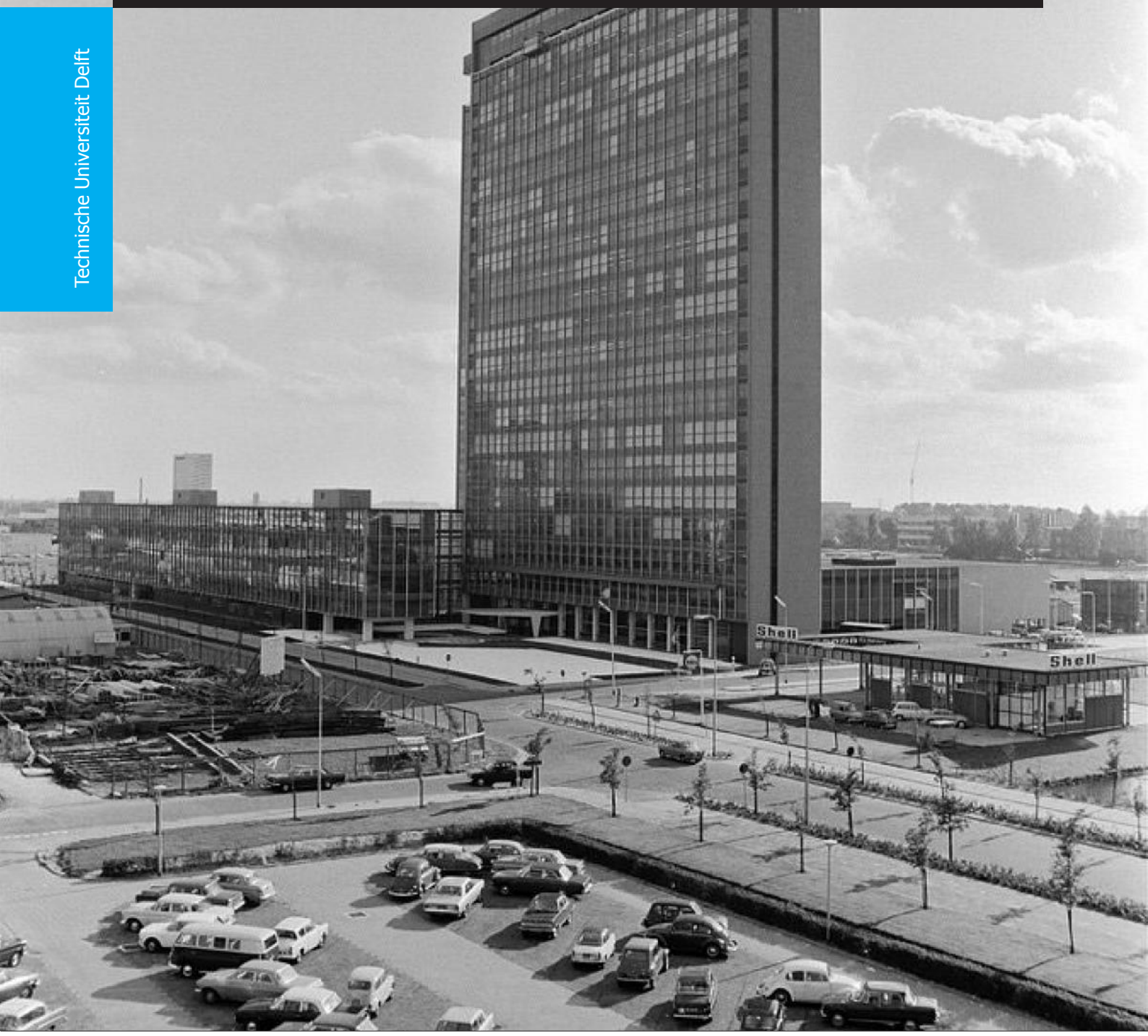
# Enhancing the Diversity Adjusting Strategy with Personality Information in Music Recommender Systems

Feng Lu

Technische Universiteit Delft




TU Delft Delft University of Technology

Challenge the future

# Enhancing the Diversity Adjusting Strategy with Personality Information in Music Recommender Systems

by

**Feng Lu**

in partial fulfillment of the requirements for the degree of

**Master of Science**
in Computer Science

at the Delft University of Technology,
to be defended publicly on Friday August 31, 2018 at 2:00 PM.

| | | |
|---|---|---|
| Supervisor: | Dr. Nava Tintarev | |
| Thesis committee: | Prof. dr. Geert-Jan Houben, | TU Delft |
| | Prof. dr. Martha Larson, | TU Delft |
| | Dr. Nava Tintarev, | TU Delft |

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft Delft University of Technology

# Abstract

Current research on personality and diversity based Recommender Systems (RecSys) are mostly separated. In most diversity-based Recommender Systems, researchers usually endeavored to achieve an optimal balance between accuracy and diversity while they commonly set a same diversity level for all users. Different diversity needs for users with different personalities are rarely studied. Another branch of research on personality-based Recommender Systems mostly emphasize utilizing personality information to enhancing the rating prediction accuracy so as to solve the 'Cold-Start Problem'. While few of them have in depth investigated whether and how it influences users' other preference needs (such as diversity needs).

This thesis presents the work how we combine these two branches of research together. Anchored in the music domain, we investigate how personality information can be incorporated into the Music Recommender Systems to help adjust the diversity degrees for people with different personalities. We first conducted a pilot study to investigate the correlation between users' personality factors and their diversity needs on the music recommendations. Results showed that there exits significant correlations between them, especially when we consider the personality factor 'Emotional Stability'. Based on such findings, we then 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 demonstrated that our proposed method is an effective solution to generate personalized recommendation lists with relatively higher diversity.

**Keywords** Recommender Systems, Diversity, Personality, Re-ranking, Music Recommendations

# Preface

It was six years ago, I selected the Computer Science as my major by chance. I gained almost all the honors I could acquire during my four years of college life. Two years ago, I was rejected by all the American 'dream schools' I applied. Feeling frustrated, I came to the Netherlands, came to TU Delft, hoping that I can start a meaningful life here.

TU Delft does not fail me. Everything here is fulfilling and rewarding, both in life and education. While I still felt confused. I still felt that I was a freshman in academia. Then, one year ago, I met Nava, my supervisor, who illumined my academic career. I came to her like a blank sheet on academic research. While Nava is strict, farsighted but rather patient. I still remember that I have revised my proposal for four times. It was a horrible time but rather meaningful when I look back. Thus, I would like to first express my gratitude to my supervisor, Dr. Nava Tintarev, who taught me almost everything I need to learn in academic research. Hopefully, such 'intellectual legacy' I gained from Nava will carry on. I would also like to give my thanks to Prof. Martha Larson and Prof. Geert-Jan Houben for being my thesis committee. Prof. Martha's suggestions helped me look deeper into my work. Thanks Daniel Morales, my Italian friend, for his generous help during the very early stage of the research. Thanks all the epsilon members in Nava's group for continuously giving me advice on the work.

In addition, I would like to thank my girlfriend, Jianing Li, for her perpetual accompany during the last five years. Thanks all my friends in TU Delft for giving a unforgettable memory here in Delft. Last but not least, thank you, my parents, for supporting me all the time.

This thesis is the final work of my master study in Computer Science at Delft University of Technology. Most of the evaluation work is carried out on the Dutch national e-infrastructure with the support of SURF Cooperative (SURFsara).

*Feng Lu*
*Delft, August 2018*

# Contents

# 1

# Introduction

## 1.1. Motivation

It has long been argued that simply improving the rating prediction accuracy of the Recommender Systems (RecSys) does not always mean a better user experience [1]. Personalized recommendations help users cope with the information overload problem by filtering the relevant content for users online. While, over time, using such recommender systems will slightly decrease the diversity of content that users consume [2], limiting users' exposure to more diverse items and views. This leads to the so called 'filter bubble' problem [3] (see Section 2.1). In order to cope with such problem, metrics such as diversity and novelty thus have been proposed to offer an extra evaluation of the quality of Recommender Systems. This has led to the emergence of many diversity-based Recommender Systems, which are proposed by researchers endeavoring to achieve an optimal balance between accuracy and diversity [4, 5]. In addition, as another branch of research, following some prior research showing that personality is an enduring and primary factor that influences human's real-world social behaviors [6] and there exists a connection between people's personality traits and their tastes and preferences [7, 8], a few researchers recently have shown increased interests to cover psychological aspects in Recommender Systems, especially exploring the relationship between the personality and user preferences [9–11].

Studies also show that personalities influence human decision making process and interests for music and movies [7, 12, 13], which implies that personality information should be considered if we want to deliver personalized recommendations. While, in another aspect, since users' attitudes towards new or diverse experiences vary considerably [14], 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.

However, existing research on personality-based and diversity-based Recommender Systems are mostly separated [15]. Most current diversity-oriented recommender systems [16–18] 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 $\theta$ and re-ranked the generated recommendation list according to the calculated scores. While this balance is commonly fixed to all users, which means that they rarely consider that users may have different diversity needs. While for personality-based recommender systems, since most of the research work [19–21] in this field is designed to address the Cold-Start Problem, they always set a fixed diversity degree for the recommended lists for all the users. While 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 [13, 22] or book recommendations [14], people with different personalities may also need recommendations with different diversity degrees in Music Recommendations.

Considering the above reality in both branches of research, we ask the question of whether we can combine these two research together. For this reason, we have proposed a personality-based diversification algorithm to enhance the diversity adjusting strategy for people with different personalities. In the following section, we will discuss our research questions and research steps in details.

## 1.2. Research Questions & Steps

**Research Questions** Since our work is to study how we can incorporate users' personality information into the diversity adjusting algorithm in music recommendations, our main research question studies:

- *RQ: How does personality information affect how diversity degrees should be applied in Music Recommender Systems?*

In order to address the main research question, two sub-research questions are proposed:

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

By asking the first research question, we first conducted a Pilot Study to investigate whether there exists a relationship between users' personality information and their diversity needs on music preference. A relation model is built based on the Pilot Study results. To address the second research question, we first proposed a personality-based diversification algorithm referred to this relation model and then evaluated its effect both on recommendations diversity and accuracy. Our proposed diversification method will adjust the diversity degrees adaptively in music recommendations according to users' distinct personality information.

**Research Steps** Based on our research questions, we divided our research work into two steps: a) Pilot Study and b) Diversity Adjusting Strategy. The full illustration of my research steps can be checked in Figure 1.1.

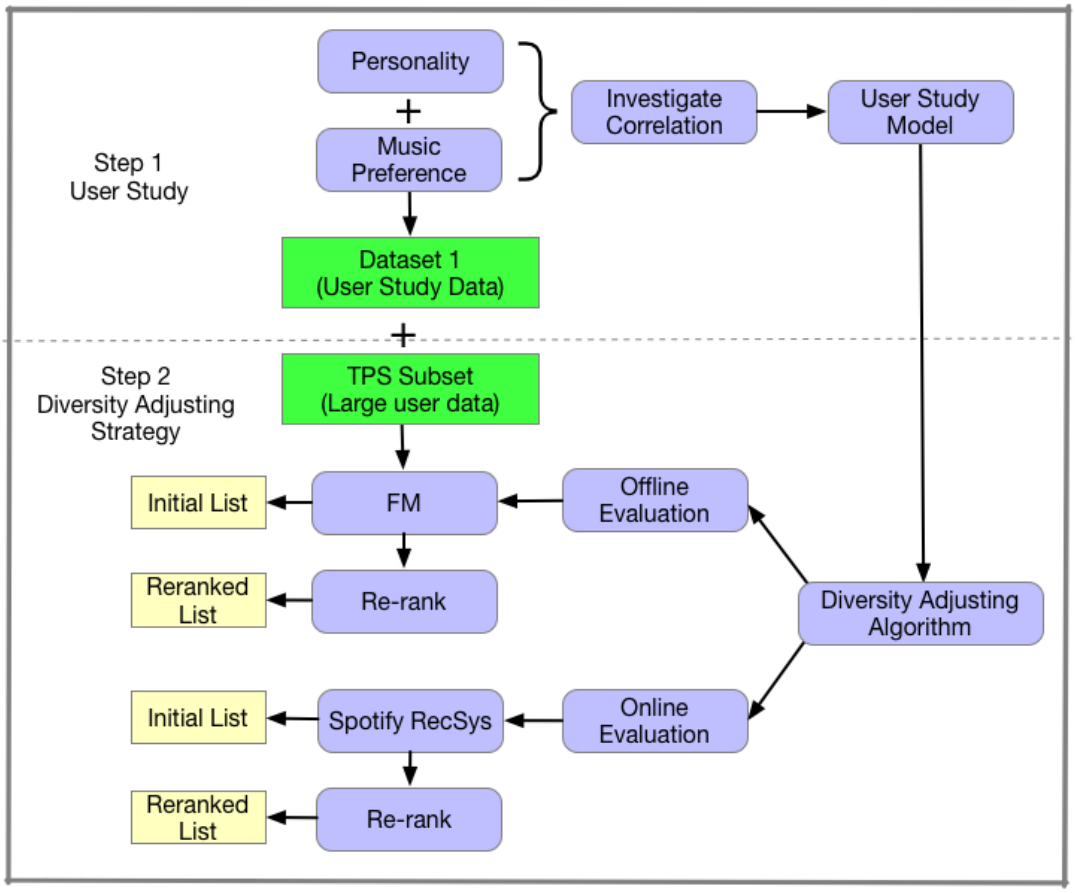



Figure 1.1: Research Steps

In the first step, a pilot study (introduced in Chapter 3) is conducted, in which users' personality information and music preferences are collected. We investigated the relation between these two objectives to construct our pilot study model. This relation model is used in our diversity adjusting algorithm in the second research step (introduced in Chapter 4). We then conducted both offline and online evaluations to evaluate the efficiency and effectiveness of our proposed algorithm.

## 1.3. Contributions

The main contributions of our research are twofold. First of all, we investigated the relation between users' personality factors and their diversity needs (both single attribute diversity and overall diversity) on music preference and found that there exist certain positive correlations between these two objectives. To the best of our knowledge, in the music domain, we are the first to conduct such systematic pilot study on the correlation between personality and users' diversity needs. In the movie domain, Wu and Chen et al. [13] conducted a similar research on movie recommendations. However, our research are dissimilar in several aspects (like domain difference and algorithm difference).

Our second contribution lies in our proposed personality-based diversification algorithm. By using a more flexible re-ranking strategy, our algorithm can adaptively set different diversity levels for people with different personalities in music recommendations. Evaluation results show that users are more satisfied with the recommendation list generated by our algorithm.

## 1.4. Ethical Concerns

More or less, almost all personalized information filtering systems may cause a feedback loop by which people become isolated from new and diverse information due to the so called 'filter bubble' problem (see Section 2.1). Besides preventing users from views and content that are different from their own, such filtering algorithms (including our proposed diversification method) would cause some other ethical problems as well if not used properly.

The first problem is the privacy problem. Information systems need to construct a user profile model before they can apply the filtering algorithm to the user. To refine the user model, recommender systems may need to gather users' behavior/personal information. From the privacy perspective, the data collection process should be transparent and clear to users. The user profile itself is also another privacy concern. Being able to get access to a perfect user model would enable anyone to predict user's decisions for a wide of range of conditions [23]. In our research, all participants are anonymous. All information we collected (including the personality information) cannot be used to infer who the users are. Users are fully aware of the data collecting process and how their user data will be used in our research.

The second problem is the control (autonomy) problem. As users' growth of knowledge will be greatly influenced by the filtering algorithms [24] and the personalized filtering system usually only captures a snapshot of the user at one moment (e.g. personality in our case), personalized filtering would manipulated users' behavior to some degree. This is unavoidable in any information presentation system [23]. Behavior manipulation will also cause the 'filter bubble' problem since the filtering system itself is biased on the user profile. Our diversification algorithm is based on users' current personality information, which can also cause the 'filter bubble' problem if the personality information is not up-to-date. In such case, users' listening behaviors might be controlled by the filtering algorithm if it is not properly used.

Besides, although personality is considered as stable [25–27], it is changeable [28]. Personality stability is the result of the interplay between the individual and her/his environment [29]. Thus, users' personality might also be reshaped by our filtering systems if users' behavior is manipulated by systems based on our algorithm. As mentioned in the last paragraph, our diversification algorithm is based on users' current personality information. In order to mitigate such effects, we suggest that recommender systems based on our diversification algorithm should update users' personality information from time to time.

## 1.5. Thesis Organization

The following chapters are structured as follows:

- **Chapter 2: Literature Review**

In this chapter, we introduce the background and related work regarding our research. We will explain the motivations, used metrics and related research in details.

- **Chapter 3: Pilot Study**
  In this chapter, we introduce the first step of our research: the Pilot Study, in which the whole experiment design will be introduced. Results and discussion of the pilot study are also included.
- **Chapter 4: Diversity Adjusting Strategy**
  In this chapter, we introduce the second step of our research: the Diversity Adjusting Strategy, in which we will explain our personality-based diversification algorithm in details. The offline evaluation of the algorithm is also included in this chapter.
- **Chapter 5: Online Evaluation**
  Besides the offline evaluation, we also conducted an online evaluation for our proposed method, which will be introduced in this chapter.
- **Chapter 6: Discussion and Future Work**
  In this chapter, we will conclude the whole research. Discussion for both Pilot Study and evaluations for our diversity adjusting strategy will be included. Future work and suggestions for later researchers will follow.

# 2

# Literature Review

This thesis describes how we can use personality information to improve the diversity adjustment strategy in the domain of Music Recommendation, in which two important aspects are mostly concerned: diversity and personality.

Before we move on to the main work of the research, we first introduce the background and related work of the whole project and the reason why we anchor the research in the diversity and personality aspects. We first describe the 'filter bubble' problem and its negative effects in recommendation problems. We also explain how diversification methods intend to address such effects. Then, we explain the reasons why we choose to focus on the diversity and personality aspects of the recommender system (RS). Then, we introduce diversity and personality separately. When talking about diversity, the general concept of diversity as well as the research on diversity based recommender systems will be introduced. For personality, we will introduce the basic personality models and their acquisition methods. Research on personality based recommender systems then follows. At last, limitation and contribution will be shown.

## 2.1. Filter Bubble

The term 'filter bubble' was first raised by Eli Pariser [3] to describe a feedback loop by which people become isolated from new information due to the influence of online personalized information filtering. Personalized recommendations help users cope with the information overload problem, filtering relevant content for them. As a result, we keep seeing content that are relevant to our interests/views, we cease to be exposed to more diverse content contrary to our own. Such situation was coined as 'filter bubble' by Pariser, which would reduces user creativity and learning ability, and strengthens the user belief [3]. Tetlock [30] found similar effects of 'filter bubble' in the political research area. In their research, they found that normal people would give more accurate predictions than the experts when people with different background are asked about views on political and economic issues. The low prediction accuracy of the experts might be caused from the fact that their views might be strengthened and biased after years of study. Recommender systems can both increase and narrow the diversity of the content shown to users. Over time it is found that recommender systems will slightly decrease the diversity of content that users consume [2].

Isolating users in a filter bubble has its positive and negative effects [24]. For positive effects, users can get relevant information faster but not causing social data overload [31]. However, from the negative side, it becomes even harder for users to be aware of the distortion of the content they received [3], which means the information is biased without users realizing it. The second problem is the information equivalent of obesity [24], which means that such system will always surround users with ideas/content which they are already familiar with. Surprising and contradictory information is hardly provided to them, making it harder for users to learn and think. The third problem is the behavior manipulation [23]. The growth of user knowledge and user behavior is greatly influenced by the personalization algorithm of the system, which could also be controlled by such algorithm.

In order to cope with the 'filter bubble' problem, there are two common strategies [32]. One approach is to build diversity aware filtering algorithms and corresponding recommender systems.

Helberger et .al [33] suggested to exposure diversity as a design principle in the design of recommender systems to break potential 'filter bubbles'. Tintarev [34] introduced a diversity aware recommendation model for selecting and presenting a diverse selection of news for users, which aims to maximize the amount of diverse content that users are exposed to by considering both item and user diversity. The second approach is to provide techniques (e.g. interfaces) that can help users search for diverse exposure. As shown in [24], interactive visualization was found to increase users' awareness of the filter bubble as well as the understandability of the filtering mechanism.

In this research, we are focusing on the first approach to propose a diversity aware filtering algorithm that helps users relieve the negative effects of 'filter bubble'.

## 2.2. Why Diversity and Personality in Recommendation

For a long time, research on Recommender Systems have always been focusing on the aspect of prediction accuracy which measures the matching error between the predicted ratings of the recommendation list and users' actual ratings. However, recently, more and more research have argued that user's satisfaction is not always correlated with accuracy alone [1, 21, 35]. Metrics such as diversity and novelty thus have been proposed to offer an extra evaluation of the quality of Recommender Systems. In addition, as another branch of research, following some prior research showing that personality is an enduring and primary factor that influences human's real-world social behaviors [6] and there exists a connection between people's personality traits and their tastes and preferences [7, 8], a few researchers recently have shown increased interests to incorporate psychological factors into the Recommender Systems, especially using the personality factors [9–12].

In this section, we explore the reasons why we should consider diversity and personality in the research of Recommender Systems. While before that, we first review the process of the general recommendation problem.

