, I am satisfied with the Recommendation List A/B"*

All of these questions are referred to the ResQue User-Centric Evaluation Framework [22], which are are responded on a 5-point Likert scale, from 1 to 5, meaning from "Disagree strongly" to "Agree strongly". We then compute and compare the average ratings for each question on both lists. Considering that users may give the same ratings for both lists, we added two more sub-questions regarding the Recommendation Quality and Recommendation Diversity:

- Recommendation Quality (Q3): *"Which Recommendation List is more interesting to you (match more of your interests)?"*
- Recommendation Diversity (Q6): *"Which Recommendation List is more diverse to you?"*

These two questions rated with categorical answers: "List A", "List B", or "Hard to tell".

5.5.4 ***Procedure Design.*** Similar to our pilot study, four main parts are included in the website: The user basic information, personality test, recommendation and feedback, and user comment. The user basic information, personality test, and the last user comment parts are similar. For the recommendation and feedback part, we provide two channels to obtain users' original interest: a) utilize Spotify history; or b) Type in manually. If users choose to use their Spotify listening history, we will use two of their top-played artists, two of their top-played tracks, and the top-played genre for generating the recommendations. Users can alternatively choose to type in their interests manually. In this way, we request users to type in at least one artist seed, one track seed, and one genre seed.

After we obtain users' music preference, we then feed these seeds into the Spotify recommendation system to generate the initial recommendation list (100 tracks). The first list L_1 is constructed by directly taking the top-10 items from the initial list. The second list L_2 is generated based on our personality-based diversification algorithm. We select the top-50 tracks as the input list for re-ranking. To minimize any carryover effects, we show these two lists in random order to users (displayed as List A and List B). For each track, users can click on the play button to listen to a 30 seconds' preview. The track name and the corresponding artist name are also shown in the list. For each track, users need to rate as 'Like' or 'Dislike' for both lists. After rating all the 20 tracks, users are asked to fill in the feedback questionnaire (see Section 5.5.3).

6 ONLINE EVALUATION RESULTS

To evaluate users' actual satisfaction towards our personality-based diversification method, we conducted this online evaluation.

6.1 Participants

We conducted our online evaluation with 25 participants recruited at a university. Participants' ages ranged from 21-30 years old. Table 6 summarizes their demographics.

6.2 Feedback Questions

Figure 3 shows the comparison of the two lists on three aspects. We used a paired t-test for questions on a 5-point Likert scale (Q1, Q2, Q4, Q5, Q7, and Q8). And we applied Chi-Squared Test for the questions with categorical answers (Q3 and Q6).

Table 6: Demographic profiles of 25 participants for the online evaluation.

Gender	Male (13); Female (8); Prefer Not to Answer (4)
Age	21-30 (25)
Education	College (4); Graduate School (21)

Table 7: Precision@10 and ILD@10 for the two lists. Pairwise t-tests significant at $p < 0.05$.

	Initial List L_1	**Re-ranked List L_2**
Precision@10	0.58 (std: 0.15)	0.668 (std: 0.14)
ILD@10	0.48 (std: 0.06)	0.57 (std: 0.07)

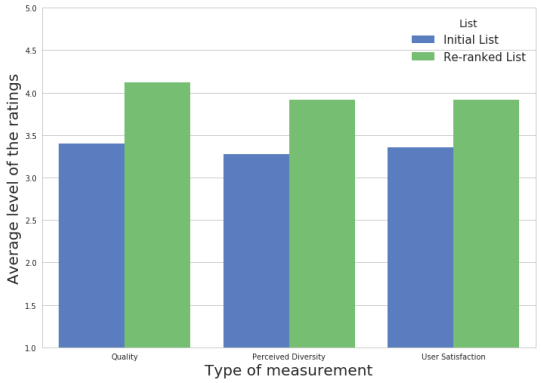

Figure 3: Full comparison for Recommendation Quality (Accuracy), Diversity and User Satisfaction. Student t-Test is also used. $p < 0.05$.