### 2.2.1. A general overview of the recommendation problem

In general, the recommendation problem can be formalized as follows [12, 36]. Let *U* be the set of all users and *P* be the set of recommended items. Function *Pref(u,p)* is defined to measure the possibility of one item $p_i$ is liked by user $u_i$. Then, for each user $u \in U$, what the recommender system is doing is to find such item $p_u^{'} \in P$ that maximizes the inferred preference value:

$$\forall u \in U, p_u^{'} = argmax_{p \in P} Pref(u,p) \tag{2.1}$$

Function $Pref(u,p)$ here is loosely defined. Commonly, since the recommender system only observes a limited sample of the whole user preference, recommendations are operated on the incomplete knowledge, which means that the $Pref(u,p)$ function cannot be accurate enough to reflect users' whole actual interests. Such incomplete knowledge is also one of the reasons why we should not design our system to be as 'accurate' as possible alone.

To better construct and improve the $Pref(u,p)$ function, on the one hand, besides accuracy, we need wider concepts (such as diversity) to measure users' satisfactory to the system generated recommendations. In traditional recommendation problem, we mainly care about whether the prediction $p_u^{'}$ is accurate enough. While since we have known that the function $Pref(u,p)$ is not well-learned, in diversity-oriented recommendation problem, we also care about whether $p_u^{'}$ is diverse enough to reflect user's diversity needs inherent in $Pref(u,p)$ function. On the other hand, other context elements such as personality and moods of the users can also help to better infer the $Pref(u,p)$ function.

### 2.2.2. Why Diversity

As we have known, Recommender Systems should not be just as accurate as possible alone [1]. The concept of diversity is increasingly regarded as important as it may help users discover unexpected items that might be of interest to them [1]. Granted, diversity is not the only dimension of recommendation utility one should consider aside from accuracy, while it is a fundamental one. The motivations to introduce diversity in recommendations are various [37].

From the user's aspect, considered as a direct source of user satisfaction, diversity are generally desirable per se [37]. Studies on consumer behaviors showed that humans have a natural variety-seeking drive [38], which is also explained as a strategy to deal with the uncertainty about one's own

future preference when one will actually consume the choices [39]. Moreover, when users approach a recommender system, it is highly possible that they have a intention to discover new and unexpected items, which means that the items in recommendations should be diverse potentially. Lacking diversity in recommendations may result from the over-fitting problem and too much personalization, which is also known as the so-called filter bubble [3].

For the system side, users' activities on the system cannot reflect the whole actual user preferences. As the system only observes a limited sample of the whole user activity, recommendations are operated on the incomplete knowledge [37]. In addition, since users' interests are complicated and dynamic, predicting the user needs is therefore an inherently difficult task, unavoidably causing an error rate. Diversity is believed to be a approach to cope with such uncertainty by optimizing the chances that some items might at least satisfy the user. Thus, from the system perspective, diversity can be viewed as a strategy to optimize the gain drawn from accuracy in matching the actual user needs in an uncertain environment [37].

### 2.2.3. Why Personality

Through the last twenty years, a number of studies have revealed the inner relationship between users' personality and the user preferences. People with different personalities tend to have distinguished preferences. Studies also showed that human decision-making process and interests can be influenced by people's own personality [7]. Due to these inherent inter-related patterns between users' personalities and their behaviors, it seems quite natural to incorporate these differences in the Recommender Systems if we want to deliver personalized recommendations.

In the landmark work [7], Rentfrow and Gosling explored how music preferences are related to people's personality. By analyzing the statistical results on a large-scale dataset, they categorized each music piece into one of the four categories: reflective & complex (such as blues and folk music), intense & rebellious (such as rock music), upbeat & conventional (such as country and pop music) and energetic & rhythmic (such as rap and electronic music). After that, they empirically revealed that these four musical preferences are not only linked to the different levels of complexity and energy of musical compositions, they are also associated with users' personality factors defined in Five-Factor Model (FFM), which will be introduced in section 2.5.1.

Cantador's work [40] also observed the relations between user preferences and personality in various domains like movies, music, and books, where they not only found the relations between personality traits and individual domains but also in cross domains as well. Raghav et al. [41] from the GroupLens Research Group also showed substantial effects in multiple categories for various personality types and demonstrated that researches along the lines of incorporating personality in Recommender Systems are promising in contexts such as providing better cold-start recommendations, and delivering personalized recommendations with novel, diverse and serendipitous items.

All these researches support the hypothesis that people's personalities are related to their preferences. Since current algorithms of Recommender Systems are highly dependent on users' preferences, it therefore seems quite reasonable to incorporate the personality information into these systems.

## 2.3. Concept and Measurement of Diversity in Recommender Systems

Diversity is usually considered as the inverse of similarity [42], which refers to recommending a diverse set of items to users so as to help them discover unexpected and surprising items more effectively [1]. This concept has been introduced into the field of Recommender Systems as one of the possible solutions to address the over-fitting problem. The importance of diversity lies in its two purposes: addressing the over-fitting problem in Recommender Systems and increasing users' satisfaction with the recommendation list [43].

As evidenced by a large number of publications addressing diversity, this topic has discussed by a large number of research groups in the last few years.

### 2.3.1. Definition and Evaluation Metric of Diversity

In Recommender Systems (RS), diversity can be defined at two levels: *intra-list diversity*, which is the average pairwise dissimilarity between recommended items within the same list, and *inter-list diversity*,

which measures the dissimilarity between different recommendation lists. In this research, we focus on *intra-list diversity*.

### Intra-List Diversity

Research that is focused on the definition and evaluation of the intra-list diversity starts with Bradley and Smyth [44] who define the diversity as the averaged pairwise distance (dissimilarity) between all items in the recommendation set, which can be calculated as follows:

$$D(R) = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n}(1 - Similarity(c_i, c_j))}{n * (n-1)/2} \tag{2.2}$$

where $c_1..c_n$ are items in a set of recommendation list and R is the recommended list. While this metric is quite basic and highly dependent on the definition of item similarity.

Later, using the same idea, Ziegler et al. proposed the *Intra-List Similarity* as the aggregated pairwise similarity of items in the recommended list intending to capture the diversity of a list:

$$ILS(R) = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n} Sim(c_i, c_j)}{2} \tag{2.3}$$

where the function $Sim(c_i, c_j)$ measures the similarity between item $c_i$ and $c_j$.

Fleder and Hosanagar followed up and in 2007 proposed to use the Gini-coefficient (index) to measure the sales diversity, which measures how unequally different items are chosen by users when a particular recommender system is used. It is calculated as:

$$G = \frac{1}{n-1}\sum_{j=1}^{n}(2j - 1 - n)p(i_j) \tag{2.4}$$

where $i_1, \cdots i_n$ is the list of items ordered according to increasing $p(i)$. $p(i)$ presents the proportion of user choices, which can be different according to specific applications (for instance, $p(i)$ can be the proportion of a specific attribute value in a whole list). The index is 0 when all items are chosen equally often, and 1 when a single item is always chosen. While as pointed out in [43], Gini-index is still questionable whether it could be applied to environment like file/music/book recommendation since [45] only evaluated the effect on the diversity of sales.

Another measure of distributional inequality is the Shannon Entropy:

$$H = -\sum_{i=1} np(i)log(p(i)) \tag{2.5}$$

The entropy is 0 when a single item is always chosen or recommended, and $log(n)$ when n items are chosen or recommended equally often.

In 2011, another better metric to measure diversity in Recommender System was raised by Vargas [46] to measure the product of item's relevance, similarity and positions in the ranked list, which can be calculated as follows:

$$ILD(i_k|u, R) = C_k^{'}\sum_{l} disc(l|k)p(rel|i_l, u)dist(i_k, i_l) \tag{2.6}$$

where $C_k^{'}$ is a normalization factor, $disc(l|k)$ is a function that considers the distance in the ranked list between elements $i_l$ and $i_k$ in a browsing scenario, $p(rel|i_l, u)$ is the user-relative relevance probability and $dist(i_k, i_l)$ stands for the diversity metric. The user-relative relevance probability $p(rel|i_l, u)$ is an estimation for rating preferences based on a utility function $g(u, i) = max(0, r(u, i) - \tau)$, where $\tau$ represents the 'indifference' rating value. The relevance probability can be computed as:

$$p(rel|i_l, u) = \frac{2^{g(u,i)} - 1}{2^{g_{max}}} \tag{2.7}$$

This measure is fairly domain independent and also covers relevancy.

In 2013, Castagos et al. [47] performed a user study that compared the user's acceptance and satisfaction with presented diversified recommendation lists and found that although diversification could reduce the user's acceptance rate, it did increase the user's satisfaction with the system. This study confirms that the definition used in [44] is viable for use in real-life applications.

In 2014, Vargas et al. [48] focused on genre as one of key attributes of diversity evaluation and proposed a Binomial framework to measure genre diversity of each recommendation list:

$$BinomDiv(R) = Coverage(R) \cdot NonRed(R) \tag{2.8}$$

This method is quite complex which we will not explain in details.

#### Inter-List Diversity

In addition to the definitions that refer to the diversity within a single recommendation list for users, other measurements that measure the dissimilarity between different recommendation lists can be assigned to the class of *inter-list diversity*.

For instance, in 2010, Lathia et al. [49] proposed the concept of *Temporal Diversity* to measure the ability of the recommender systems not to provide the same or similar recommendations over time. They defined the diversity between two lists (at depth N) as the size of their set theoretic difference over N:

$$diversity(L1, L2, N) = \frac{|L2 \setminus L1|}{N} \tag{2.9}$$

where L1 and L2 are two recommendation lists and

$$L2 \setminus L1 = \{x \in L2 | x \notin L1\} \tag{2.10}$$

meaning the members of L2 that are not in L1.

At the same year, the concept of *aggregate diversity* was raised by Zhou et al. [50], which was defined as the average pairwise distance between recommendation lists generated for different users.

Since this study does not focus on the Inter-List Diversity, we will not discuss this concept in details.

### 2.3.2. Complementary Metrics besides Diversity

Adding diversity to the recommender system does not mean that we will neglect the accuracy of the system. To balance the dilemma between diversity and accuracy and better evaluate the quality of recommender systems, in addition to the measurement of diversity, most of the times, we need to measure the prediction accuracy of the systems as well. Two popular measures of accuracy are the Mean Absolute Error (MAE):

$$MAE = \frac{1}{|R_{test}|} \sum_{r_{ui} \in R_{test}} |f(u,i) - r_{ui}| \tag{2.11}$$

and the Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{|R_{test}|} \sum_{r_{ui} \in R_{test}} (f(u,i) - r_{ui})^2} \tag{2.12}$$

where $R_{test}$ is a testing set of user ratings and $r_{ui}$ is the true rating of user u on item i. Function $f(u,i)$ is the recommendation function that predicts the rating of a user u for a new item i.

However, when ratings are not available, measuring the rating prediction accuracy is not possible. In such cases, we transfer the the problem of finding the best item into the task of recommending a list of items that likely interest him or her. The performance of such method can be computed using the measures of precision and recall:

$$Precision = \frac{1}{|U|} \sum_{u \in U} |L(u) \cap T(u)| / |L(u)| \tag{2.13}$$

$$Recall = \frac{1}{|U|} \sum_{u \in U} |L(u) \cap T(u)| / |T(u)| \tag{2.14}$$

Table 2.1: Classification of the possible result of a recommendation of an item to a user

| | Recommended | Not recommended |
|---|---|---|
| Preferred | True-positive (tp) | False-negative (fn) |
| Not preferred | False-positive (fp) | True-negative (tn) |

where U is the set of users and L(u) stands for the recommendation list for user u. T(u) is the subset of test items that a user u found relevant.

In information retrieval, precision and recall are also commonly used to measure performance. In such context, Precision is a measure of how many errors we make in classifying samples as being of class A [51], which is defined as:

$$P = \frac{\#TruePositives}{\#TruePositives + \#FalsePositives}. \tag{2.15}$$

Recall measures how good we are in not leaving out samples that should have been classified as belonging to the class [51], which is computed as:

$$R(TruePositiveRate) = \frac{\#TruePositives}{\#TruePositives + \#FalseNegatives}. \tag{2.16}$$

Here, True Positives, False Negatives, and False Positives are three possible results of a recommendation of an item to a user, which can be referred to Table 2.1. Another related measure used in classification is called specificity (True negative rate), which is computed as:

$$Specificity = \frac{\#TrueNegatives}{\#TrueNegatives + \#FalsePositives} \tag{2.17}$$

As we can see from the equations, sometimes precision and recall are contradictory. A measure to summarize the precision and recall is called F-measure, which is computed as follows:

$$F = \frac{2 \times precision \times recall}{precision + recall} \tag{2.18}$$

## 2.4. Previous Approaches and Algorithms to Diversity Based Recommender Systems

This section focuses on the research on diversity oriented Recommender Systems that are mostly related to my study. Since they usually emphasize on applying diversity on different sides of a recommender system, we have assigned those research into different classes.

### 2.4.1. Diversification with Re-Ranking

From the algorithm's perspective, related research on diversity based recommender systems usually focus on achieving the optimal balance between two objectives, accuracy and diversity. To achieve this balance, most of the approaches adopt a method called re-ranking algorithm, in which the final diversified list of recommendations is created by reordering recommendations after they have been generated by any other recommendation algorithms (either collaborative or content-based). This means that the such re-ranking methods assume that initial recommendations are already diverse and just need to be reordered in order to achieve the maximum possible effect. The most widely used re-ranking function to add diversity into the recommendations adopts a linear combination of the item's similarity to the target list and the relative diversity degree to the items that are already in the recommendation list, which looks like follows [4, 16, 44]:

$$Q(t, c, R) = Sim(c, t) * (1 - \theta) + Div(c, R) * \theta \tag{2.19}$$

where the $Q(t, c, R)$ represents the final quality of the item c to the target list t in recommendation list R. $Sim(c, t)$ stands for the similarity function to compute the similarity between item c and target list t and $Div(c, R)$ represents the diversity of c relative to those items so far selected, $R = \{r_1, ..., r_m\}$.

In this research, we are focusing on enhancing the diversity degree by such re-ranking approaches. Some of the related works are shown as follows.

### Attribute based Recommendations

To increase the recommendation diversity, some of the approaches have focused on users' preferences for items' attributes. For instance, Ziegler et al. proposed the topic diversification approach [16] towards balancing top-N recommendation lists, which is a heuristic algorithm based on taxonomy (item's attribute) 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 list, one ranking the items that are similar to user's attribute-based preference and the other ranking the items in reverse. Both offline and online experiments were performed to test their system in the Book Recommendation domain. In offline experiments, they compared precision, recall, and intra-list similarity scores (ILS, see equation 2.3) and showed that their topic diversification appears detrimental to both user-based and item-based collaborative filtering (CF) along precision and recall metrics. While their online user survey showed that users could still perceive the positive effect on diversity and coverage although the accuracy declined in the offline experiment.

Considering the drawbacks of attribute-based diversification such as lack of item attributes or computational overhead of attribute retrieval, Yu et al. improved Ziegler's method by adopting an explanation-based method to diversify recommendations [52]. Instead of using taxonomy similarity, they defined the diversity as the distance between explanations for why the items are recommended. The explanation for a recommended item depends on the underlying recommendation strategy (either item-based or collaborative filtering) used. Technically, they did not use the attribute-based method, but they compared their system with the attribute-based diversification system from [16]. They evaluated their system using the Yahoo! Movies database and adopted two diversification algorithms, algorithm swap and algorithm greedy. Results showed that, compared with attribute-based diversification, their approach not only can achieve a similar level of diversification with better performance, but also it can apply in scenarios where the attributes are not available. While the big limitation of their work lies in that they did not define the concept of 'good balance' between relevance and diversity, which means that the way to set diversity levels for different users is not studied.

Vargas et al. [48] focused on the attribute genre to increase the diversity into the final recommendations. They adopted an objective function which combines item's relevance and binomial diversity as shown in function 2.19. The binomial diversity is defined as the combination of item's coverage score with non-redundancy score to measure the genre diversity in a recommendation list. For the genre distribution, as they claimed that the optimal approach for providing diverse recommendations is to make a random selection, they adopted a binomial distribution as the model for the genre distribution. Their experiments on two movie recommendation datasets validated the consistency and the quality of their proposed binomial framework.

### Optimization based Recommendations

Since most diversity-based Recommender Systems focus on achieving the optimal trade-off between diversity and accuracy, such balance can also be traded as an optimization problem from the algorithm's perspective. Hurley and Zhang [53] represented such trade-off between diversity and accuracy as a binary optimization problem and applied this approach to the top-N recommendation problem. They used the average pairwise dissimilarity of all items as the diversity measurement metric and relaxed the binary optimization problem with trust region algorithm. Their evaluation on the Movielens dataset showed that this method could increase the likelihood of the system recommending novel items, while maintaining good performance on the core items.

Smyth and McClave [4] compared three optimization strategies for balancing the similarity and diversity, which include bounded random selection, greedy selection, and bounded greedy selection. Their experimental results show that the bounded greedy selection strategy offers the best performance, not only in efficiency, but also in the way that it trades-off similarity for diversity for reasonable values of k. The same strategies were used by Bradley and Smyth [44] to improve diversity in a content-based recommender system.

In contrast to previous approaches, Parambath et al. [54] did not rely on an explicit trade-off between a relevance objective and a diversity objective since they argued that the estimations of relevance and diversity are implicit in the coverage criterion. Instead, they viewed items as nodes in a similarity graph, and defined the coverage of a set of items as the similarities between pairs of nodes from another set of items. They formulated the relevance and diversity trade-off as finding a set of unrated items that covers the set of items that were positively rated by the user. Relevance is

measured by covering a set of unrated items similar to the set of items that were positively rated by the user. While diversity is obtained by defining the coverage as a submodular function, where the diversity degree only improves largely when the list covers a new positively rated item. They tested their system on the MovieLens and Yahoo! Movies datasets and results showed that their algorithm performed well both in relevance and diversity compared with Maximal Marginal Relevance (MMR) and Max-Sum Diversification (MSD).

Recently, Juovac et al. proposed a generic Personalized Ranking Adaptation (PRA) re-ranking framework that can be used to include diversity in an optimization function [55]. The algorithm first estimates user tendency for various criteria like diversity and novelty, then iteratively and greedily re-ranks items in the head of the recommendations to match the top-N set with the user tendencies. This method is quite flexible and not restricted to a specific underlying item ranking algorithm. They evaluated their algorithm in both Movie and Music Recommendations on Recall values (at list length 10) and effectiveness. For the criteria of user tendency, they used list diversity (ILS), item popularity, and item release years in both domains. Their results showed that balancing the quality factors with PRA can be done with a marginal or no loss in ranking accuracy.

#### Rating based Recommendations

Since collaborative filtering (CF) recommender systems are operated based on user ratings, to improve such rating-based method, Zeng et al. [56] proposed a recommendation algorithm to increase recommendation diversity by considering both the effects of similar and dissimilar users under the framework of collaborative filtering. They adopted the simplest (while quite effective) method, *common neighbors*, to compute the similarity between two users. Namely, *common neighbors* mean that two users are regarded as more similar if they have collected more common objects. Accordingly, the dissimilarity are defined as the number of different objects that two users have collected. The final prediction function for an item to the target user is generated by combing the positive score from similar users and the negative score from dissimilar users linearly. They tested the performance of their algorithm on three datasets: MovieLens, Netflix and Amazon. Results showed that their method performs much better than the traditional collaborative filtering algorithm both in accuracy (measured by precision and Ranking Score) and diversity.

To improve recommendation diversity in collaborative filtering recommender systems, Mourão et al. [57] defined the Oblivion Problem that aims to identify which items have been preferred in the past and exhibit a high probability of being consumed by this user again in the present. They first verified such problem occurs in real domains and then performed a utility analysis checking the relevance of the forgotten items to RSs. Their testing on a sample of Last.fm dataset demonstrated that this approach helps to increase the diversity of the returned recommendations.

Boim et al. [58] proposed a novel technique for diversifying the recommendations by clustering items based on a unique notion of *priority medoids* that provides a natural balance between the need to present highly ranked items vs. highly diverse ones. The concept of priority-medoids is an adaptation of the classical notion of medoids. A standard medoid ((also called the cluster's representative)) of a given cluster is an element in the cluster whose sum of distances to the other items in the cluster is minimal. In Boim's research, they focused on the representatives with high ratings. Priority-medoids are therefore defined as the representatives that have highest rating in their corresponding clusters besides the requirements defined in standard medoids. They estimated item diversity by comparing all the ratings given to the items by the users. Using this approach they were able to create item clusters and create recommendation lists with higher diversity.

### 2.4.2. Diversification with Models

Although the re-ranking algorithms seem to be a effective solution to the diversification problem, we still cannot guarantee that the initial recommendation lists generated by the recommendation algorithms have been diverse enough. Spotting such disadvantage of the re-ranking algorithm, some of the other researchers adopted some preprocessing methods. Instead of postprocessing the data like reordering the list after it has been generated, they want to optimize the diversity degree of the list during the generation process itself. We will give a brief introduction to the typical works.

Shi et al. [59] combined matrix factorization technique with the portfolio theory in text retrieval, in which they captured the range of user interests and uncertainty of the user profiles by exploiting the variance of the latent user factors and they adapted that as the level of diversification to the user. Su

et al. [60] also proposed a diversification model based on learning-to-rank approach and integrated diversity into the matrix factorization model. They considered the diversity as a set-oriented concept and constructed a set-oriented CF model to quantify this concept. In the objective function of this model, they introduced a set diversity bias component that allowed to learn user specific need for diversity, which was computed by the average dissimilarity of the item latent features.

## 2.5. Personality

Psychologically, the personality stands for people's differences in their enduring emotional, interpersonal, experiential, attitudinal and motivational styles [61], which can also be considered as a user profile. The most important aspect of personality lies in its context and domain independent property, which makes it relatively stable and predictable through different context (such as time and location) and domains (such as movies and music domains).

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).

This section is mainly referred to Tkalcic and Chen's review in Recommender Systems Handbook [62] and the introduction of personality in Nunes's thesis [63].

### 2.5.1. Personality Models

Currently, one of the most well-known and commonly used personality model is the Big-Five Factor Model (Five Factor Model, FFM) [64]. In this model, personality is defined as five factors: Openness to Experience (O), Conscientiousness (C), Extroversion (E), Agreeableness (A), and Neuroticism (N). According to [62], specifically, Openness to Experience (O) distinguishes people between imaginative, creative (high score) and down-to-earth, conventional (low score). Conscientiousness (C) leads people to become prudent or impulsive. Extraversion (E) tells the degree of engagement with the external world. Agreeableness (A) then reflects the person's cooperation and social harmony. For Neuroticism (N), people with high score tend to be more sensitive and nervous than others.

The roots of the FFM lie in the lexical hypothesis, which states that things that are most important in people's lives eventually become part of their language [62]. By studying these languages, a set of adjectives that describe permanent traits are concluded as the five dimensions of FFM.

Although the Big Five factors have represented the personality model in a broad level, they do not guarantee that all dimensions of the personality traits have been exploited. In order to distinguish each factors more clearly, further studies have added more facets to each factors. Facets are used by psychologists in order to enrich Big Five dimensions with more fine-grained characteristics. For instance, NEO-PI-R [65] has 6 facets for each factor as shown in Table 2.2.

There are also other kinds of personality models such as the RIASEC model [66], which was used in an e-commerce prototype and Thomas-Kilman conflict mode personality model [67], which has been developed to model group dynamics.

While the Big-Five model has been shown to have excellent reliability in practice in the field of Recommender Systems [63], which makes it quite popular among the researches in personality-based Recommender Systems. In this research, we only consider the FFM.

### 2.5.2. Acquisition Methods

Since we have defined the personality model, the acquisition of personality parameters is another major issue in the design of personality-based Recommender Systems. Current acquisition methods can be classified into two groups:

- explicit methods, which are usually extracted from computer-based questionnaires
- implicit methods, which usually extract personality information from social networks

Explicit methods generally provide more accurate assessment of users' personality traits as they are more instructive than implicit methods. But they consume more time for people to finish the questionnaire.

While the implicit methods offer an unobtrusive way to extract people's personality via social networks as they do not disturb people with questionnaires. However, the accuracy of these techniques is not as high as those explicit methods and depends on the quality of the source.

Table 2.2: NEO-PI-R Facets of Big Five [65]

| Big Five Factor | Facet |
|---|---|
| Extraversion (E) | Warmth, Gregariousness, Assertiveness, Activity Excitement-Seeking, Positive Emotions |
| Agreeableness (A) | Trust, Straightforwardness, Altruism, Compliance, Modesty, Tender-Mindedness |
| Conscientiousness (C) | Competence, Order, Dutifulness, Achievement Striving, Self-Discipline, Deliberation |
| Neuroticism (N) | Anxiety, Angry Hostility, Depression, Self-Consciousness, Impulsiveness, Vulnerability |
| Openness (O) | Fantasy, Aesthetics, Feelings, Actions, Ideas, Values |

### Explicit Personality Test

In explicit methods to extract the personality traits, psychologists usually use computer-based questionnaires, which have either a large or a small amount of questions. The number of questions in the questionnaire is directly related to the granularity of the desired extracted traits from each person's Personality [63].

The most used explicit Personality Tests based on the Big Five factors are:

- 240-items NEO-PI-R (Revised NEO ( Neuroticism-Extraversion-Openness) Personality Inventory) [64]
- 300-items NEO-IPIP (NEO International Personality Item Pool) [68]
- 60-items NEO-FFI (NEO Five-Factor Inventory) [26]
- 44-items BFI (Big Five Inventory) [65]
- 10-items TIPI (Ten-Item Personality Inventory) [69]

Among all these inventories mentioned above, NEO-PI-R is considered as one of the most robust, used and well-validated commercial inventory in the world [70], which can not only measure the five factors, but also the six facets of each factor (30 facets in total)) as shown in Table 2.2. Most of the instruments that have less items than NEO-PI-R, are less fine-grained. They are mainly based on 5 factors of FFM and do not have defined facets. For instance, the NEO-FFI instrument, which measures the five factors only (but not their related facets), is a 60-item truncated version of NEO-PI-R.

In addition, Goldberg has proposed the creation of a public domain scale called IPIP - The International Personality Item Pool [71]. The IPIP's web page [1] contains questionnaires with 50 and 100 items, depending on the number of questions per factor (10 or 20). Johnson [68] created the NEO-IPIP based on Goldberg's IPIP Website and made it a free-of-charge version of NEO-PI-R.

John and Srivastava [65] developed a shorter list containing 44 items, called Big Five Inventory (BFI), by which each personality factor is measured by eight or nine questions. This questionnaire is also recognized as a well-established measurement of personality traits.

As the time to answer a reputed fine-grained Personality Inventory (like NEO-PI-R or NEO-IPIP) may be limited, much shorter instruments of questionnaires should also be provided.

A typical used shorter questionnaire is the Ten Item Personality Inventory (TIPI)[69], in which 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. The scores for these questions can be then mapped into the FFM model. The whole questionnaire inventory can be checked in Table 2.3.

Although different extraction techniques have been developed so far, the choice of these methods is highly context and application dependent, such as the time and finance limitation.

More explicit methods can be referred to Nunes's thesis [63].

[1] http://ipip.ori.org

Table 2.3: The ten-items personality inventory questionnaire[69]. 'R' denotes reverse-scored items.

| **Big5 model factor** | **Assessment: I see myself as** |
|---|---|
| Extraversion | Extraverted, enthusiastic. |
| Agreeableness | Critical, quarrelsome. |
| Conscientiousness | Dependable, self-disciplined. |
| Neuroticism | Anxious, easily upset. |
| Openness to Experiences | Open to new experiences, complex. |
| Extraversion - R | Reserved, quiet. |
| Agreeableness - R | Sympathetic, warm. |
| Conscientiousness - R | Disorganized, careless. |
| Neuroticism - R | Calm, emotionally stable. |
| Openness to Experiences - R | Conventional, uncreative. |

### Implicit Personality Acquisition

Implicit methods to extract personality traits are usually based on social media streams. For instance, in the study conducted by Quercia et al. [72], a strong correlation between features extracted from users' micro-blogs and their respective FFM factors was shown.

Lankveld et al. [73] also observed the correlation between FFM parameters and the users' behaviour in a videogame.

Wu et al. [74] focused on deriving users' personality from their implicit behavior in movie domain and hence enabling the generation of recommendations without involving users' efforts. They identified a set of behavioral features through experimental validation, and developed inference model based on Gaussian Process to unify these features for determining users' Big-Five personality traits.

In [75], the authors developed a method to predict users' personality from their Facebook profile.

Although implicit methods have the advantage to extract people's personality without disturbing people with questionnaires, as the accuracy of these techniques is not as high as those explicit methods, we will not consider these methods in this study. More implicit methods can be referred to Tkalcic's review in [62].

## 2.6. Previous Approaches and Algorithms to Personality Based Recommender Systems

Among the many Personality-Based Recommender Systems research, Rentfrow's research [7] seems to be the landmark in the field of personalized music recommendation. In this research, the relationship between personality factors and users' preference on music genre was shown. It also showed that personality can be utilized to solve the cold-start problem in RS.

Hu and Pu also conducted a number of works [12, 19–21, 76] over personality-based Recommender Systems. These works show that users' personality can be efficiently used to enhance the prediction accuracy of collaborative filtering Recommender Systems and address the cold-start problem as well. More specifically, since new users usually do not have enough overlapping rated items to calculate their similarities in collaborative filtering Recommender Systems, Hu and Pu proposed to use the Pearson correlation coefficient to calculate the user similarities based on users' personalities (FFM factors) [20].

In their study[76], they first acquired explicit FFM parameters for each user and then calculated the user distances (dissimilarities) using the Pearson correlation coefficient:

$$simp(u,v) = \frac{\sum_k (p_u^k - \overline{p_u})(p_v^k - \overline{p_v})}{\sqrt{\sum_k (p_u^k - \overline{p_u})^2 \sum_k (p_v^k - \overline{p_v})^2}} \tag{2.20}$$

where u and v are the FFM vectors for two different users and the n-dimension vector $p_u = (p_u^1, p_u^2, ..., p_u^n)^T$ is the personality descriptor (individual FFM factors) for each user. They then combined the user dis-

tance with existing rating-based user similarities $simr(u, v)$ with weight $\alpha$:

$$sim(u, v) = \alpha * simr(u, v) + (1 - \alpha) * simp(u, v) \tag{2.21}$$

They compared the proposed approach to a rating-based user similarity metric collaborative filtering recommender system and showed that the personality-based algorithm outperformed the rating-based in terms of mean absolute error, recall and specificity. They also showed that such personality-based Recommender Systems can increase users' loyalty towards the system and decrease their cognitive effort when compared with common Recommender Systems in research [19]. While one of the limitations of their experiments lies in the relatively small data set, which could make the results incidental. Moreover, the density of the user-item matrix of the dataset they used is quite low (sparse), which cannot guarantee the same result with a dataset with more more co-rated items by users. They also did not explore different domains to check whether their findings were domain-independent, only anchoring their experiment in the area of music.

Marko et al. adopted a similar approach[77], in which they defined a different metric for user distances:

$$d_W(b_i, b_j) = \sqrt{\sum_{l=1}^{5} w_l (b_{il} - b_{jl})^2} \tag{2.22}$$

where the vectors $bi = (b_{i1}, ..., b_{i5})$, $bj = (b_{j1}, ..., b_{j5})$ represent the personality values of two users $u_i$, $u_j$ and $w_l$ are the weights. They performed their experiment on a 52 users consuming 70 content items dataset to check whether the personality-based user similarity measure (USM) performs better in the cold-start problem (CSP). Their results showed that the personality USM is statistically equivalent to the rating based USM which makes it a good candidate for a complete replacement of the rating based USM. The limitation of their experiment is also similar, in which they only verified their results on a specific dataset. They cannot guarantee the presented approach is useful also in other domains.

Most of previous research are focusing on utilizing the personality information to address the cold-start problem. While in this research, we are not meant to address such problem. We will study the relation between users' personality and their diversity needs on music recommendations and use such relation to improve the diversity adjustment strategy.

## 2.7. Limitations of Previous Work

As we can see, although many studies have been conducted to show that adding personality information or adjusting the diversity degrees to the RS will lead to better performance, a few studies have combined these two things together.

For research on diversity based recommender systems, previous work usually suffer from neglecting the personalization aspect. Some of them only consider the diversity over a single item attribute but rarely consider that different users may have diversity needs on different item attributes. While others commonly set the same diversity control parameter for all users, meaning that they seldom studied whether users would be affected by their own characteristics (such as personalities) in terms of the need for recommendation diversity with different levels.

For most research on personality based recommender system, former work just mainly emphasize on addressing the cold-start problem. They also rarely consider that different people with different personalities may have different diversity needs. Only a few studies have studied whether personality information can be used to affect the diversity adjustment strategy.

Tintarev et al. [14] apply a user-as-wizard approach to study how people diversify a set of items when they recommend them to their friends, in which they particularly emphasize the personality trait "openness to experience". Their study did not prove the effect of Openness to Experience on the overall diversity participants applied, while they observed that users who are low on Openness to Experience might prefer thematic diversity to categorical variation.

Wu and Chen also conducted a series of research[13, 22] on studying the relationship between personality and diversity in the field of Movie Recommendation. In their study, they first conducted a user study aiming of identifying the relationship between personality and users' preferences for recommendation diversity. They defined the personality traits based on the popular big-five factor model and measured the diversity degrees based on Gini-index and Jaccard coefficient. Spearman's rank

| **Overall Diversity** | **Attribute's Diversity** (w.r.t. the most important attribute) | **n** (the number of diverse movies) |
|---|---|---|
| High need | High need | **7** |
| High need | Middle need | **6** |
| Middle need | High need | **6** |
| High need | Low need | **5** |
| Low need | High need | **5** |
| Middle need | Low need | **4** |
| Low need | Middle need | **4** |
| Low need | Low need | **3** |

Figure 2.1: Adjustment of the number of diverse items in [22]

correlation coefficient was used to reveal the diversity score's correlation with the user's five personality values. Results showed that significant correlations exist between some values. For instance, the "director" attribute is significantly positively correlated with the personality factor neuroticism and "country" attribute is negatively significantly correlated to the personality factors agreeableness [13]. Based on the survey findings, they then proposed a personality-based diversity-adjusting strategy for recommender systems [22]. They evaluated their system based on an online experiment, in which they tested the system's recommendation accuracy, system competence, and users' overall satisfaction. Results showed that their system obtained significantly higher evaluation scores than the normal recommender system for all three aspects.

## 2.8. Contribution

Compared with the methods mentioned above, the key contribution of our research work is the proposal of a diversity degree adjustment strategy which considers both people's diversity needs on different item attributes and with different levels.

Although Wu and Chen's work [13, 22] inspires our research work, our research is dissimilar in several aspects. Different from their work, our research work will anchor in the domain of Music Recommender Systems. Users will have more understanding of the recommended songs when listening to the previews compared to the dull text of introductions about movies in Wu's work [22]. Such understanding of the recommended items is important as it can make users' overall satisfaction evaluation to the system more accurate.

Another big difference lies in our more flexible re-ranking strategy and the way to set diversity levels. In Wu and Chen's work [22], before their system generates the recommended list, they asked about users' preferences on different item attributes and performed a conjoint analysis to infer the weight of those attributes, which then can determine the most important attribute for later re-ranking process. Their re-ranking approach is based on the diversity need of this 'most important attribute'. While their approach is limited and fixed, in which they actually constructed and combined two recommended lists: the original list generated by the original RS and a new recommended list generated according to users' diversity needs. The number of diverse items is controlled according to a fixed rule, which can be referred as Figure 2.1. Specifically, the number of diverse items are mechanically adjusted referred to the specific diversity needs of the user. As the maximum number of diverse items is fixed, their approach is not suitable if the final recommended list is enlarged.

In our strategy, the final selected attributes to perform diversity degree adjustment strategies will also be determined directly from users' personality information. Besides, a better way to adjust the diversity degrees in the final recommendation list is to add diverse items greedily and adaptively instead of setting a absolute number for diverse items. We use a re-ranking method to adjust the diverse items in the recommendation list.

In addition, we will also adopt Factorization Machine in the offline experiment. Factorization machines allow researchers to incorporate more features (such as personality) into Recommender Systems

easily. Studies [78, 79] show that it allows parameter estimation under very sparse data (which is also a problem for traditional RS) and has linear complexity at the same time. Such advantages are helpful when my training dataset is sparse and large. Personality information and other metadata (such as genre and artist attributes) can also be treated as extra features to improve the prediction accuracy of the system.

Based on the survey, our main research question is:

*RQ: How does personality information affect how diversity degrees should be applied in Music Recommender Systems?*

To address the main research question, two sub-research questions are proposed:

*sub-RQ1: Is there an underlying relationship between people's personality and their needs for recommendation diversity in Music domain?*

*sub-RQ2: What is the effect (on diversity and accuracy) of adjusting the diversity degrees in Music Recommender Systems based on users' personality information?*

# 3

# Pilot Study

The previous chapter has introduced the background and related work in the two branches of research: personality and diversity, in which we have also defined the gap as the lack of combined research in these two fields, especially in the music domain. In order to narrow this gap, we have also raised two sub-research questions to be solved.

In this chapter, we intend to discuss the first research question:

- RQ1: Is there an underlying relationship between people's personality and their needs for recommendation diversity in Music domain?

To answer this question, a pilot study is conducted, in which the relationship between users' personality factors (based on the FFM) and their diversity needs on music recommendations is explored. Generally, the user survey is designed to obtain users' music preferences (a list of preferred tracks) and their personality scores via a TIPI personality test. After computing the diversity scores on the six selected attributes (*Release Times*, *Artists*, *Number of Artists*, *Genres*, *Tempo* and *Key*) for tracks within the preferred list, we then compared these diversity scores with the corresponding personality scores (for each user) to explore the relation between users' personality factors and their diversity needs.

In the following sections, we first introduce the motivation of this pilot study and the relation we are going to research. The hypotheses based on previous research then follow. Materials used for our pilot study will also be introduced. Then we show how these materials are further explored and how users can participate in our survey. Both quantitative and descriptive results of the study will be presented. Discussion of the results and limitations of the study are also included. Finally, a conclusion of the study will be made.

## 3.1. Motivation for the Pilot Study

In this section, we explain the motivation and several hypotheses of our pilot study. The motivation of the pilot study is based on the gap we found in the previous research. While the hypotheses we made in this pilot study are also mainly based on the findings shown in previous work.

As mentioned in Chapter 2, previous work [7, 80–83] have shown that there exists a significant correlation between users' personality traits and their music preference, especially regarding music genres. For instance, in [80], Ferwerda et al. found that people's emotional state influences the type of music they are listening to. They also showed that people who are more extraverted and open to experience tend to listen to happy music when they feel sad. While for people who are neurotic (emotional unstable), they are more inclined to listen to more sad songs when they are down in spirits. In [81], Langmeyer et al. found that people who are more open to new experiences prefer reflective, complex, intense and rebellious music (e.g., classical and rock), while they dislike conventional types of music like pop music. In [83], Ferwerda et al. again found that neurotic users show more positive correlations with alternative music than any other genres of music. They also found that open users are more inclined to listen to a wide variety of music genres.

While these mentioned works all did not touch the field of users' diversity needs. Few works have directly studied the relation between people's personality traits and their diversity needs on music preferences. As mentioned in Chapter 2, in the Movie's domain, Chen and Wu [13] found that personality factors have a significantly causal relationship with users' diversity preference (on both item's individual attributes and overall combined attributes). For instance, they found that attribute "actor/actress" is positively correlated to openness while conscientiousness is significantly negatively correlated with the overall diversity. While in the Music domain, this study is missing and the correlation between users' personality factors and their music diversity preferences is unexplored. To the best of our knowledge, the only related research is Ferwerda et al's recent work [84] on studying the relation between users' personality traits and their satisfaction and attractiveness of diversified recommendation lists. They found that conscientiousness is positively related to a higher degree of diversification, while agreeableness is related to a mid-level diversity of the recommendations.

Thus, since referencing previous research is not enough to construct the relation between users' personality factors and their diversity preferences in Music domain, it is necessary for us to carry out a tailored pilot study for our own research. In this pilot study, a relation model that reflects the relation between personality factors and diversity degrees of music preference is constructed.

### 3.1.1. Hypotheses of the Pilot Study

Since the pilot study is made to verify whether there are some relations between users' personality factors and their diversity needs for music, a few hypotheses should be raised beforehand. Based on previous related research, our hypotheses are:

- H1: Personality factor **Emotional Stability** (opposite of Neuroticism) has a highly positive correlation (e.g. Spearman's Correlation Coefficient > 0.2) with users' diversity needs for **Genres**.
- H2: Personality factor **Openness to Experience** has a highly positive correlation with users' diversity needs for **Artists**.
- H3: Personality factor **Extraversion** has a mid-level positive correlation (e.g. Spearman's Correlation Coefficient around 0.15) with users' **overall diversity needs**.
- H4: Personality factor **Agreeableness** has a mid-level correlation with users' **overall diversity needs**.
- H5: Personality factor **Emotional Stability** has a highly positive correlation with users' **overall diversity needs**.

H1 and H5 are mainly based on Ferwerda's and Langmeyer's work [80, 81, 83], in which findings show that people's emotional state significantly influence the type of music they listen to.

H2 is made based on Chen and Wu's work [13]. Their work showed that personality factor Openness to Experience has a significant correlation with the diversity of Actor/Actress of the movies they prefer. Although their work was conducted in the movien domain, the correlation could be similar in the music domain.

Previous work [81] studied the relation between personality factor Extroversion and users' music preference but did not show evidence for the relation between Extroversion and users' diversity needs. Thus, our hypothesis H3 is based on the intuition that more extroverted people would prefer to listen to more types of music.

In [84], Ferwerda used the ANOVA method to investigate the impact of personality traits on participants' evaluations of the different levels of diversified music recommendation lists. They found that agreeable participants show to be more attracted to and more satisfied with the medium diversification (e.g. for satisfaction, $F(2, 11)=9.660$, $p<0.05$)) than for the high diversification (e.g. for satisfaction, $F(2, 11)=4.036$, $p<0.05$)). This finding becomes the basis of our hypothesis H4.

For personality factor Conscientiousness, previous work have different findings on the correlation between Conscientiousness and users' diversity needs. In [13], Chen et al. conducted their research in the Movie domain and found that Conscientiousness is consistently significantly correlated with users' diversity needs in the negative way, which suggests that users who are more flexible, spontaneous, disorganized, and lack in patience will be subject to choose diverse items. While in [84], Ferwerda et al. found that Conscientiousness is related to a preference for a higher degree of diversification in

music recommendation. Normally, this inconsistency could be caused by the domain difference and different diversity metrics they used. Since there exists such an inconsistency regarding the personality factor Conscientiousness, we do not make a specific hypothesis for Conscientiousness.

## 3.2. Materials of the Pilot Study

This section introduces the materials we need in our pilot study. Specifically, two types of materials are included: Tracks with associated attributes and Personality Profile.