Recommendation Quality. Specifically, for recommendation quality (Q1 and Q2), the average ratings for the two lists are 3.4 (initial list, std=0.98) and 4.12 (re-ranked list, std=0.65) (t=-3.00, p=0.004). Q3 further compares the recommendation quality of the two lists with categorical answers. Results show that 8.0% users think the Initial List is better in matching their interests, 52.0% users think the re-ranked list is better, other 42.0% users think it is hard to tell (for Chi-Squared Test, statistic=7.76, $p < 0.05$).

Recommendation Diversity. Table 7 shows the Precision@10 and ILD@10 results for both lists.

For perceived recommendation diversity (Q4 & Q5), the average ratings for the two lists are 3.28 (initial list, std=0.96) and 3.92 (re-ranked list, std=0.89) (t=−2.39, p=0.02). Q6 further compares the recommendation diversity of the two lists with categorical answers. Results show that 16.0% users think the initial List is better in matching their interests, 48.0% users think the re-ranked list is better, other 36.0% users think it is hard to tell (for Chi-Squared Test, statistic=3.92, p=0.14).

User Satisfaction. For user satisfaction (Q7 & Q8), the average ratings for the two lists are 3.36 (initial list, std=0.93) and 3.92 (re-ranked list, std=0.97) (t=-2.03, $p < 0.05$).

7 DISCUSSION AND LIMITATION

From the online evaluation results, we see that our re-ranked recommendation list outperforms the initial recommendation list in all three aspects (recommendation quality, diversity, and user satisfaction). For the two categorical questions Q3 and Q6, results for Q3 is in line with the results shown in Figure 3. While for Q6, the p-value for Chi-Square Test is larger than 0.05, which means that there is no significant difference for Q6 when we asked users which list is more diverse to them. The reason behind this phenomenon may lies in our limited sample size. The precision of the two lists has no big difference (around one relevant track difference). While considering that our algorithm has raised the diversity level of the recommendation at the same time, we still can say that the re-ranked list is better in users' perspective and our personality-based diversification algorithm has enhanced the diversity adjusting strategy in music recommendations.

One limitation of our research lies in the limited sample size both in pilot study and online evaluation. If more participants are recruited in our pilot study, the correlation between personality factors and diversity needs may be stronger. Similarly, more users included in our online evaluation might also yield better results. Later researchers are suggested to repeat our research with more participants. Another limitation lies in that we did not include more features (e.g. more audio features like loudness) in our pilot study.

8 CONCLUSION

In this paper, we proposed a solution to address the research gap between research in diversity-based recommender systems and personality-based recommender systems. We proposed an algorithm to adjust the diversity degrees in music recommendations adaptively for users with different personalities. The adjustment was based on a pilot user study which explored the relationship between users' personality factors and their diversity needs on music preferences. To assess the effectiveness of our algorithm, we conducted both offline and online evaluations. Results suggest that our diversification method not only increases the diversity degrees for recommendations, but it also gains more user satisfaction.

In future work, more (audio) features with a larger participant pool will be studied. Instead of using the explicit personality test, we also plan to try implicit personality extraction method (e.g. via social media) in later work. Moreover, besides the re-ranking algorithm, we also plan to try different diversification strategies (e.g. optimization based diversification) with personality to check whether they would yield better results.

REFERENCES

[1] Immanuel Bayer. 2016. fastFM: A Library for Factorization Machines. *Journal of Machine Learning Research* 17, 184 (2016), 1–5.

[2] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

[3] Keith Bradley and Barry Smyth. 2001. Improving recommendation diversity. In *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland*. 85–94.

[4] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 335–336.

[5] Òscar Celma Herrada. 2009. Music recommendation and discovery in the long tail. (2009).

[6] Li Chen, Wen Wu, and Liang He. 2013. How personality influences users' needs for recommendation diversity?. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 829–834.