### 3.2.1. Tracks

To further explore our users' music preference besides their listening history, we need to study the diversity needs of their preference lists. Before we compute the actual diversity scores for each of these lists (check Section 3.3.1), we need to first figure out what kind of attributes are used for computation. Since merely knowing the track ids is not enough for computing the solid diversity scores, we enriched our first material (tracks) with a number of associated attributes. Generally, six attributes of the tracks are selected. These attributes are: *Release Times*, *Artists*, *Number of Artists*, *Genres*, and two audio features (*Tempo* and *Key*). By computing the diversity degrees of each of these six attributes for the songs selected by each user, we can generally obtain users' diversity needs of their music preference. In this pilot study, the explanation of these attributes are all referred to the Spotify Web API Reference[1].

- *Release Times*: Release time is an attribute referred from *release_date* from Spotify. It is an album level attribute along with each track, which represents the date the album was first released, for example, '1981-12-15'. The precision of the release time varies among different tracks. Some precision value can be accurate to 'day', while some others might only have 'year'. To maintain the consistency of the comparison, in this pilot study, all release times are converted to the 'years' they were released. The reason to choose this attribute lies in the findings from Wu and Chen's work [13]. In their work, they found that the attributes *Release Time* and *Actor/Actress* have a positive correlation with personality factors Conscientiousness and Openness respectively. Although their research lies in the Movie Recommendation domain. The findings could be similar in the Music domain.
- *Artists*: Attribute artist is referred from the track level Spotify attribute *artists*, which stands for the artists who performed the track. A single track can include several different artists. For each track, attribute *artists* is an array of artist objects, which contain various information regarding these artists (e.g. names, ids). We only compare the ids of the these artists in this pilot study. The reason to choose our next attribute *Artists* is the same as *Release Times*.
- *Number of Artists*: This attribute represents the total number of artists for each track. Each track has different numbers of artists. Some people might prefer solo, some might prefer duet, or some even might prefer chorus. People who are more agreeable might have more diversified choices on this attribute.
- *Genres*: As mentioned before, several studies [7, 81, 82] have found that attribute *genre* has a significant correlation with users' personalities. Unfortunately, Spotify does not provide track or album level genre information. The only genre information we can use here is the artist level genre information. The main artist's genre information is used in this pilot study as the attribute *Genres*.
- *Tempo*: *Tempo* is a track level audio feature, which represents the overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. As far as we know, there is no research that has studied the relation between the diversity degrees of tracks' audio features and personality factors. Thus, the choice of these two audio features are based on the intuition that some typical musical properties (e.g. average tempo and key) may have a strong correlation with some specific music genres or artists (e.g. rock music usually have high level of tempo). Since we assume that there might exists a correlation between users' diversity needs for genres and their personality factors, such correlation might also exits with some audio features.

[1]Spotify Web API Reference: https://beta.developer.spotify.com/documentation/web-api/reference/

- *Key*: *Key* is also a track level audio feature, which represents the average key of the track. Integers map to pitches using standard Pitch Class notation [2].

For audio features, we did not include the feature *Energy*, which represents a perceptual measure of intensity and activity of the track. The reason lies in that, in Spotify, *Energy* is a mixed feature, which means that several features have contributed to this attribute including dynamic range, perceived loudness, timbre, onset rate, and general entropy. We do not know how these features are mixed and what the consequences it might lead to the computation of the correlations (e.g. different personalities may have different impacts on different features within the attribute *Energy*). Thus, we turned to another simpler feature: *Tempo*.

### 3.2.2. Personality Profile (Model and Extraction Method)

The second material for our user survey is each user's personality profile, which can be further divided into the personality model applied to them and the extraction method we used.

For the personality model, we have chosen the Big-Five Factor Model (FFM) [85]. In this model, personality is defined as five factors: Openness to Experience (O), Conscientiousness (C), Extroversion (E), Agreeableness (A), and Neuroticism (N, or Emotional Stability). The general explanation of the Big-Five Factor Model can be referred to Chapter 2. While the specific indication of the scores and analysis for each personality factor can be referred to Section 3.3.2.

For the extraction method of these personality factors, since the participants of the pilot study have limited time to finish the survey, we have chosen the Ten Items Personality Inventory (TIPI) [3], which is a short personality test inventory containing ten personality-related questions in total. The whole list of these ten questions can be found in Chapter 2. The specific way to convert the scores from the inventory to the actual personality scores for each factor can be found in Section 3.3.2.

## 3.3. Variable Computation

After we have obtained the track attributes and personality profiles from the users, since we want to study the correlation between users' personality factors and the diversity needs for music, we need further computation and analysis on our raw materials (tracks with associated attributes & TIPI scores, see Section 3.2). Specifically, for each track attribute, we need to compute the diversity score within the users' preference list; for TIPI scores, we need to convert all of the question scores into the personality factor scores based on FFM. The final two kinds of variables we are going to study for the analysis of our correlation are: **a)** the diversity scores of the six track attributes (see Section 3.3.1) and **b)** the actual personality factor scores (see Section 3.3.2).

This section includes the description of methods how we analyze the two materials obtained from users. The way to collect these information from users can be referred to the Procedure Design in Section 3.4.

### 3.3.1. Diversity Metrics

Since we are analyzing the correlation between **a)** users' diversity needs for music and **b)** personality, the first thing we need to define is how we compute the diversity degrees. Generally, two Diversity Measuring Metrics are used in our pilot study: the *Intra List Diversity* (ILD, also known as Intra List Dissimilarity) and *Shannon Entropy*. In order to obtain the overall diversity degrees of the six attributes, we also included some weighted combination methods to compute the overall diversity.

- *Intra List Diversity*: Intra List Diversity is the most commonly used diversity metric in the research of recommender systems, which is defined as the averaged pairwise distance (dissimilarity) between all items in a list. This metric is quite basic while highly dependent on the definition of item similarity. In our pilot study, it is defined as:

$$Div = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n}(1 - Similarity(c_i, c_j))}{n * (n-1)/2} \tag{3.1}$$

[2]Pitch class: https://en.wikipedia.org/wiki/Pitch_class
[3]TIPI: https://gosling.psy.utexas.edu/scales-weve-developed/ten-item-personality-measure-tipi/

where $c_1..c_n$ are items in a single list, n is the total number of items in the list. The Similarity function is defined as:

$$Similarity(c_i, c_j) = \begin{cases} 1 & \text{if } c_i = c_j \\ 0 & \text{if } c_i \neq c_j \end{cases} \tag{3.2}$$

This metric is used to compute the diversity degrees for attributes *Release Times*, *Number of Artists*, *Tempo* and *Key*.

- *Shannon Entropy*: Shannon Entropy (or Shannon index) is another diversity index used in our study. Instead of directly computing the pairwise distance between all the items in a list, it considers the proportion of an item in the whole list, which is computed as:

$$Div = -\sum_{i=1} np(i)log(p(i)) \tag{3.3}$$

The entropy is 0 when a single item is always chosen or recommended, and $log(n)$ when n items are chosen or recommended equally often. p(i) is the proportion of the item i in a list.

Considering that, for each user, the total number of artists and genres for his/her whole music preference is not small (a track may have several artists and genres), and the genre information for each track is not exactly the true genre for the track (since we used the artists' genre information), it is better to compute their proportion in the whole list instead of using the ILD directly. Thus, we used *Shannon Entropy* to compute the diversity degrees for attributes *Artists* and *Genres*.

- *Overall Diversity*: The overall diversity of the list is measured by combining all the diversity degrees of the six attributes. 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 several typical sets of weights to the six attributes in reference to [86] (e.g. one of the assignment is equal weights $\{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}$). The overall diversity is then calculated as:

$$Div_{overall} = \sum_{i=1}^{n} W_i * Div(attr_i) \tag{3.4}$$

where the $W_i$ is the weight on the i-th attribute, n is the total number of selected attributes. The whole assignments of the weights for the overall diversity in our pilot study are shown as follows:

**Overall_Div1** We assigned three different sets of weights (w.r.t. [86]) when combining the six attributes. The first assignment is called 'Equal weights method', which means that the value of the weights distribution is the equal weights vector defined by

$$w_i = 1/m, i = 1, 2, ..., m \tag{3.5}$$

$w_i$ stands for the weight for each attribute, $m = 6$ in our study. Thus, the weight assignment for Overall_Div1 is $\{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}$ on the six attributes {Release Times, Artists, Number of Artists, Genres, Tempo, Key}.

**Overall_Div2** The second assignment is called 'Rank-order centroid (ROC) weights', which considers the rank order of the attributes. Each weight is calculated as:

$$w_i = \frac{1}{m}\sum_{k=1}^{m}\frac{1}{k}, i = 1, 2, ..., m \tag{3.6}$$

Thus, the weight assignment for Overall_Div2 equals to {0.41, 0.24, 0.16, 0.10, 0.06, 0.03}.

Table 3.1: TIPI scale scoring ('R' denotes reverse-scored items)

| Personality Factor (FFM) | Question in TIPI |
|---|---|
| Extraversion | 1, 6R |
| Agreeableness | 2R, 7 |
| Conscientiousness | 3, 8R |
| Emotional Stability | 4R, 9 |
| Openness to Experiences | 5, 10R |

**Overall_Div3** The third assignment is called 'Rank-sum (RS) weights', in which each attribute is weighted in proportion to its position in the rank order. That is,

$$w_i = \frac{m+1-i}{\sum_{k=1,m} k} = \frac{2(m+1-i)}{m(m+1)}, i = 1, 2, ..., m \tag{3.7}$$

where (i) is the rank position of attribute i, and $\sum_i w_i = 1$. Thus, the weight assignment for Overall_Div3 equals to {0.29, 0.24, 0.19, 0.14, 0.09, 0.05}.

### 3.3.2. Personality Analysis

The other variable is the personality profile. Here we describe how to convert the TIPI question scores into the actual personality scores. Initially, since we are adopting the TIPI personality test, we can only obtain users' raw answers to the ten questions (ratings of 1-7 for each question). The way to convert the question scores to the personality scores is described as follows (using TIPI scoring [4]):

1. Each personality factor is assessed by two questions in TIPI as described in Table 3.1. First, we recode the reverse-scored items (i.e., recode a 7 with a 1, a 6 with a 2, a 5 with a 3, etc.). The reverse scored items are 2, 4, 6, 8, & 10.
2. Take the AVERAGE of the two items (the standard item and the recoded reverse-scored item) that make up each scale.

A short example using the Extraversion scale: A participant has scores of 5 on question 1 (Extraverted, enthusiastic) and and 2 on question 6 (Reserved, quiet). First, we recode the reverse-scored question (i.e., question 6), replacing the 2 with a 6. Second, we take the average of the score for question 1 and the (recoded) score for question 6. So the TIPI Extraversion scale score would be: (5 + 6)/2 = 5.5.

The actual personality score for each factor is also ranging from 1 to 7. The scores of each factor can be further mapped into different personality levels which are compared with norms based on a sample of 1813 respondents from the research of Gosling's [69] [5]. The full mapping can be referred to Table 3.2. Below we list the general indication of these scores for each factor according to the psychological study reported in [87]:

- **Extroversion**: Low scores on Extroversion indicates that the person is introverted, reserved, and quiet. While high scores indicates that the person is sociable, outgoing, energetic, and lively.
- **Agreeableness**: Low scores on Agreeableness indicates that the person less concerns with others' needs than with his/her own. While high scores indicates that the person is pleasant, sympathetic, and cooperative.
- **Conscientiousness**: Low scores on Conscientiousness indicates that the person is careless and disorganized. While high scores indicates that the person is reliable and hard-working, who usually sets clear goals and pursue them with determination.

[4] TIPI scoring: https://gosling.psy.utexas.edu/scales-weve-developed/ten-item-personality-measure-tipi/, retrieved date: June 3, 2018

[5] TIPI excel spreadsheet: http://gosling.psy.utexas.edu/wp-content/uploads/2014/09/excelscoreTIPI.xls, retrieved date: June 3,2018

Table 3.2: Mapping of the range of personality scores (1-7) to levels [69]

| | Low | Medium Low | Medium High | High |
|---|---|---|---|---|
| **Extroversion** | 0 - 2.99 | 2.99 - 4.44 | 4.44 - 5.89 | >5.89 |
| **Agreeableness** | 0 - 4.12 | 4.12 - 5.23 | 5.23 - 6.34 | >6.34 |
| **Conscientiousness** | 0 - 4.08 | 4.08 - 5.4 | 5.4 - 6.72 | >6.72 |
| **Emotional Stability** | 0 - 3.41 | 3.41 - 4.83 | 4.83 - 6.25 | >6.25 |
| **Openness** | 0 - 4.31 | 4.31 - 5.38 | 5.38 - 6.45 | >6.45 |

- **Emotional Stability/Neuroticism**: Low scores on Emotional Stability (high scores on Neuroticism) indicates that the person is sensitive and emotional, who is easily upset. While high scores (low scores on Neuroticism) indicates that the person is exceptionally calm, composed and unflappable.
- **Openness to Experience**: Low scores on Openness indicates that the person is down-to-earth, practical, and conservative. While high scores indicates that the person is curious, imaginative, and creative.

## 3.4. Procedure Design

This section describes how the users take part in the user survey and how users' data are collected. The survey is web-based, in which the users take a short personality test (TIPI) and select some preferred songs.

### 3.4.1. Survey Design and Participation Procedure

To effectively collect users' personality information and music preference, a website [6] is designed for the user survey. The website is constructed using Flask framework [7] and deployed on the Heroku Cloud Application Platform [8]. Generally, four main parts are included in the survey:

- User's Basic Information: The first part of the survey collects some of users' basic information including their age range, gender, education level, nationality and profession. We also ask about how often they use the the online music service and recommendation service normally.
- Personality Test: As mentioned, 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). The whole list of the ten questions can be found in Chapter 2. Through analyzing the scores of these ten questions, we can map them into the actual scores for each personality factors (see Section 3.3.2).
- Music Preference Collection: Users' music preference is collected by means of Spotify Web API, which means that users need to log into their Spotify account in our user survey. After users have logged into their Spotify account, maximally 50 recently listened songs will be displayed to them. Users are asked to select at least 20 preferred songs that they normally listen to and can best describe their music taste. In case that their preferred songs are not in the recently played list, a search function is also provided for the users to search for the songs they like. In addition, users are also asked to rate their selected songs from 1 to 5 (least preferred to most preferred).
- User Comment: A user comment section is also included in our survey, in which users can comment on the problems they have encountered during the survey or any additional suggestions. This section is in the form of free text, and users can choose to skip this section.

[6] Survey address: https://music-rs-personality.herokuapp.com
[7] Flask: http://flask.pocoo.org
[8] Heroku: https://www.heroku.com/

All participation in this research study is voluntary. All participants should go through the four parts one by one in order to complete the survey. Users may choose to interrupt during the survey. We only use the data from the users who have completed the whole procedures.

### 3.4.2. Data Gathering Channels

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) [9]. 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.

### 3.4.3. Ethical Clearance

All our pilot study procedures are approved by the Delft Human Research Ethics Committee (HREC) [10]. For the whole project, we submitted the research application to the HREC which included all the procedures we were going to conduct before we conducted the actual user survey. A risk assessment was also included in the application, which discussed some ethical issues this research might contain. We also included all the key questions (e.g. personality test) in the survey and the consent form offered to users. After several weeks' examining and waiting, our research was successfully approved by HREC.

**Consent Form** The consent form offered to users can be accessed via the first page of our survey website [11]. In the form, we first describe the purpose/nature of this research and why participants are recruited. We declare that the participation of survey is voluntary and users can choose not to participate and withdraw at any time. We also describe the whole procedures users need to take during our survey. All users are informed that their information is confidential and anonymous (including IP addresses). The results of this study will only be used for the creation of scholarly publications. All users need to agree with the consent form before they can take the survey.

**Ethics Related Designs** To eliminate any ethical/legal issues, we only store the data related to our research purpose, which includes

- Users' basic information: age, gender, nationality, profession, education level, frequency of using online music service, frequency of using music recommendations and whether they like the recommendations.
- Personality test results: ten question scores for the TIPI test.
- Users' preferred songs: at least 20 songs (Spotify track id) selected by users and their corresponding ratings.
- User comments

Users need to log into their Spotify account during the survey, in which we utilize the Spotify Web APIs. All users' logging information will not be stored. All the information we collected from users cannot be used to infer who the users are.

**Design Flaw** When we collect users' basic information, we have collected some personal information that is not used in our study, which includes the frequency of using online music service, frequency of using music recommendations and whether they like the recommendations. We planned to study the correlation between these factors with personality factors. While since these factors (e.g. frequency of using online music service) are not the main research scope of our work, we did not used these factors in the end.

[9] Crowdflower: https://www.figure-eight.com
[10] Delft Human Research Ethics Committee (HREC): https://www.tudelft.nl/over-tu-delft/strategie/strategiedocumenten-tu-delft/integriteitsbeleid/human-research-ethics/
[11] Consent form: http://music-rs-personality.herokuapp.com

Table 3.3: Demographic profiles of 148 participants (numbers in the bracket stand for the total number of users for each case).

| | |
|---|---|
| **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) |
| **Profession Domain** | Student (53); Enterprise (62); Institution (15); Others (18) |

Table 3.4: Personality distribution in Pilot Study v.s. general population distribution in Gosling's work [69]

| | Pilot Study | General Population in Gosling's work [69] |
|---|---|---|
| Extraversion Mean (std) | 4.05 (1.25) | 4.44 (1.45) |
| Agreeableness Mean (std) | 5.05 (1.17) | 5.23 (1.11) |
| Conscientiousness Mean (std) | 5.16 (1.27) | 5.40 (1.32) |
| Emotional Stability Mean (std) | 4.78 (1.28) | 4.83 (1.42) |
| Openness to Expreriences Mean (std) | 5.11 (1.07) | 5.38 (1.07) |

## 3.5. Results

Following the procedures mentioned in Section 3.4 and variable computation methods mentioned in Section 3.3, we conducted our user survey and investigated the Spearman's rank correlation coefficient between users' personality factors and their diversity needs for music. In this section, we display the results of our pilot study. Both quantitative and descriptive results are shown.

### 3.5.1. Participants

Generally, around 185 people are recruited to participate in the survey via the two channels mention in Section 3.4.2. Most of them (155 users) are recruited via Crowdflower, the rest of them are students from several universities. We conducted our first round of data-selection with the help of the filtering functions provided by Crowdflower. From our database, we can see that more than 1400 crowdcourcing workers have started our survey, while most of their data are discarded at the very beginning due to several reasons. For instance, some of them have not finished all the procedures; some of them provided the wrong verification codes; some of them just did not spend enough expected time on our survey (more than 4 minutes).

After the second round of data-selection and data-cleaning, totally 148 users' data are used for the analysis to construct the relation model. The reasons to discard these data are varied: most of them (around 30 users) are discarded due to the violation of setup of the test question; some of them (around 3 users) are discarded due to the obviously random filling (e.g. select all 7 for all TIPI questions); some of them (around 2 users) are discarded due to the obvious contradiction on some specific questions (e.g. some one says he scores 7 on 'Extraverted' and 7 on 'Reserved' at the same time); the rest them (around 2 users) are discarded due to some other reasons like maliciously filling out the survey for several times.

The general demographic properties of these 148 participants are shown in Table 3.3. All participants are volunteered to take part in the survey or recruited from the crowdsourcing platforms.

In addition, we also studied the personality distribution for participants in our pilot study. We compared it with the general population personality distribution reflected in Gosling's work [69], which involves 1813 users' personality information. The result is shown in Table 3.4, which is in line with the distribution we expect of the general population.

Table 3.5: Spearman Correlation coefficient between personality factors/demographic values and diversity degrees w.r.t. single attribute (p-value is obtained from the Spearman Rank-Order Correlation Coefficient Test, *p-value<0.05 and **p-value<0.01)

| | Div(Release times) | Div (Artists) | Div(Artists number) | Div (Genres) | Div (Tempo) | Div (Key) |
|---|---|---|---|---|---|---|
| **Extraversion** | -0.03 | 0.10 | 0.00 | 0.07 | 0.11 | **0.21**** |
| **Agreeableness** | -0.12 | 0.09 | **0.25**** | 0.00 | 0.09 | 0.05 |
| **Conscientiousness** | 0.01 | 0.11 | 0.13 | -0.01 | 0.11 | 0.06 |
| **Emotional Stability** | 0.11 | **0.22**** | 0.15 | **0.25**** | **0.24**** | 0.17* |
| **Openness** | -0.15 | -0.04 | 0.07 | 0.06 | 0.08 | 0.08 |
| **Gender** | 0.00 | -0.03 | 0.06 | 0.06 | -0.17* | -0.13 |
| **Age** | **0.28**** | -0.16 | -0.14 | 0.03 | -0.02 | -0.10 |

Table 3.6: Spearman Correlation coefficient between personality factors/demographic values and overall diversity (p-value is also obtained from the Spearman Rank-Order Correlation Coefficient Test, **p-value<0.01)

| | Overall_Div1 | Overall_Div2 | Overall_Div3 |
|---|---|---|---|
| **Extraversion** | 0.11 | 0.11 | 0.12 |
| **Agreeableness** | 0.09 | 0.08 | 0.06 |
| **Conscientiousness** | 0.08 | 0.08 | 0.07 |
| **Emotional Stability** | **0.31**** | **0.28**** | **0.29**** |
| **Openness** | 0.03 | 0.01 | 0.02 |
| **Gender** | 0.01 | 0.00 | 0.00 |
| **Age** | -0.05 | -0.10 | -0.09 |

### 3.5.2. Relation between Personality Factors and Diversity Degrees of Music Preference w.r.t. Single Attribute

When studying the correlation between personality factors and each attribute's diversity degrees, we first calculated the personality scores for each user according to the method we introduced in Section 3.3.2. Then, we computed the diversity scores for each attribute within the list of tracks a user has selected according to the metrics discussed in section 3.3.1. 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 3.5.

From the results, we can see that the personality factor ***Emotional Stability*** does have a significant correlation with the diversity degree of the attribute ***Genres***, which means that our hypothesis **H1** holds in this case. While the factor ***Openness*** seems to have no significant correlation with any of our selected attribute, thus we cannot provide support for the hypothesis **H2** in our pilot study.

### 3.5.3. Relation between Personality Factors and Overall Diversity of Music Preference

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. The three different overall diversity scores are combined and computed according to different weighting assignment methods discussed in Section 3.3.1.

Results for the correlation between the overall diversity and user's personality values are shown in

Table 3.6. From the results, we can see that there exits a significant correlation between personality factor ***Emotional Stability*** and the **overall diversity** no matter how the attributes' weights are varied, which means that our hypothesis **H5** holds in our experiment. While our survey does not show much evidence for correlation between the overall diversity and factors ***Extroversion*** and ***Agreeableness***. This means that we also cannot find support for our hypotheses **H3** and **H4** in our pilot study.

### 3.5.4. User Comments

Although most of the user comments consist of positive feedback on our user survey, we still can find some unexpected problems in practice. Here, we list some of the interesting comments made by participants regarding their feelings about the user survey:

- Two of our users (1% of the whole sample) found that some functions (drag and drop of songs) do not work properly with Firefox browser.
- One of our users also found that it is not so clear how to remove the selected songs from the selected songs pool.
- One user commented that it is not very user-friendly that the page will always go back to the top of the page every time he searched for a song.
- Two of the users (1% of the whole sample) also complained that there are many duplicated songs in their recently listened songs' list. This is actually a problem from Spotify since I just simply used Spotify API to grab those songs with user's authorized accessing token.
- Three of our users (2% of the whole sample) commented that it is quite hard to rate his/her selected 20 songs since every song is already his/her favourite.

## 3.6. Discussion

**Validation of Hypothesis H1.** In our hypothesis **H1**, we consider whether the personality factor **Emotional Stability** (opposite to Neuroticism) is highly correlated with users' diversity needs for **Genres**. This assumption is validated by our pilot study. As we can see from Table 3.5, Emotional Stability has a significant positive correlation with diversity degrees of 'Artists', 'Genres' and audio feature 'Tempo', which means that people who are more emotional stable prefer more diversification in 'Artists', 'Genres' and 'Tempo'. This suggests that people who are exceptionally calm, composed and unflappable are more inclined to listen to diverse genres of music. They might also prefer playlists with diverse artists and tracks with different beats.

**Validation of Hypothesis H5.** Besides the correlation with the diversity needs for attribute **Genres**, from Table 3.6, we also see that **Emotional Stability** is positively correlated with the **overall diversity** no matter how the attributes' weights are varied, which supports our hypothesis **H5**. This means that people who are more sensitive and emotional (score low on Emotional Stability or score high on Neuroticism) are inclined to listen to less diverse music in general. This finding is in line with the research conducted by Ferwerda et al. [80, 83], in which they found that neurotic users tend to listen to sad songs when they are in negative emotional state. While neurotic users would prefer angry and fearful music when they are feeling angry or disgusted. In addition, for genre preference, Ferwerda also found that neurotic users tend to prefer alternative music than any of the other genres. In some degree, our pilot study supports such conclusion, showing that neurotic users do prefer less diversification on music preference in general. The reason behind this phenomenon may lie in that neurotic users are inclined to listen to the same kind of music to maintain their original emotional state.

**Validation of Hypothesis H2.** Our second hypothesis **H2** supposes that personality factor **Openness to Experience** is highly positively correlated with users' diversity needs for **Artists**. This assumption is intuitively acceptable and verified by Chen's research in Movie domain [13]. Intuitively, people who are more open to new experiences should be more likely to enjoy novelty, variety, and change since they are more curious, imaginative, and creative. In [13], Chen et al. found that personality factor Openness to Experience is positively correlated with the diversity needs for actor/actress

in movie recommendations. In [83], Ferwerda et al. found that open users tend to listen to a wide variety of music genres.

However, unfortunately, we cannot provide support for this hypothesis in our pilot study. We did not found significant correlation between Openness and any of these six attributes. This may be due to some external factors. Firstly, our user sample is not that big enough due to our financial budget. Secondly, some defects on the metadata of the attributes may also cause the correlation not obvious. For instance, for the attribute 'Release Times', the precision of this attribute varies for each track. Due to the incomplete metadata crawled from Spotify, we have to convert all the 'Release Times' into 'Years'. Since the music world changes quickly, better intervals used to compute diversity degrees between different tracks may lies in 'Seasons' or even 'Months'. Our 'Genre' metadata is also not ideal due to the difficulty to find exact track genre information, especially for newly released songs.

**Validation of Hypotheses H3 and H4.** For the third and fourth hypothesis **H3** and **H4**, we assume that the personality factors **Extraversion** and **Agreeableness** have a mid-level correlation with users' **overall diversity** needs. These two assumptions are partially verified by our study, in which we do find these two factors are correlated with some of the attributes while not with the overall diversity. For **Extraversion**, our study shows that it is positively correlated with the diversity needs of the audio feature 'Key', which suggests that people who are more sociable, outgoing, and energetic prefer music lists with different average pitches. For **Agreeableness**, we found that this personality factor is significantly correlated with the diversity of number of artists in users' music preference. This is a positive correlation, which indicates that people who less concern with others' needs and well-being prefer songs with less diversification on the singing forms (solo, duet or chorus). Normally, people are more tough, critical, and uncompromising when they score low on Agreeableness, which may cause them to prefer more the kind of music they are originally fond of.

**Regarding personality factor Conscientiousness.** For personality factor **Conscientiousness**, as mentioned in Section 3.1.1, previous work have different opinions on the correlation between Conscientiousness and users' diversity needs. While in our pilot study, we did not find any significant correlation between personality factor Conscientiousness and users' diversity needs on music preference. Like the case of personality factor **Openness**, this result still cannot lead to the conclusion that these two things have no correlations at all mainly due to the limited size of our user survey.

**Extra demographic findings.** Besides the five personality factors, we also found that demographic value '**Age**' is highly positively correlated with the attribute '**Release Times**', which suggests that people who are younger prefer less diversity on the freshness of the music they are listening to. This result is also intuitive due to the fact that young people tend to prefer more recent music. Old people may be more nostalgic, which leads to the preference of adding more old songs to their play-lists, causing their lists be more diverse. This finding is also helpful in the designing of our further Diversity Adjusting Strategy.

## 3.7. Limitations

As mentioned in the Section 3.6, one of the limitations in our user survey lies in the limited size of the user sample. Due to this reason, we did not find support for some correlations in our hypotheses (for instance, for the personality factor Conscientiousness and Openness).

Another limitation lies in the availability of the attributes (features) used for music preference. In our pilot study, since the exact genre information for tracks is impossible to obtain via the Spotify Web API, we adopted the artist's genre information. Although these genre information still reflect the correlation with some personality factors, these features could be not that accurate to reflect the actual genres of the specific track. For the attribute (feature) 'Release Times', as mentioned in Section 3.2.1, we have converted all of them into 'years', which is also not the best choice if the best intervals used to compute diversity degrees for 'Release Times' between different tracks lies in 'Seasons' or even 'Months'.

One additional limitation is the selection of the audio features in our user survey. To the best of our knowledge, there is no previous research that has studied the relation between audio features and personality factors in the music domain. Chances are that there exists some correlation between other

audio features (like 'Energy' and 'Loudness') with the personality factors that we have not studied. Due to the time and budget limitation, we leave these for further research.

## 3.8. Conclusion

This pilot study has studied the relation between people's personality factors and their diversity needs in the music domain. For now, we have discussed and studied the first research question:

- RQ1: Is there an underlying relationship between people's personality and their needs for recommendation diversity in Music domain?

Through the results of our study, we can see that there exists significant correlations between some of people's personality factors and diversity degrees of some specific attributes of the tracks, especially when we consider the personality factor '**Emotional Stability**'. Moreover, we see that **Emotional Stability** is also significantly correlated with users' overall diversity needs for music. These findings are partially correspond to Ferwerda et al.'s former work [80, 83], which showed that users' personality factors (especially Emotional Stability) are correlated with their music preferences (especially on genres).

For the next step, our research shifts to the designing of the specific Diversity Adjusting Strategy for Music Recommender Systems according to the findings in this pilot study. The system will incorporate personality factors as a moderating factor into the adjusting of diversity degrees within the final recommendation lists. Our next research question is:

- RQ2: What is the effect (on diversity and accuracy) of adjusting the diversity degrees in Music Recommender Systems based on users' personality information?

Our general strategy is to first using an existing recommendation algorithm (either via a Factorization Machine or a Collaborative Filtering Recommender System) to generate an initial recommendation list for the user. Then, we apply a re-ranking strategy to the list based on our relation model. The diversity of the list will be adjusted according to the specific personality level of the user. A detailed explanation of the system can be supplied in the next Chapter.

# 4

# Diversity Adjusting Strategy

In the previous chapter, we have conducted a pilot study to investigate the relationship between people's personality factors and their diversity needs in music recommendations. We found that there exits a significant correlation between these two objects via our user survey, especially when we consider the personality factor 'Emotional Stability'.

Since we have found such correlation between personality factors and users' diversity needs, our next mission is to explore how we can apply such correlation in the diversity adjustment strategy in Music Recommender Systems (RecSys). Thus, in this chapter, we are going to address our second research question:

- RQ2: What is the effect (on diversity and accuracy) of adjusting the diversity degrees in Music Recommender Systems based on users' personality information?

To answer this question, we propose an adaptive diversification algorithm by incorporating users' personality factors into a re-ranking function. The re-ranking method is a diversification algorithm which re-orders the recommendation list by balancing both similarity and diversity, which we will have a recap in Section 4.1 and discuss it in details in Section 4.2. Given a user's specific personality information, we can then use this algorithm to adjust the individual diversity degree for him/her in Music Recommendations.

In the following sections, we first offer a recap of the general re-ranking method that is widely used in the research of diversity-based recommender systems. Then, we show how we incorporate uses' personality information into the re-ranking function in order to adjust diversity degrees according to users' specific personalities. After that, we introduce the whole Diversity Adjusting Strategy algorithm in details. The corresponding offline evaluation of this proposed personality-based diversification method follows. The online evaluation of this algorithm will be followed in the next chapter.

## 4.1. Recap of the Re-ranking Diversification Method

Our Diversity Adjusting Strategy is based on the idea of the re-ranking diversification algorithm. In Chapter 2, we have had a basic introduction to the re-ranking diversification method and reviewed several related papers that used such diversification model. In this section, we look into this method a little bit further.

In conventional recommender systems, recommendation lists are generated by sorting the predicted ratings of unseen items in descending order. Top-N items (N top-ranked items) in this sorted list are usually recommended as the final list. Generally, this method satisfies offline metrics (such as prediction accuracy) but has little impact on user satisfaction. Rearranging the sequence of the final list in order to include some diversified or surprising items into the top-N list may increase the user satisfaction and interests compared to the similar items to users' original interests. In order to pop up the diverse items from the bottom of the list to top-N list, we need a method to 're-rank' the recommendation list.

As mentioned in Chapter 2, most re-ranking algorithms usually focus on achieving an optimal balance between two objectives: accuracy and diversity. To achieve this, the idea of Maximal Marginal Relevance (MMR) [88] from Information Retrieval (IR) used for search diversity to the RecSys re-ranking

task is widely used. Other methods are also used, but we only focus on MMR in this research. This algorithm itself is a greedy approach with the following objective function:

$$Q(c, P, R) = Sim(c, P) * (1 - \lambda) + \lambda * Div(c, R) \tag{4.1}$$

where the $Q(c, P, R)$ represents the final quality of the item c to the user's original interests P in recommendation list R. $Sim(c, P)$ stands for the similarity function to compute the similarity between item c and users' original interests P and $Div(c, R)$ represents the diversity of c relative to those items so far selected in the re-ranked list $R = \{r_1, ..., r_m\}$. $\lambda$ is used for controlling the balance between the similarity function and the diversity function. High $\lambda$ means that the function will lean more to the diversity side, while low $\lambda$ means that the function will lean more to the similarity side for the recommendations.

As we can see, the final diversified list of recommendations using the re-ranking method is created by reordering recommendations after they have been generated by other recommendation algorithms. This means that the such re-ranking methods assume that initial recommendations are already diverse and just need to be reordered in order to achieve the maximum possible diversity levels.

## 4.2. Diversification Algorithm

In the last section, we have had a recap of the re-ranking diversification method, which uses an objective function to control the balance between similarity and diversity. We also used a similar objective function in our diversification algorithm to control the diversity degrees for recommendations. The difference is that, in order to adaptively adjust the diversity degrees for users with different personalities, we further incorporate users' personality information into the objective function. In this section, we reveal how we apply this objective function on our diversification algorithm.

Normally, the recommendation process of a RecSys can be divided into two steps: first the RecSys generates the predicted values for all unrated items for each user and secondly these items are sorted in order to meet some specific metric. The second step is to sort the items 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 [16]. Specifically, greedy heuristics are used in our work, which have been demonstrated to be efficient and effective [16, 89].

This greedy algorithm will iteratively selects an item from the original list O and then puts it at the end of the current re-ranked list R until the size of R meets a size N (N equals to 10 in our case) and the re-ranking process is complete. The core of the greedy algorithm lies in the objective function 4.1 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, the size of the input list should 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 $\lambda$ in equation 4.1 is controlled by personality factors in our algorithm. To adjust the diversity degrees more flexibly, we also introduce three parameters $\theta_1$, $\theta_2$, and $\theta_3$ to control the computation of the overall diversity. All of these four parameters ($\lambda$, $\theta_1$, $\theta_2$, $\theta_3$) are affected by the personality factors. The specific way of computing these four parameters will be introduced in Section 4.3. More explanation on the objective function can also be found in Section 4.2.1. The diversification algorithm is shown in Algorithm 1.

As mentioned, at each step (line 2), we select an item that minimizes the the objective function defined in 4.1 (line 4) and add it to the current re-ranked list R (line 5). We do not touch the first item of the original list (line 1) based on the assumption that the top ranked item is always good.

### 4.2.1. More Explanation on the Objective Function

In previous sections, we have already had a general impression on the composition of the objective function and how it is used in our diversification algorithm. In this section, we look into this function a little bit further.

Given this re-ranking objective function 4.1, we are going to incorporate the personality information into this function. In this section, we show how we define the similarity function $Sim(c, P)$ and diversity

**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 $\lambda$, $\theta_1$, $\theta_2$, ..., $\theta_n$)
**Output:** Top-N re-ranked list R

$R(1) \Leftarrow O(1)$
**while** |R| < N: **do**
  $Div_{overall}(c,R) = \sum_{i=1,2,...,n} \theta_i * Div_i(c,R)$
  $c^* = \text{argmin}_{c \in O \setminus R}\, Obj(c,R) = Sim(c,P) * (1-\lambda) + \lambda * Div_{overall}(c,R)$
  $R = R \cup \{c^*\}$
  $O = O \setminus \{c^*\}$
**end while**
**return** R

function $Div(c,R)$ in our own scenarios. Then, in Section 4.3, we reveal the principles of how we adaptively adjust the diversity degrees of the this re-ranking function by considering users' personality information via controlling the parameters such as $\lambda$.

**Similarity.** The left part of the function $Sim(c,P)$ considers the similarity aspect of the item c to users' initial interests P. Normally, the similarity values can be presented in two ways:

- The predicted rating of item c to user u (e.g. ranging from 1 to 5);
- The rank of item c in the final list according to their predicted ratings sorted in the descending order.

Previous work defined their similarity function differently according to their own scenarios. For instance, Ziegler et al. [16] used the rank of the item as the $Sim(c,P)$ function in their topic diversification method. While, people like Di Noia et al.[89] and Vargas et al. [90] prefer to use the rating estimation of the item as the $Sim(c,P)$ function. In our own task, considering that in our online evaluation we used Spotify Music Platform, which does not provide a rating function, so we choose to use the rank of items as our similarity measurement for the specific items. Thus, our $Sim(c,P)$ function becomes:

$$Sim(c,P) = Rank(c,O) \tag{4.2}$$

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

**Diversity.** 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 $R$. Recalling that in our Pilot Study (Chapter 3, Section 3.3.1) we defined the overall diversity as the weighted combination of several diversity degrees for different track attributes, we also utilize similar overall diversity definition (see Chapter 3, Section 3.3.1) here in our diversity adjusting algorithm. But the items used for computing the attribute diversity are a little bit different. The new diversity function is defined as follows:

$$Div_{overall}(c,R) = \sum_{i=1,2,...,N} \theta_i * Div_i(c,R) \tag{4.3}$$

where N represents the total number of attributes we used for computing the overall diversity $Div_{overall}(c,R)$, $\theta_i$ represents the weight for each attribute diversity degree. $Div_i(c,R)$ represents the different diversity degrees for different attributes, which can be further defined as:

$$Div_i(c,R) = \frac{\sum_{j=1}^{n}(1 - Similarity(c,R_j))}{n} \tag{4.4}$$

where $R_1..R_n$ are items in the current re-ranked list $R$, n is the total number of items in the list. The Similarity function is defined as (for one specific track attribute):

$$Similarity(c,R_j) = \begin{cases} 1 & \text{if } c = R_j \\ 0 & \text{if } c \neq R_j \end{cases} \tag{4.5}$$

In our final overall diversity computation, we used three attributes, which are exactly the attributes that have close correlations with the personality factors found in our Pilot Study. These attributes are **Genre**, **Number of Artists (the total count of Artists for a track)** and **Key (Audio Feature)**. Since in our Pilot Study, we found that **Genre** has a positive correlation with the personality factor **Emotional Stability**, **The total count of Artists** has a positive correlation with **Agreeableness**, and the audio feature **Key** has a positive correlation with **Extraversion**, we believe that using these three attributes as a combination can give distinct diversity degrees for people with different personalities.

## 4.3. Variables

In the previous section, we have given a detailed explanation on our diversification algorithm. However, we still have not explained how we specifically adjust the personality-related parameters $\lambda$, $\theta_1$, $\theta_2$, $\theta_3$ in the objective function. Moreover, since our diversification algorithm is built upon a re-ranking algorithm, its final diversity degree is also affected by some re-ranking related parameters such as the size of the input list. In this section, we introduce all of the variables that might have an impact on the accuracy or the diversity degrees for the final re-ranked list. Offline evaluations on the impact of these variables will be introduced in Section 4.4.

### 4.3.1. Re-ranking Related Parameters

Some parameters in the re-ranking algorithm itself might have a direct impact on the final diversity degrees of the re-ranked list. These parameters are:

- **The size of the final Top-N re-ranked list (N).**
  The size of the final re-ranked list N might have a impact on both the accuracy and the diversity of the recommendation list. Intuitively, the more items contained in the recommendation list, the more likely that the recommendation list will be more diverse and contain more items users may prefer.

- **The size of the input list (LS).**
  The size of the input list Ls has a similar impact on the recommendation diversity as parameter N. Intuitively, when we use more items for re-ranking, it is more likely that the final re-ranked list will be more diverse.

- **The size of the unrated items used for testing (K).**
  Different from LS and N, K is a parameter defined in our Testing Methodology. We will have a detailed explanation on this parameter in Section 4.4.4.

### 4.3.2. Personality Related Parameters

In Section 4.2.1, we have defined our similarity function and diversity function. But we still have not incorporated the personality information into this equation 4.1. Furthermore, in equation 4.1 and equation 4.3, we still have four parameters $\lambda$, $\theta_1$, $\theta_2$, $\theta_3$ left for further assignment. The influence of our personality information is exactly exerted on these four parameters. In our former Pilot Study (Chapter 3, Section 3.5), the findings are shown in two aspects:

- For single attribute diversity, we have:
  **Correlation 1** Personality factor Extraversion has a positive correlation with the diversity degree of the audio feature Key.

  **Correlation 2** Personality factor Agreeableness has a positive correlation with the diversity degree of Artists Number.

  **Correlation 3** Personality factor Emotional Stability has a positive correlation with the diversity degrees of Artist, Genre and Tempo.

- For overall diversity, we have:
  **Correlation 4** Personality factor Emotional Stability has a positive correlation with the overall diversity degree.

Based on these correlations, the four parameters of equation 4.1 and 4.3 are adjusted according to the following rules:

- For parameter $\lambda$ in equation 4.1, since it is correlated with controlling the weight of the overall diversity degree, according to the correlation 4, it should be adjusted depending on the personality factor **Emotional Stability**. Noted that for each personality factor, we mapped it into four different levels: Low, Medium Low, Medium High, and High (The mapping method from personality scores to these four levels can be found in Chapter 3, Section 3.3.2). Thus, the mapping from Emotional Stability into $\lambda$ can be defined as:

Table 4.1: Mapping from Emotional Stability to $\lambda$

| **Emotional Stability Level** | Low | Medium Low | Medium High | High |
|---|---|---|---|---|
| $\lambda$ | 0.2 | 0.4 | 0.6 | 0.8 |

As shown in Table 4.1, when the Emotional Stability Level becomes higher, the $\lambda$ also becomes larger, which will cause the equation 4.3 play a more important role within the equation 4.1.