[7] Li Chen, Wen Wu, and Liang He. 2016. Personality and Recommendation Diversity. In *Emotions and Personality in Personalized Services*. Springer, 201–225.

[8] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 39–46.

[9] Tommaso Di Noia, Vito Claudio Ostuni, Jessica Rosati, and Paolo Tomeo. 2014. An analysis of users' propensity toward diversity in recommendations. In *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 285–288.

[10] Daniel M Fleder and Kartik Hosanagar. 2007. Recommender systems and their impact on sales diversity. In *Proceedings of the 8th ACM conference on Electronic commerce*. ACM, 192–199.

[11] Ishan Ghanmode and Nava Tintarev. 2018. MovieTweeters: An Interactive Interface to Improve Recommendation Novelty. In *IntRS@ RecSys*.

[12] Samuel D Gosling, Peter J Rentfrow, and William B Swann. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[13] Rong Hu and Pearl Pu. 2010. A study on user perception of personality-based recommender systems. *User Modeling, Adaptation, and Personalization* (2010), 291–302.

[14] Rong Hu and Pearl Pu. 2010. Using personality information in collaborative filtering for new users. *Recommender Systems and the Social Web* 17 (2010).

[15] Rong Hu and Pearl Pu. 2011. Enhancing collaborative filtering systems with personality information. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 197–204.

[16] Jianmin Jia, Gregory W Fischer, and James S Dyer. 1998. Attribute weighting methods and decision quality in the presence of response error: a simulation study. *Journal of Behavioral Decision Making* 11, 2 (1998), 85–105.

[17] Oliver P John and Sanjay Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* 2, 1999 (1999), 102–138.

[18] Jayachithra Kumar and Nava Tintarev. 2018. Using visualizations to encourage blind-spots exploration. In *IntRS@ RecSys*.

[19] Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality* 60, 2 (1992), 175–215.

[20] Sean M McNee, John Riedl, and Joseph A Konstan. 2006. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*. ACM, 1097–1101.

[21] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[22] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 157–164.

[23] Steffen Rendle. 2010. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 995–1000.

[24] Peter J Rentfrow and Samuel D Gosling. 2003. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology* 84, 6 (2003), 1236.

[25] Barry Smyth and Paul McClave. 2001. Similarity vs. Diversity. In *Proceedings of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development (ICCBR '01)*. Springer-Verlag, London, UK, 347–361.

[26] Nava Tintarev and Judith Masthoff. 2013. Adapting recommendation diversity to openness to experience: A study of human behaviour. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 190–202.

[27] Marko Tkalcic and Li Chen. 2015. Personality and Recommender Systems. *Recommender Systems Handbook* (Jan. 2015).

[28] Marko Tkalcic, Matevz Kunaver, Andrej Košir, and Jurij Tasic. 2011. Addressing the new user problem with a personality based user similarity measure. In *First International Workshop on Decision Making and Recommendation Acceptance Issues in Recommender Systems (DEMRA 2011)*. 106.

[29] Marko Tkalcic, Matevz Kunaver, Jurij Tasic, and Andrej Košir. 2009. Personality based user similarity measure for a collaborative recommender system. In *Proceedings of the 5th Workshop on Emotion in Human-Computer Interaction-Real world challenges*. 30–37.

[30] Saúl Vargas. 2011. New approaches to diversity and novelty in recommender systems. In *Fourth BCS-IRSG symposium on future directions in information access (FDIA 2011), Koblenz*, Vol. 31.

[31] Saúl Vargas and Pablo Castells. 2013. Exploiting the diversity of user preferences for recommendation. In *Proceedings of the 10th conference on open research areas in information retrieval*. 129–136.

[32] Wen Wu, Li Chen, and Liang He. 2013. Using personality to adjust diversity in recommender systems. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM, 225–229.

[33] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*. ACM, 22–32.