- Parameters $\theta_1$,$\theta_2$, and $\theta_3$ in equation 4.3 are correlated with the single attribute diversity (Genre, Artists Number, and Key). Here, $Div_1(c,R)$, $Div_2(c,R)$, and $Div_3(c,R)$ represent the diversity degree for Genre, Artists Number, and Key separately. According to Correlation 1,2, and 3, parameters $\theta_1$,$\theta_2$, and $\theta_3$ should be adjusted depending on personality factors Emotional Stability, Agreeableness, and Extraversion separately. The mapping functions from these personality factors into $\theta$ values are also similar to Table 4.1:

Table 4.2: Mapping from Emotional Stability (ES)/Agreeableness (A)/ Extraversion (E) to $\theta_1/\theta_2/\theta_3$

| **Personality ES/A/E Level** | Low | Medium Low | Medium High | High |
|---|---|---|---|---|
| $\theta_1/\theta_2/\theta_3$ | 0.2 | 0.4 | 0.6 | 0.8 |

Different from $\lambda$, we take one more computation step for $\theta_1/\theta_2/\theta_3$: Normalization. Thus, the final $\theta_1/\theta_2/\theta_3$ are computed as follows:

$$\theta_i = \frac{\theta_i}{\sum_{j=1,2,3}\theta_j}, \quad i = 1,2,3 \tag{4.6}$$

As shown in Table 4.2, when the Emotional Stability/Agreeableness/Extraversion Level are higher, the $\theta_1/\theta_2/\theta_3$ also become larger, which will cause the corresponding attribute diversity function $Div_i(c,R)$ play a more important role within the objective function 4.3.

## 4.4. Offline Evaluation

In the previous section, we have introduced our diversity adjusting strategy in details, in which we used a greedy re-ranking algorithm to adjust the diversity degrees specifically for people with different personalities. We incorporated users' personality information into an objective function that balances the final list's similarity and diversity degrees. To test the effectiveness and efficiency of our proposed re-ranking method, we designed both offline and online evaluation. The online evaluation can be found in the next chapter. In the offline evaluation, in order to generate initial recommendations with high quality, we used a Recommender System called Factorization Machine (FM). Since our re-ranked list is generated based on the initial recommendation list, better quality of the initial recommendation list will increase the quality of the re-ranked list as well.

In this section, we introduce the offline evaluation of our diversity adjusting strategy. Recalling that our second research question studies how diversity degrees in Music Recommender Systems can be adjusted according to users' different personalities, we want to know whether our proposed personality-based diversification algorithm can adjust the diversity degrees appropriately for different users. Since

the re-ranking results will be greatly influenced by different re-ranked related parameters like the size of the input list (LS), or the size of the final top-N list (N), we first tested the effect of these re-ranked related parameters on our re-ranking method. Furthermore, in our objective function 4.1 and 4.3, we exert the influence of the personality factors mainly on parameters $\lambda$, $\theta_1$, $\theta_2$ and $\theta_3$, which will also greatly influence the final diversity degrees. The offline evaluation also allows us to study the impact of these parameters.

### 4.4.1. Factorization Machine

Before we introduce the offline evaluation procedures, since we are using the Factorization Machine, we first give a little introduction to this state-of-the-art RecSys. We will also show the advantages of using FM compared with conventional RecSys such as Collaborative Filtering Recommender Systems.

Factorization Machine is a general-purpose supervised machine learning algorithm raised by Rendle et al. [78, 79] in 2010. Served as an extension of a linear model, it is designed to capture the interactions between high-dimensional features in extremely sparse datasets, which is exactly the case in the recommendation problem. The basic idea of Factorization Machine is to learn a polynomial kernel by representing high-order terms as a low-dimensional inner product of latent factor vectors. By utilizing feature engineering, Factorization Machine has proven to be an extremely powerful tool with enough expressive capacity to generalize other factorization models such as Matrix/Tensor Factorization.

#### Mechanisms

Most recommendation problems assume that we have a rating dataset formed by a collection of <user, item, rating> tuples (or user-item matrix). This is also the foundation of many current popular recommendation algorithms such as Collaborative Filtering algorithms and Content-based algorithms, which have proven to yield really nice results. However, besides these user-item matrixes, we still have plenty of metadata (such as track attributes, context information) unused for predictions. In conventional recommendation algorithms, it is hard to directly incorporate these information to yield better prediction results. While in Factorization Machines, feature incorporation becomes much easier since in FMs extra features can be included in the model in a natural way and higher order interactions can also be modelled.

A typical second order (here, order means the highest degree of feature interactions) FM model is defined as follows:

$$\hat{y}(x) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} < v_i, v_j > x_i x_j \tag{4.7}$$

Where the $v_i$, $v_j$ represent k-dimensional latent vectors associated with each variable (i.e. users and items) and the bracket operator $< v_i, v_j >$ represents the inner product. Parameters $w_0$ and $w_i$ are the global bias and the weights (associated with each variable) to be learned during the training process. While $x_i$ represents the value of each variable.

The first part of equation 4.7 models the linear interactions while the nested sum captures pairwise interactions. The effect of pairwise interactions $w_{ij}$ is modelled as the inner product:

$$w_{ij} =< v_i, v_j >= \sum_{z=1}^{k} v_{iz} v_{jz} \tag{4.8}$$

Factorization Machines can be used for both classification and regression tasks. For regression tasks, the model is trained by minimizing the squared error between the model prediction $\hat{y}(x)$ and the target value y:

$$L(\hat{y}(x), y) = (\hat{y}(x) - y)^2 \tag{4.9}$$

For binary classification tasks, it is trained by minimizing the crosss entropy loss:

$$L(\hat{y}(x), y) = -\ln \sigma(\hat{y}(x) y) \tag{4.10}$$

Where $\sigma$ represents the sigmoid function. To avoid the overfitting problem, $L^2$ regularization is also applied in both cases.

### Advantages & Limitations

As mentioned, one of the advantages of FM lies in its computational efficiency on large sparse data sets than traditional algorithms like linear regression. In real-life recommendation problems, the user count and item count are typically very large while since users do not rate all items, the actual user-item matrix is really sparse, which could be a problem for traditional recommendation problems like the Collaborative Filtering algorithms. While FM model equation can still be computed in linear time. Training and prediction in FMs are faster compared to other factorization models while their performance is still competitive.

Since FMs utilize feature engineering, it is also able to incorporate extra context information or other metadata as additional features into the machine to make better predictions, which is a efficient way to address the Cold-Start Problem. Another benefit of FMs lies in its capability for estimation of higher order interaction effects even if no observations for the interactions are available.

One limitation of FM lies in its nature of factorization models, which have to be devised individually for each problem and each set of categorical predictors [91]. FMs also cannot be used for datasets with implicit or unary feedback. FMs are also not optimized for ranking, which is also why we need to re-rank the list afterwards.

We used fastFM [1] [92] as the basic Factorization Machine in our offline evaluation.

## 4.4.2. Datasets

We used two datasets in our offline evaluation:

1. Pilot Study Dataset collected from our own Pilot Study (discussed Chapter 3). This dataset contains:

   - Number of all Users: 148
   - Number of distinct Tracks: 3071
   - Number of total Ratings: 3465 (ratings ranging frim 1-5)

   In this dataset, each user has rated at least 20 tracks. We also know the personality information for each user.

2. A complementary dataset with much larger user data: The Echo Nest Taste Profile Subset (TPS) [2] [93]. The original TPS dataset contains:

   - 1,019,318 unique users
   - 384,546 unique songs
   - 48,373,586 <user, song, play count> triplets

   The original TPS dataset is quite huge to be dealt with. There are also a lot of useless user data in TPS. In order to tailor this dataset to our own scenario (focusing on users with sufficient listening history), we made a few data selection beforehand. We first ruled out those tracks that have only been listened once, which means that all the tracks in our dataset should be listened by at least two distinct users. We made this rule because in recommendation problems, such isolated tracks are no help for recommending similar songs to other users. Secondly, we ruled out those users who listened less than 100 tracks in total. In Di Noia's work [89], they conducted their re-ranking algorithm in the Movie Domain and used Movielens 1M [3] for experiment. They concentrated on users who gave at least fifty ratings to construct their own dataset. In the Music Doamin, users are more possible to listen to more songs than movies. Thus, in our own case, we concentrated on users who gave at least 100 ratings. 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 [94]. Applying these two rules and the rating mapping, as a consequence, the final dataset contains:

   - Number of all Users: 4060

[1] Github fastFM: https://github.com/ibayer/fastFM
[2] The Echo Nest Taste Profile Subset: https://labrosa.ee.columbia.edu/millionsong/tasteprofile
[3] Available at http://grouplens.org/datasets/movielens

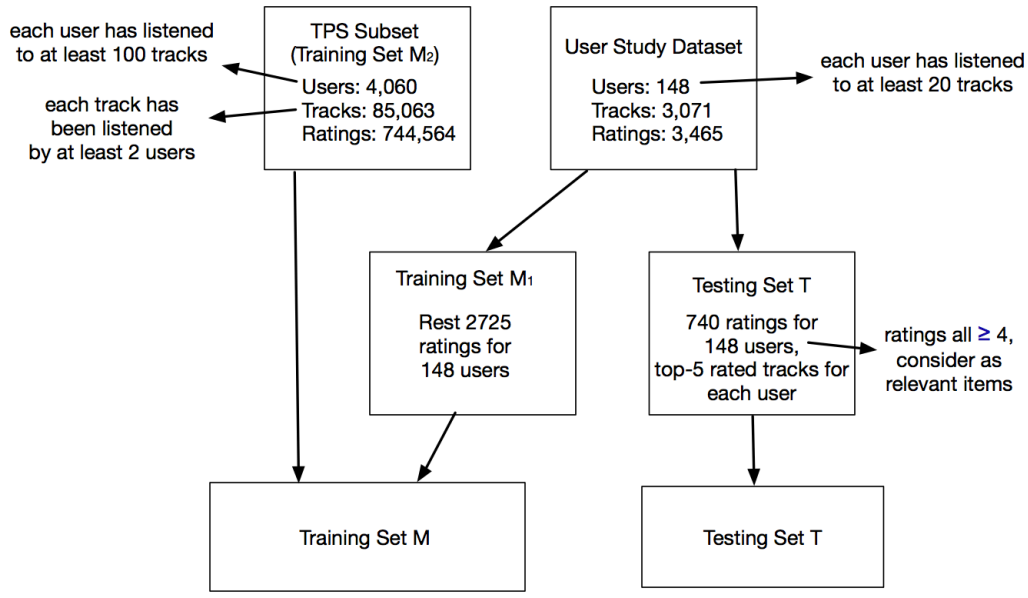



Figure 4.1: The splitting and combining process of the training and testing sets

- Number of distinct Tracks: 85,063
- Number of total Ratings: 744,564 (ratings ranging from 1-5)

These two datasets (Pilot Study Dataset and TPS subset) are used for training and testing the RecSys (FM). Specific training and testing process can be found in the following two sections.

### 4.4.3. Training the RecSys

Normally, in offline recommender system evaluation methodologies, a subset of the known ratings is usually held off as the ground truth for testing [95]. This subset of known ratings usually play the role of known relevance in the accuracy computation: highly rated (usually higher than a threshold value) items are considered as relevant items, while unrated items are taken as non relevant.

Following this idea, we first split our Pilot Study Dataset into two subsets: Training Set $M_1$ and Testing Set T. The Testing Set T contains the top-5 rated tracks (ratings all $\geq 4$) for each user, which we will consider as the relevant items to each user. The remaining user data of the Pilot Study Dataset forms the first Training Set $M_1$. We use the whole TPS subset as the second Training Set $M_2$, which is again combined with $M_1$ to form our whole Training Set M. The whole splitting and combining process can be referred to Figure 4.1.

What should be noted here is that, although the TPS Subset is large, this dataset is out of date. The TPS Dataset was constructed before 2012, which means that all the tracks contained in this dataset are released before 2012. This situation becomes a problem when we combine the TPS Subset with our own Pilot Study Dataset. All the data in our Pilot Study Dataset are collected in 2018, means that it is highly possible that most of the tracks in our own Pilot Study Dataset are released after 2012. When we use TPS Subset for training the RecSys and use the subset of our Pilot Study Dataset for testing, it is highly possible that there exists an item Cold-Start Problem when we make recommendations for users.

To relieve the potential effect of the item Cold-Start Problem, we fully utilized the benefits of bringing Factorization Machines as our RecSys. We added two time-independent track metadata as additional features when training the FM: the audio features Tempo and Loudness. By adding these two extra features, item similarities can be better compared. We then feed the whole Training Set M into the FM for training the RecSys.

### 4.4.4. Testing Methodology

After training the RecSys, we used this RecSys to generate recommendations for users in the Testing Set. In this section, we introduce the metrics we used for evaluate the effectiveness of the recommendations.

#### Hit Rate

The first metric we used in our offline evaluation is defined as hit rate, which measures the accuracy of the recommendations. Given the large item count (number of distinct tracks) and the small number of listening history per user, we used a similar methodology given in Cremonesi et al. [96]. Instead of using all unseen items (all items not used for training for each user) for prediction and 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 other items that this user has not rated. We assume that these unrated items will not be of interest to user u, representing the irrelevant items. The task of the RecSys is then to rank these K+1 items for each user. We take the top-N items from the ranked list and evaluate the corresponding performance. The whole measurement can be referred as follows:

1. For each item i in Testing Set T, we randomly select K (e.g. K = 100) additional items unrated by user u. We assume that these unrated items will not be of interest to user u, representing the irrelevant items.
2. We predict the ratings for these K+1 items.
3. We rank the K+1 items according to their predicted ratings from the highest to lowest.
4. We then generate the two recommendation list for each user:
   - To construct the first recommendation list $L_1$, we simply select the Top-N rated items as the initial recommendation list.
   - For the second recommendation list $L_2$, we first re-rank the Top 5*N ( 5*N < K) items using our personality-based diversification algorithm. Then we select the Top-N rated items as the final re-ranked list.
5. We check whether this item i is in the Top N list $L_1$ and $L_2$. If in, we consider it as hit, if not, we consider it as miss. Chances of hit will increase with N. When N equals to K, we always have a hit.
6. We repeat this process (process 1-5) for all the items (740 items for 148 users) in the Testing Set T. And we count the total number of hit for all the 740 items.
7. The final hit rate is computed as:

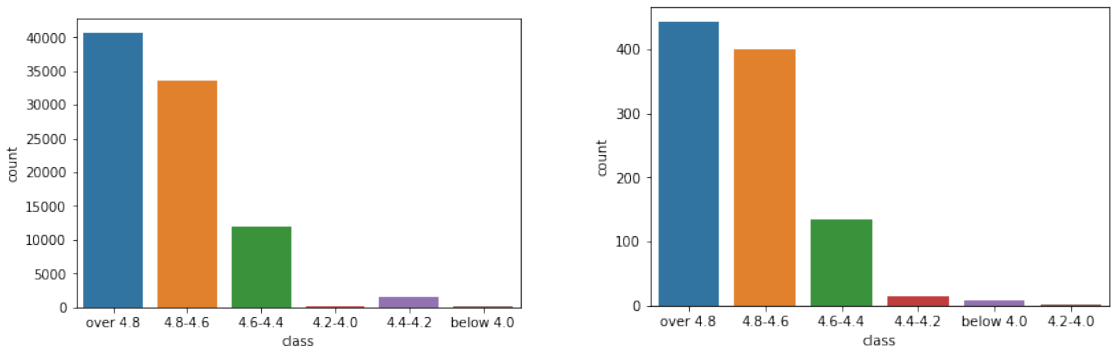
$$H(N) = \frac{\#hit}{|T|} \tag{4.11}$$

   where |T| is the total number of items in the Testing Set T and $\#hit$ is the total number of hits for all the items.

#### Why do we use K unrated items instead of all of them?

We use this testing methodology for different reasons. On the one hand, for each user, we used only around 15 <user,track,rating> triplets for training and used 5 other triplets for testing. Given the large unrated item count (around 88,000 distinct tracks), it is possible that all the testing items would have a low rank (with lower predicted ratings generated by the RecSys) in the list since the training may not be sufficient for users. On the other hand, even for items with highly predicted values, since the total number of predicted items is so large, it is still possible that these items will not rank in the Top-5 or Top-10 lists, which means that in these cases, we will end up with very tiny and hard-to-compare recall/precision values given the huge item count.

So, in sum, if we use all of the unrated items to make recommendations for each user in the Testing Set, we could obtain a very tiny hit rate value for both lists given the huge item count (around 85000 tracks in total). This is not what we want since it is meaningless to compare two very tiny values.




(a) Using all unrated items for recommendation

(b) Using K unrated items for recommendation (K=1000)

Figure 4.2: Comparison between using all unrated items and using K unrated items for predictions. X-label values represent the rating range of one specific item and the y-label values represent the total of number of items in this rating range.

To relieve such effect, we reduced the number of unrated items for recommendation to a relatively smaller K (e.g. K=1000). In this case, we enlarged the hit rate difference between the initial list and the re-ranked list in Top-N recommendations.

Here is an example to illustrate the predicted rating distribution for all the items used for recommendation in the two cases: a) using all unrated items for recommendation; b) using K (relatively smaller) unrated items for recommendation.

**Example** For user 'u1148' in the Testing Set, the five predicted ratings (ranging from 1-5) for his/her relevant items in the Testing Set are: [4.97022497, 4.88030287, 4.88370363, 4.86271008, 4.86455691]. Figure 4.2(a) and Figure 4.2(b) show the comparison of the rating distributions for one recommendation for this user in two cases.

From Figure 4.2, we see that when we use all the unrated items for prediction, around half of the unrated items (≈40000) are rated over 4.8, which means that it is highly possible that the five relevant items will not be ranked in the top-10 list since around half of the items are in the same rating range. In this case, the hit rates for both initial list and re-ranked list could be very low.

While if we decrease the number of unrated items used for recommendation as shown in Figure 4.2(b), the distribution of the rating range does not change much, while the total number of highly rated items decreases a lot. This will give us two reasonable hit rate values for comparing the quality of the two lists.

#### Intra List Diversity (ILD)

Besides hit rate, we also compare the diversity degrees for both recommendation lists. The diversity metric we used is still the Intra List Diversity. For quick review, we still provide its computation way here. The computation of overall diversity is similar to the overall diversity degree computation we defined in 4.3. The difference lies in that, for each attribute diversity, the new diversity value Div is computed as follows:

$$Div = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n}(1 - Similarity(c_i, c_j))}{n * (n-1)/2} \tag{4.12}$$

where $c_1..c_n$ are items in the list, n is the total number of items in the list. The Similarity function is defined as:

$$Similarity(c_i, c_j) = \begin{cases} 1 & \text{if } c_i = c_j \\ 0 & \text{if } c_i \neq c_j \end{cases} \tag{4.13}$$

### 4.4.5. Hypotheses

Before we run the offline evaluation procedures, we have made a few hypotheses:

- H1: Larger size (N) of the final Top-N list results in a higher hit rate for both recommendation lists.

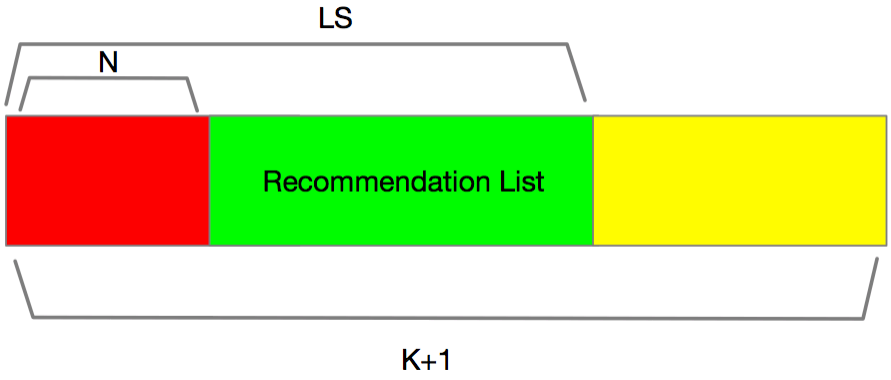



Figure 4.3: Relation between the size of final Top-N list N, the size of the re-ranked list LS, and the size of the total predictions K+1

- H2: Larger K value (size of the selected unrated items for prediction) results in a lower hit rate for both recommendation lists.
- H3: Larger size (LS) of the re-ranked list (the list used for applying the re-ranking algorithm) results in a higher hit rate for both recommendation lists.
- H4: Our re-ranked list performs better than the initial list both in hit rates and diversity degrees.

The relation between N, LS, and K can be referred to Figure 4.3.

### 4.4.6. Results

In this section, we show the results of our offline evaluation. Since some parameters in our personality-based diversification algorithm will greatly influence the final diversity degrees of the recommendation lists, for each possible and meaningful parameter, we first tested its influence on the final hit rate for the whole Testing Set. To compare the diversity degrees of the two lists, when generating each recommendation list at step 4 in the testing methodology, we also computed the Intra List Diversity (ILD) for each list.

We tested five kinds of parameters in total:

- Three re-ranking related parameters: the size of final Top-N list (N), the number of unrated items used for prediction (K), and the size of the input list used for re-ranking (LS).
- Two kinds personality-related parameter: $\lambda$ and $\theta_i$.

After we tested the impact of the parameters, we conducted a full comparison of lists with different restrictions on those parameters.

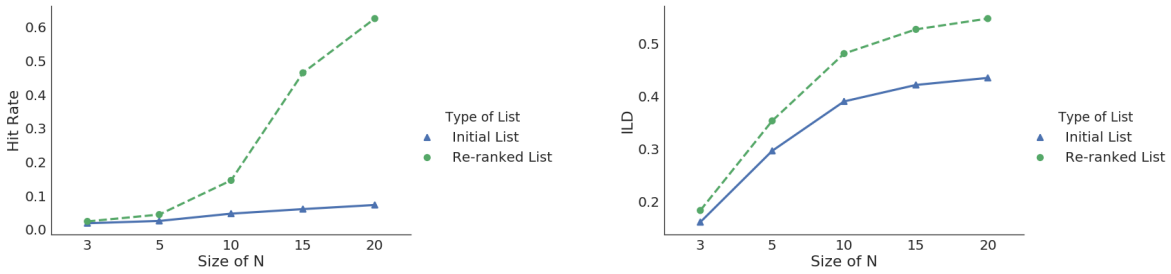
Here we show the influence of different parameters on the final hit rate and ILD on both lists.

#### The influence of the size of final Top-N list (N)

To test the influence of N, we fixed K = 100, Re-rank List Size LS = 5*N. The initial list represents the Top-N list directly generated from the RecSys. The re-ranked list is the list which is generated using our diversification algorithm. The results are shown in Table 4.3. Visualization of the results are shown in Figure 4.4.

Table 4.3: Influence of N on the Hit rate and ILD for both lists

| | N = 3 | | N = 5 | | N = 10 | | N = 15 | | N = 20 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 |
| Initial List | 0.018 | 0.161 | 0.024 | 0.296 | 0.046 | 0.485 | 0.059 | 0.421 | 0.072 | 0.434 |
| Re-ranked List | 0.023 | 0.183 | 0.043 | 0.353 | 0.145 | 0.481 | 0.464 | 0.527 | 0.624 | 0.547 |




(a) Hit Rate comparison for different N

(b) ILD comparison for different N

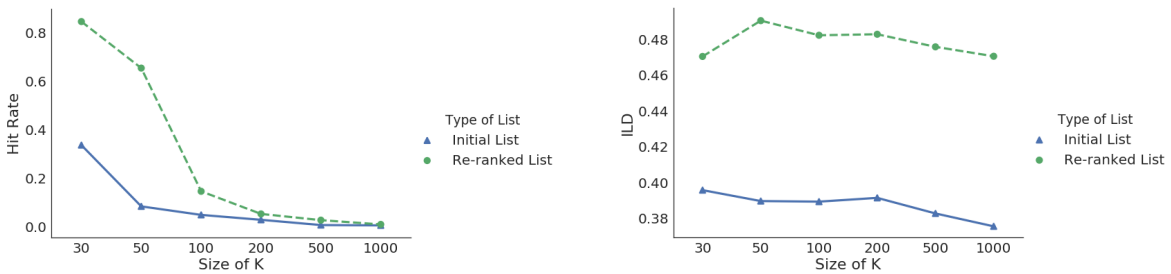
Figure 4.4: Visualization of the influence of N on the Hit rate and ILD for both lists

### The influence of the number of selected unrated items used for prediction (K)

To test the influence of K, we fixed N = 10, Re-rank List Size LS = 5*N. The initial list represents the Top-N list directly generated from the RecSys. The re-ranked list is the list which is generated using our diversification algorithm. The results are shown in Table 4.4. Visualization of the results are shown in Figure 4.5.

Table 4.4: Influence of K on the Hit rate and ILD for both lists

| | K = 30 | | K = 50 | | K = 100 | | K = 200 | | K = 500 | | K=1000 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 |
| Initial List | 0.335 | 0.396 | 0.082 | 0.390 | 0.047 | 0.389 | 0.027 | 0.391 | 0.005 | 0.383 | 0.004 | 0.375 |
| Re-ranked List | 0.845 | 0.470 | 0.653 | 0.490 | 0.145 | 0.482 | 0.051 | 0.483 | 0.026 | 0.476 | 0.008 | 0.470 |



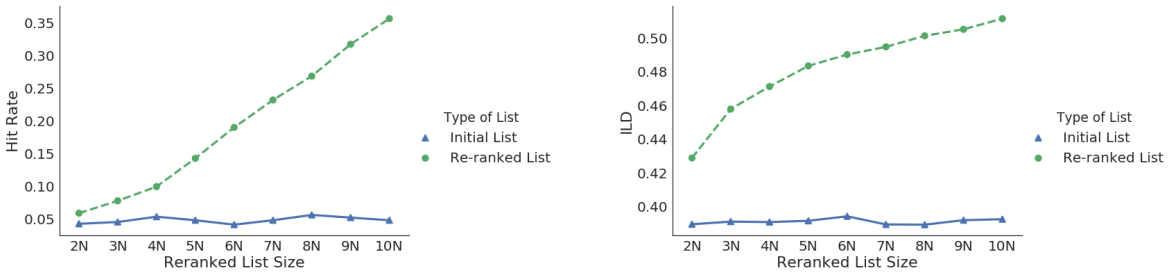

(a) Hit Rate comparison for different K

(b) ILD comparison for different K

Figure 4.5: Visualization of the influence of K on the Hit rate and ILD for both lists

Table 4.5: Influence of LS on the Hit rate and ILD for both lists

| | Rerank List Size | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2N | | 3N | | 4N | | 5N | | 6N | | 7N | | 8N | | 9N | | 10N | |
| | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 |
| Initial List | 0.042 | 0.389 | 0.045 | 0.391 | 0.053 | 0.390 | 0.047 | 0.391 | 0.041 | 0.394 | 0.047 | 0.389 | 0.055 | 0.389 | 0.051 | 0.392 | 0.047 | 0.392 |
| Re-ranked List | 0.058 | 0.429 | 0.077 | 0.458 | 0.099 | 0.471 | 0.142 | 0.483 | 0.189 | 0.490 | 0.231 | 0.495 | 0.268 | 0.501 | 0.316 | 0.505 | 0.355 | 0.511 |




(a) Hit Rate comparison for different LS

(b) ILD comparison for different LS

Figure 4.6: Visualization of the influence of LS on the Hit rate and ILD for both lists

### The influence of the size of the input list used for re-ranking (LS)

To test the influence of LS, we fixed N = 10, K = 100. The initial list represents the Top-N list directly generated from the RecSys. The re-ranked list is the list which is generated using our diversification algorithm. The results are shown in Table 4.5. Visualization of the results are shown in Figure 4.6.

### The influence of the personality-based parameter $\lambda$

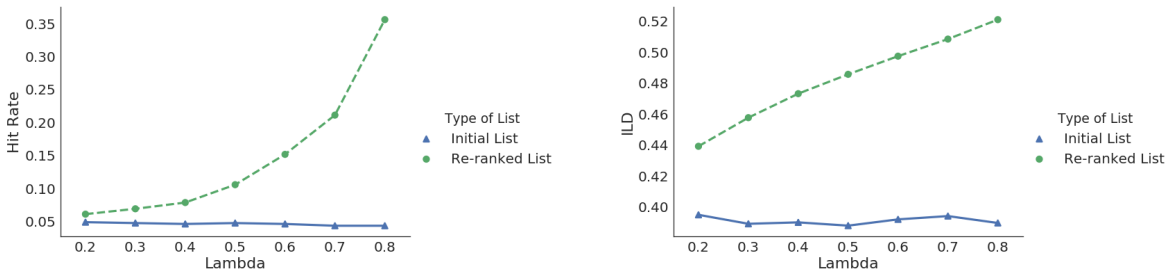
Apart from the parameters N, K, and LS, we also evaluate the influence of the personality-based parameter $\lambda$ in our objective function 4.1. To test the influence of $\lambda$, we fixed N = 10, K = 100, re-ranking list size LS = 5*N. The initial list represents the Top-N list directly generated from the RecSys. The re-ranked list is the list which is generated using our diversification algorithm ($\theta$ changes according to user's personality, $\lambda$ is tested from 0.2 -0.8). The results are shown in Table 4.6. Visualization of the results are shown in Figure 4.7.

### The influence of the individual personality-based parameter $\theta_1$, $\theta_2$, and $\theta_3$

Besides the personality-based parameter $\lambda$ which influences the overall diversity degree, we also evaluate the influence of the individual personality-based parameters $\theta_1$, $\theta_2$, and $\theta_3$ on single attribute diversity. To test the influence of $\theta_1$, $\theta_2$, and $\theta_3$, we we fixed N = 10, K = 80, reranking list size LS = 5*N, $\lambda = 0.5$. The initial list represents the Top-N list directly generated from the RecSys. The re-ranked list is the list which is generated using our diversification algorithm. For each $\theta_i$ (e.g. $\theta_1$), we tested it from 0.2-0.8. The other two $\theta_i$ (e.g. $\theta_2$ and $\theta_3$) are fixed to 0.33 for each round of test. Here, $\theta_1$, $\theta_2$, and $\theta_3$ mainly influence the diversity degrees for *Genre*, *Artist Number*, and *Key* respectively. We also recorded the overall diversity degrees in the experiment. While we found that the overall diversity degree changes little when we adjust the parameter $\theta_i$. Thus, we only provide the diversity degrees (ILD) for single attribute here. The results are shown in Table 4.7, Table 4.8, and Table 4.9. Visualization of the results are shown in Figure 4.8.

Table 4.6: Influence of $\lambda$ on the Hit rate and ILD for both lists

| | $\lambda = 0.2$ | | $\lambda = 0.3$ | | $\lambda = 0.4$ | | $\lambda = 0.5$ | | $\lambda = 0.6$ | | $\lambda = 0.7$ | | $\lambda = 0.8$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 |
| Initial List | 0.049 | 0.395 | 0.047 | 0.389 | 0.046 | 0.390 | 0.047 | 0.388 | 0.046 | 0.392 | 0.043 | 0.394 | 0.043 | 0.389 |
| Re-ranked List | 0.061 | 0.439 | 0.069 | 0.457 | 0.078 | 0.473 | 0.105 | 0.486 | 0.151 | 0.497 | 0.211 | 0.508 | 0.355 | 0.521 |




(a) Hit Rate comparison for different $\lambda$

(b) ILD comparison for different $\lambda$

Figure 4.7: Visualization of the influence of $\lambda$ on the Hit rate and ILD for both lists

Table 4.7: Influence of $\theta_1$ on the Hit rate and ILD (single attribute diversity) for both lists. Here, $\theta_1$ is correlated with the personality factor *Emotional Stability*. ILD represents the diversity degrees for the single attribute *Genre*.

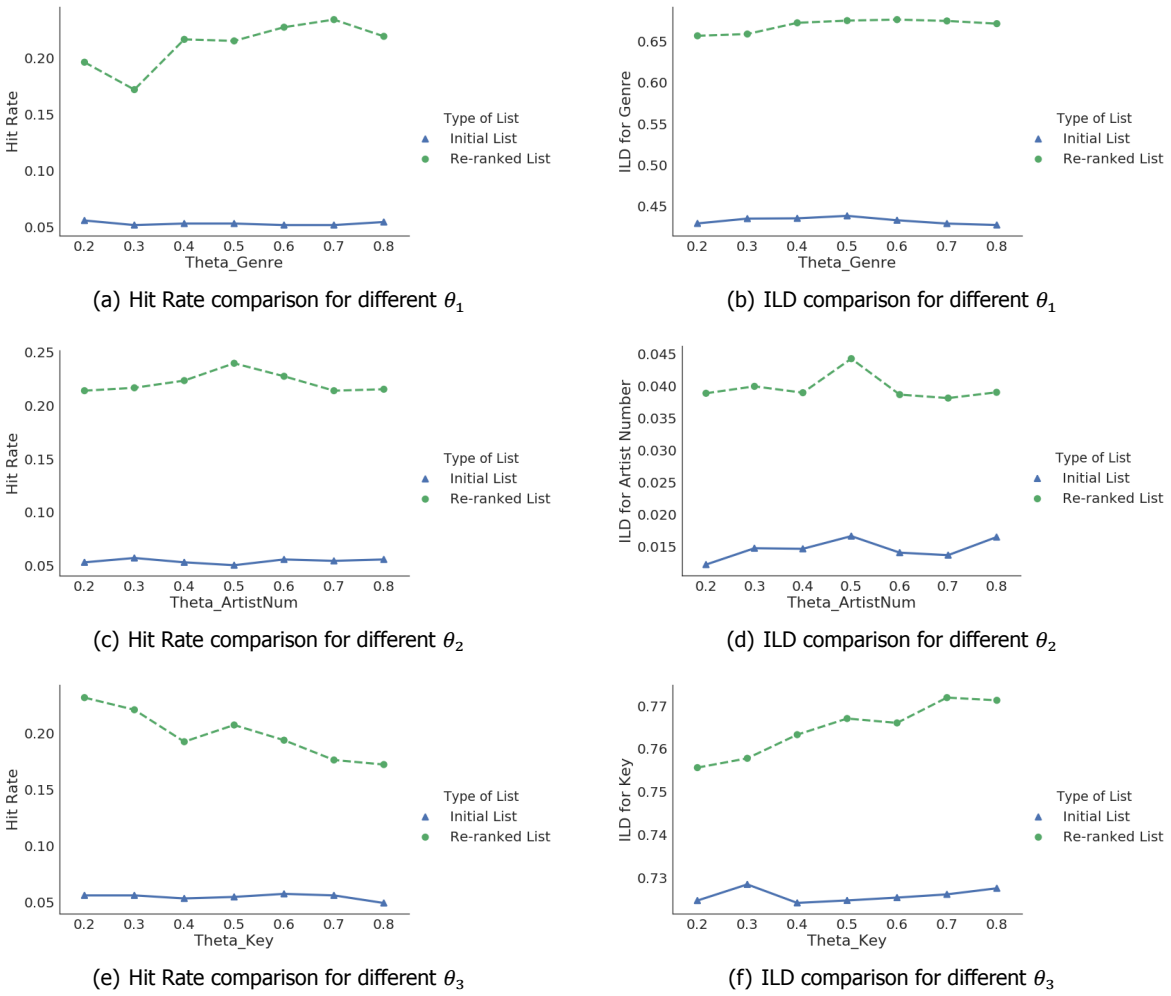
| | $\theta_1 = 0.2$ | | $\theta_1 = 0.3$ | | $\theta_1 = 0.4$ | | $\theta_1 = 0.5$ | | $\theta_1 = 0.6$ | | $\theta_1 = 0.7$ | | $\theta_1 = 0.8$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 |
| Initial List | 0.055 | 0.429 | 0.051 | 0.435 | 0.053 | 0.435 | 0.053 | 0.438 | 0.051 | 0.433 | 0.051 | 0.429 | 0.054 | 0.427 |
| Re-ranked List | 0.196 | 0.656 | 0.172 | 0.658 | 0.216 | 0.672 | 0.215 | 0.675 | 0.227 | 0.676 | 0.234 | 0.674 | 0.219 | 0.671 |

Table 4.8: Influence of $\theta_2$ on the Hit rate and ILD (single attribute diversity) for both lists. Here, $\theta_2$ is correlated with the personality factor *Agreeableness*. ILD represents the diversity degrees for the single attribute *Artists Number*.

| | $\theta_2 = 0.2$ | | $\theta_2 = 0.3$ | | $\theta_2 = 0.4$ | | $\theta_2 = 0.5$ | | $\theta_2 = 0.6$ | | $\theta_2 = 0.7$ | | $\theta_2 = 0.8$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 |
| Initial List | 0.053 | 0.012 | 0.057 | 0.015 | 0.053 | 0.015 | 0.050 | 0.017 | 0.055 | 0.014 | 0.054 | 0.014 | 0.055 | 0.016 |
| Re-ranked List | 0.214 | 0.039 | 0.216 | 0.040 | 0.223 | 0.039 | 0.239 | 0.044 | 0.227 | 0.039 | 0.214 | 0.038 | 0.215 | 0.039 |

Table 4.9: Influence of $\theta_3$ on the Hit rate and ILD (single attribute diversity) for both lists. Here, $\theta_3$ is correlated with the personality factor *Extraversion*. ILD represents the diversity degrees for the single attribute *Key*.

| | $\theta_3 = 0.2$ | | $\theta_3 = 0.3$ | | $\theta_3 = 0.4$ | | $\theta_3 = 0.5$ | | $\theta_3 = 0.6$ | | $\theta_3 = 0.7$ | | $\theta_3 = 0.8$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 | Hit rate @10 | ILD @10 |
| Initial List | 0.055 | 0.725 | 0.055 | 0.728 | 0.053 | 0.724 | 0.054 | 0.725 | 0.057 | 0.725 | 0.055 | 0.726 | 0.049 | 0.727 |
| Re-ranked List | 0.231 | 0.756 | 0.220 | 0.758 | 0.191 | 0.763 | 0.207 | 0.767 | 0.193 | 0.766 | 0.176 | 0.772 | 0.172 | 0.771 |




(a) Hit Rate comparison for different $\theta_1$

(b) ILD comparison for different $\theta_1$

(c) Hit Rate comparison for different $\theta_2$

(d) ILD comparison for different $\theta_2$

(e) Hit Rate comparison for different $\theta_3$

(f) ILD comparison for different $\theta_3$

Figure 4.8: Visualization of the influence of $\theta_1$, $\theta_2$, and $\theta_3$ on the Hit rate and ILD (single attribute diversity) for both lists

### Comparison of the five lists

After separately evaluating the influence of these parameters (N, K, LS, and $\lambda$) on the final diversity degrees, we made a full comparison of lists with different restrictions on those parameters. We have chosen five lists for comparison:

- The first list L1 is the initial list that is directly generated by the RecSys.
- The second list L2 is a re-ranked list whose $\lambda$ is fixed to 0.5 and $\theta_1$, $\theta_2$, $\theta_3$ are all fixed to 0.33.
- The third list L3 is a re-ranked list whose $\lambda$ changes according to user's personality and $\theta_1$, $\theta_2$, $\theta_3$ are all fixed to 0.33.
- The fourth list L4 is a re-ranked list whose $\lambda$ is fixed to 0.5 and $\theta_1$, $\theta_2$, $\theta_3$ are all changed according to user's personality.
- The fifth list L5 is the re-ranked list fully using our personality-based diversification algorithm, which means that the $\lambda$ and $\theta_1$, $\theta_2$, $\theta_3$ are all changed according to user's personality.

We fixed N = 10, K = 100, re-ranking list size LS = 5*N. The results are shown in Table 4.10. Visualization of the results are shown in Figure 4.9.

Table 4.10: Comparison of the five lists on the Hit rate and ILD

| | Initial List L1 | Re-ranked List L2 | Re-ranked List L3 | Re-ranked List L4 | Re-ranked List L5 |
|---|---|---|---|---|---|
| Hit rate@10 | 0.043 | 0.105 | 0.134 | 0.104 | 0.141 |
| ILD@10 | 0.390 | 0.485 | 0.483 | 0.485 | 0.483 |

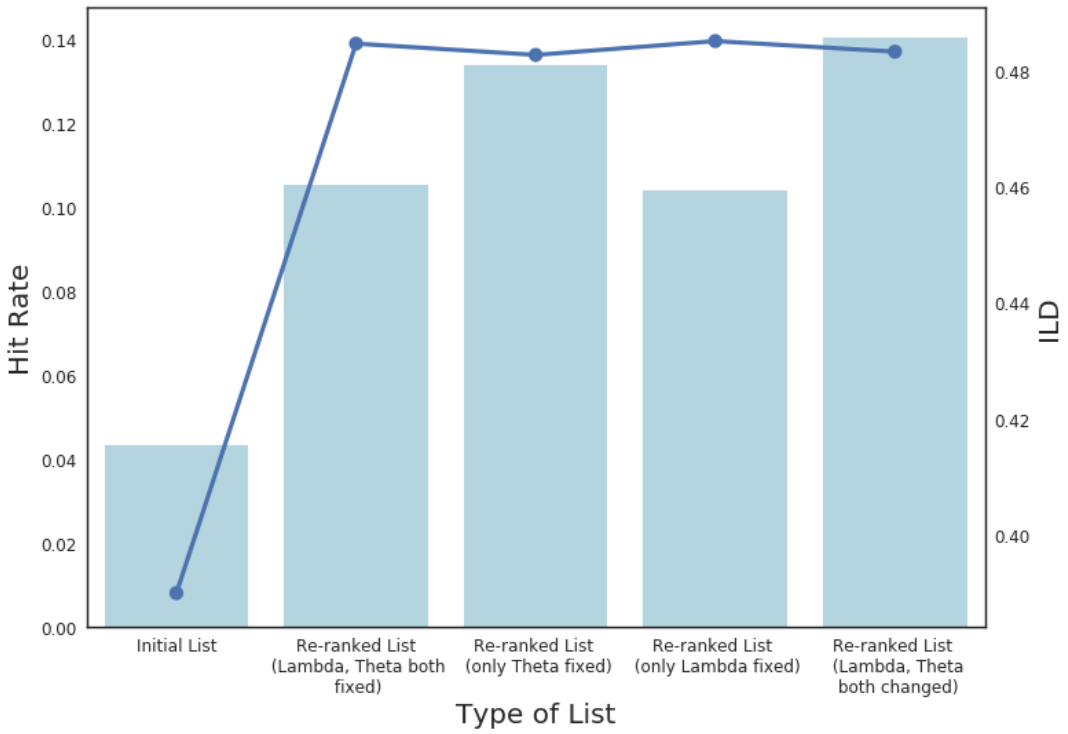



Figure 4.9: Visualization of the comparison of the five lists on the Hit rate and ILD. The bars show the hit rate of the five lists. The dot and the line show the ILD of the five lists. Xlabels from left to right represent L1, L2, L3, L4, and L5.

### Comparison with the incorrect personality information

We also compared the re-ranked list ($L_2$) generated using the correct personality information with the re-ranked list ($L_3$) generated using the incorrect personality information. To generate the list $L_3$, we reversed the impact of personality factors on the recommendations' diversity, which means that the personality is integrated into taking negative effects on the diversity adjusting. This variant list was also used by Wu and Chen in their movie recommendation research [22]. We fixed N = 10, K = 80, re-ranking list size LS = 5*N. The results are shown in Table 4.11. Visualization of the results are shown in Figure 4.10.

Table 4.11: Comparison of re-ranked list with incorrect personality information on the Hit rate and ILD

| | Initial List L1 | Re-ranke List L2 (with correct personality information) | Re-ranked List L3 (with incorrect personality information) |
|---|---|---|---|
| Hit rate@10 | 0.055 | 0.289 | 0.255 |
| ILD@10 | 0.390 | 0.482 | 0.483 |

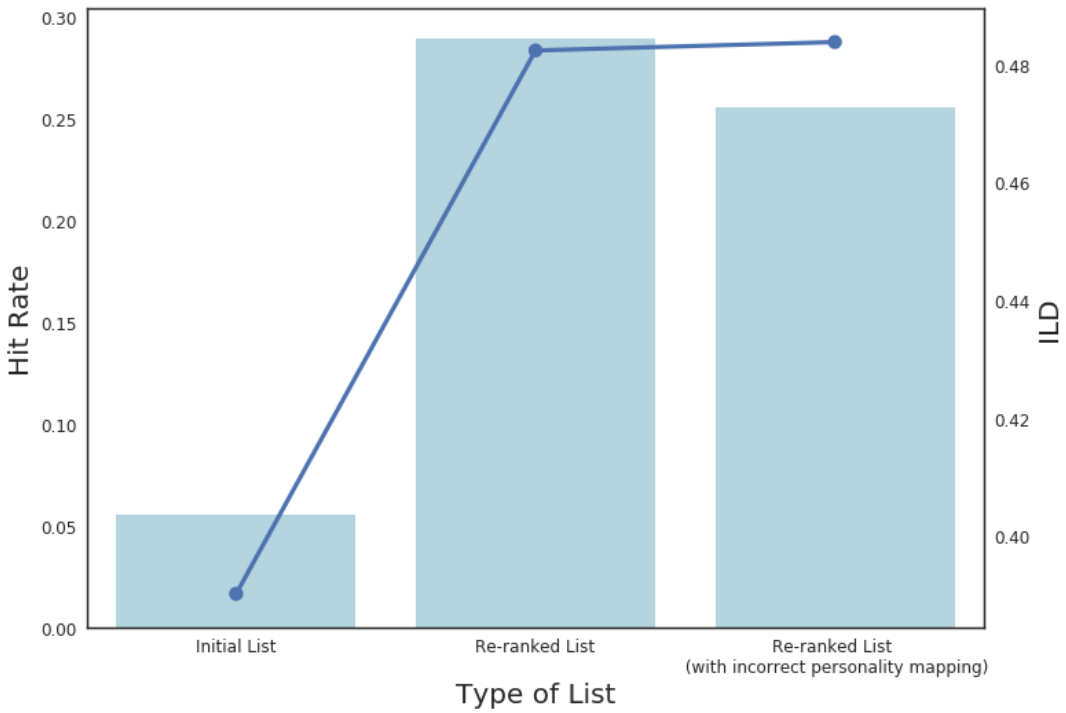



Figure 4.10: Visualization of the comparison of the three lists on the Hit rate and ILD. The bars show the hit rate of the three lists. The dot and the line show the ILD of the three lists. Xlabels from left to right represent L1, L2, L3.

### 4.4.7. Discussion

From the results, generally, in all separated parameter evaluations or the combined lists comparison (in Section 4.4.6), our re-ranked lists all outperform the initial lists both in hit rate and ILD values.

**Influence of the parameter N** When we consider the influence of different parameters separately, we found that, for parameter N, when we keep increasing the size of the Top-N list, hit rates for both lists increase and the differences between the re-ranked list and the initial list become larger. This means that the larger the final recommendation list size, the more likely that user's preferred tracks will be included. For ILD, they also increase along with the increasing of N. When we increase the recommendation list size, more tracks will be included. It is possible that the diversity degree of the reommendation list will thus be increased. Thus, our hypothesis **H1** holds in our evaluation. Considering the computational time, we chose N=10 for the final five list comparison.

**Influence of the parameter K** When we look at the Figure 4.5, clearly, we see that for the re-ranked list, both the hit rate and ILD drop when we increase the K values. This phenomenon is actually in line with the explanation we made in the Testing Methodology (Section 4.4.4) on why we want to decrease the predictions into K unrated items. When we increase the K values, more items are included in the recommendations, which will make the relevant items more hard to get into the Top-N list. Thus, our hypothesis **H2** also holds in our evaluation. Considering the computational time and performance, we chose K=100 for the final five list comparison.

**Influence of the parameter LS** The results of the influence of the size of the re-ranked list are shown in Figure 4.6. Again, similar to the parameter N, we see that both the hit rate and the ILD value are increasing when we keep increasing the size of the list. This result actually implies that the initial recommendations are not that good, leading to the result that we can even find more relevant items when we include more items from the backside of the whole list. While this result can still be acceptable since there is a potential Cold-Start Problem in our combined datasets. The ILD results are rather reasonable since the more items included in the re-ranked list, the more possible that the Top-N list will become more diverse. What should be also noted is that, the computation time increases

dramatically when we increase the LS value. Considering the computational time, we chose LS = 5*N for the final five list comparison. Given such results, our hypothesis **H3** also holds in this evaluation.

**Influence of the personality-related parameter $\lambda$, $\theta_1$, $\theta_3$, and $\theta_3$** Parameter $\lambda$ influences the role of the diversity function played in the objective function 4.1. Results in Table 4.6 show that the larger the $\lambda$, the higher the hit rate for the re-ranked list. The differences between the hit rate for the re-ranked list and the initial list also increase. This result implies that, for this dataset, we can always get better results when we apply the re-ranking methodology. The more we add the diversity into the list, the more possible that the relevant items will be included in the recommendation list. While this hit rate still cannot represent the actual user satisfaction for the final recommendations. Since in our dataset each user has only listened to around 20 songs, there are a large number of unrated songs in the final recommendations. In the offline evaluation, we consider these songs as the irrelevant items. While in real life, it is highly possible that these unrated songs are also preferred by users. This is also the reason why we need to conduct the online evaluation in the next Chapter. As expected, ILD values increased when the $\lambda$ increases (see Figure 4.7).

For the individual parameters $\theta_1$, $\theta_2$, and $\theta_3$, first of all, we noticed that the overall diversity changes little when we adjust the parameter $\theta_i$. This result can be explained by the fact that the overall diversity is mainly controlled by the parameter $\lambda$. Since the $\lambda$ is fixed to 0.5, the overall diversity will have no big changes. While if we look at the single attribute diversity change in Figure 4.8, we see that the $\theta_i$ still has a influence on the single attribute diversity, especially for $\theta_3$ (correlated with attribute *Key*).

**Five lists comparison** For the last evaluation of the five lists, we want to test whether in general the re-ranked list generated based on our proposed personality-based diversification algorithm outperforms all the other lists. The results in Table 4.10 verify our hypothesis **H4**. Although we see that the hit rate is not very high even if we consider the best result of List L5 (hit rate = 0.141), the difference between the hit rate for L5 and L1 is still considerable. We actually can raise the hit rate to a larger number for better comparison if we decrease the K value (e.g. K = 50). But we consider that as the meaningless operation. For one reason, choosing a different number of random items K only had an influence on the absolute hit rate but not on the ranking of the algorithms. We consider that K = 100 is large enough to show the differences in the results in our evaluation. For another, our Pilot Study Dataset is relatively small, a small value of the hit rate is also reasonable.

Back to the results shown in Figure 4.9, clearly, we see that all the four re-ranked lists outperform the initial list both in hit rates and the diversity degrees. Parameter $\lambda$ shows more influence on the final hit rate than the parameter $\theta$, meaning that the overall diversity degrees has more impact on the hit rate than the individual attribute diversity degrees in our objective function. The ILD values for the four re-ranked lists have no big difference, while still they all outperform the initial list.

**Comparison with incorrect personality information** Seeing that almost all kinds of re-ranking methods help enhance the recommendation accuracy and the diversity, we added one more evaluation on the re-ranked list with the incorrect personality information. As shown in Figure 4.10, we see that the re-ranked list with the correct personality information outperforms the list with incorrect personality information in accuracy. The overall diversity degrees for both lists have no big difference. Thus, it shows that our personality-based re-ranking algorithm still helps a little on the recommendation accuracy.

### 4.4.8. Limitation

One of the limitations of our offline evaluation lies in the limited size of the Training Set $M_1$ and Testing Set T. Noted that we only used around 15 tracks for the training for each user and used 5 tracks for the final testing. Given such limited size of the training data, it is possible that the RecSys is not fully trained for users in the Testing Set, making the initial recommendation list not that good. We recommend that later researchers can form a larger user dataset with personality information to repeat our evaluation to get better results.

Another limitation lies in the choice of our complementary Training Set $M_2$, whose track data are quite old. We did not find better user datasets with up-to-date track information, so we stuck to the TPS dataset in our offline evaluation. For later researchers, if they can form a larger user dataset with personality information, they can even skip adding this complementary dataset.

For the correlations we found in our Pilot Study, we did not use all of them. For instance, we also found that Emotional Stability has a positive correlation with attributes artists and tempo. We did not incorporate such correlation in our objective function to explore the potential influence. Chances are that these correlations will yield better results.

### 4.4.9. Conclusion

In this Chapter, we have introduced our Diversity Adjusting Strategy in detail, in which a personality-based diversification algorithm is evaluated. We tested this strategy via a series of offline evaluations. Results show that our proposed algorithm yields better performance (in hit rate and ILD) than the initial recommendations generated by the Recsys. So far, we have partially answered the second research question:

- RQ2: What is the effect (on diversity and accuracy) of adjusting the diversity degrees in Music Recommender Systems based on users' personality information?

We find the way to incorporate the personality information into the diversity adjusting functions and have verified its feasibility via offline evaluations.

While, as mention in Discussion, offline evaluation metrics such as recall and precision cannot always be equal to the actual user satisfaction in real life. For this reason, we conducted an online evaluation utilizing the Spotify Recommender System for our proposed personality-based diversification algorithm, which will be introduced in details in the next Chapter.

# 5

# Online Evaluation

In the previous chapter, we have introduced our Diversity Adjusting Strategy, in which a personality-based diversification algorithm is proposed. The idea of this strategy is to incorporate users' personality information into a re-ranking function to adaptively adjust the diversity degrees of the recommendation list for each user. We have conducted a series of offline evaluations to test the effectiveness and efficiency of the re-ranking algorithm. Results show that the re-ranked recommendation list generated based on our diversification methodology outperforms the initial recommendation list both in hit rate and Intra List Diversity (ILD). However, offline evaluations do not always show the actual user satisfaction. 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 (Chapter 3), we constructed a website [1] for the evaluation. We first acquired user's personality information via the TIPI personality test and then utilized Spotify RecSys [2] to generate the initial recommendations based on user's original interests. We then applied our personality-based diversification algorithm to the initial list to generate the re-ranked list. We displayed both of these two lists (in random order) to users. Users need to rate each track as 'like' or 'dislike' for both lists and give us feedback on the two lists. Our online evaluation is mainly based on these user feedback.

In the following sections, we first introduce the materials we need to obtain from the users. Then, we show how we deal with the independent and dependent variables used in our evaluation. The whole design of the system and user participating procedures then follows. After that, we reveal the results of our online evaluations. Limitations and discussion will also be included.

## 5.1. Materials

Since we want to evaluate the effectiveness of our personality-based diversification algorithm, we need to let the users adjust whether they prefer our re-ranked recommendation list or not. To generate such re-ranked recommendation list, we need to first obtain users' personality information. After that, we also need users' original interests (e.g. listening history or preferred songs) to generate the initial recommendation list. Once we obtain the initial list, we then can apply our diversification method to the list to generated the optimized list based on their personalities. Thus, two materials are needed from the users beforehand: the Personality Profile and the User Interests.

### 5.1.1. Personality Profile

Similar to our Pilot Study, we still adopt the Big-Five Factor Model (FFM) [85] as the basic personality model in our system, in which personality is defined as five factors: Openness to Experience (O), Conscientiousness (C), Extroversion (E), Agreeableness (A), and Emotional Stability (ES).

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

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

To extract the personality factors from the users, we used the Ten Items Personality Inventory (TIPI) [3] considering the limited time users will spend on our evaluation. Ten personality related self-assessment questions are asked to users, users need to rate each question from 1 to 7 (from strongly disagree to strongly agree).

Scores from the TIPI questions should be further mapped to the actual personality scores. The specific way of mapping the TIPI scores to the scores of personality factors can be found in Section Personality Analysis (Section 3.3.2) in Chapter 2. After we map the TIPI scores into the five personality factor scores, we further map these five personality factors into the four different personality levels (Low, Medium Low, Medium High, and High) according to Table 3.2 in Chapter 2. Levels for the five personality factors are used for determining the diversity degrees we set for the re-ranked list for each user individually.

### 5.1.2. User Interests & Recommendation

Besides the personality profile, we also request users to offer their music interests (or music taste). For the computation time consideration, we do not use the Factorization Machine we trained in our offline evaluation. Instead, we turn to the Spotify Recommendation System based on their open Web APIs in order to provide real-time recommendations. In this way, we therefore do not utilize users' listening history as the base for recommendations. Instead, we use the Spotify seed information as the original user interests for later music recommendations. Here, seed is a concept defined by Spotify to represent the user's input information used for generating the corresponding recommendations, such as the preferred artists or preferred tracks. However, one drawback of utilizing such Spotify Recommendation lies in its restriction on the number of the input seeds.

Three kinds of seed information are used in Spotify and so in our system: artists, tracks, and genres. The limitation on the number of the input seeds is 5, which means that users can maximally provide 5 seed values in any combination of artists, tracks and genres. These 5 seed values are used as the original user preference for generating the corresponding recommendations. Initially, 100 different tracks are generated based on these 5 seed values. To ensure that the originally generated recommendation list (100 tracks) is already diverse enough, we use at least 1 artist seed, 1 track seed, and 1 genre seed for every recommendation. That means, in order to generate recommendations, users need to provide at least one of their preferred artists, one preferred track, and one preferred genre, while the total number of the combined seeds still cannot exceed 5.

## 5.2. Independent Variables

After we obtain the two materials from users, we then generate the recommendations for them. In the evaluation, for the purpose of comparison, we generate two recommendation lists 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:

- The first list $L_1$ is constructed by directly taking the top-10 items from the initial list (100 tracks) generated by the RecSys.
- The second list $L_2$ is the re-ranked list generated based on our personality-based diversification algorithm. We use top-50 tracks as the initial input list to generate the final top-10 re-ranked list.

To minimize any carryover effects, these two lists are shown in random order to users, meaning that users do not know which list is the re-ranked list. One example of the display of the two lists can be referred to Figure 5.1. Same tracks may appear in both lists with different ranks.

## 5.3. Dependent Variables

### 5.3.1. Precision v.s. Diversity

In offline evaluation, for each recommendation, we have calculated the hit rate for each relevant item. Our evaluation results show that the re-ranked list has much higher hit rate than the initial list, which means that our diversification method can enhance the recommendation quality for users in the Testing Set. However, as we also mentioned, offline evaluation metrics cannot always reflect the

[3] TIPI: https://gosling.psy.utexas.edu/scales-weve-developed/ten-item-personality-measure-tipi/

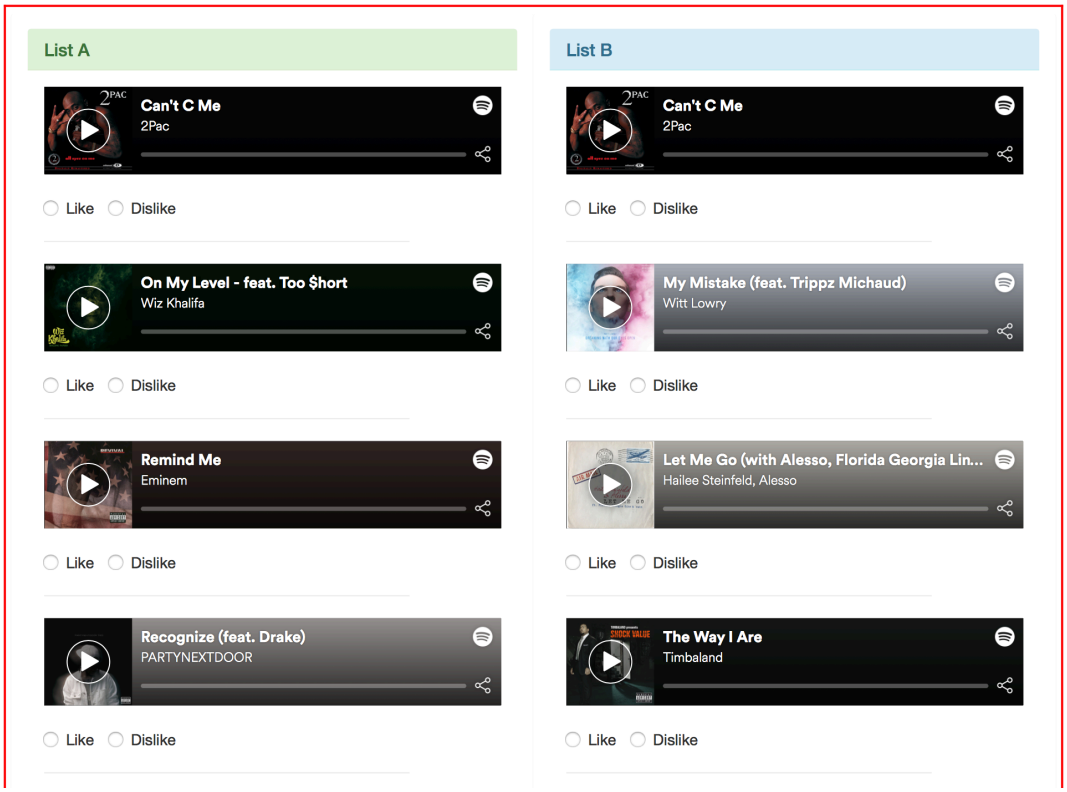



Figure 5.1: Example of the two recommendation lists shown to users. The first four tracks are shown. In total, there are ten tracks in each list.

actual user satisfaction for recommendations in real life. There are plenty of unseen items for users in recommendations. We do not know whether users will like them or not. Our whole offline evaluation is based on the assumption that the user's prior consumption (or ratings) are the ground truth. We simply assume that whatever users have consumed or rated above a certain threshold (considered as relevant items) is something that would be good to recommend. And good RecSys should be able to recommend those relevant items. However, that is not always the whole truth. We do not know whether users truly like the full recommendation list if we only follow the metrics we used in offline evaluation. The simplest way to obtain the actual recommendation quality is to ask users directly.

**Precision** Thus, in order to calculate the precision of the recommendations, we ask users to rate each track as 'like' or 'dislike'. Tracks rated as 'like' are considered as relevant items. Thus, the **Precision@10** for each list is computed as follows:

$$Precision@10 = \frac{\#\ relevant\_items}{|L_u|} \tag{5.1}$$

where the $\#\ relevant\_items$ represents the number of relevant items in a recommendation list. $|L_u|$ represents the size of the whole recommendation list ($L_1$ or $L_2$). In our case, $|L_u| = 10$.

**Diversity** For each list $L_1$ and $L_2$, we also compute the Intra List Diversity (ILD). The computation way of ILD is the same as the ILD we used in the offline evaluation. For quick reference, we agian provide its computation way:

$$ILD = Div_{overall} = \sum_{i=1,2,...,N} \theta_i * Div_i \tag{5.2}$$

where the $Div_i$ means the diversity degree for each track attribute. We used $N = 3$, $\theta_i = 0.33$ here. For each $Div_i$, it is computed as:

$$Div = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n}(1 - Similarity(c_i, c_j))}{n * (n-1)/2} \tag{5.3}$$

where $c_1..c_n$ are items in the list, n is the total number of items in the list. The Similarity function is defined as:

$$Similarity(c_i, c_j) = \begin{cases} 1 & \text{if } c_i = c_j \\ 0 & \text{if } c_i \neq c_j \end{cases} \tag{5.4}$$

### 5.3.2. 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."
- 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 [97] raised by Pearl et al. For each evaluation aspect, the same question is asked for both lists (only changes the list name A and B). Following the ResQue Framework, each question is 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". Furthermore, We have one more question to evaluate users' satisfaction on the order of the tracks in the two lists:

- Order of Tracks (Q9): "Which Recommendation List's order of tracks is better?"

We added this question because we noticed that some of the tracks in the initial list still appear in the re-ranked list but with different ranks (high ranked tracks in the initial list may rank lower in the re-ranked list, and vice verse). We want to know whether the change of the track order may affect users' overall satisfaction. Similar to the last two questions, this question is also rated with categorical answers: "List A", "List B", or "Hard to tell".

## 5.4. Procedure Design

After determining the materials we need from users and the feedback questions we need to ask users, we constructed a website for the whole online evaluation.

Similar to our Pilot Study, the website [4] is constructed using Flask framework [5] and deployed on the Heroku Cloud Application Platform [6]. We used Spotify Web APIs [7] for the extraction of track metadata (such as artist information and audio features) and initial music recommendations. Generally, four main parts are included:

- User's Basic Information: Similar to the Pilot Study, the first part of the evaluation also collects some of users' basic information including their age range, gender, education level, nationality and profession. We also ask about how often they use the the online music service and recommendation service normally.

[4] Available at https://music-rs-personality-online.herokuapp.com
[5] Flask: http://flask.pocoo.org
[6] Heroku: https://www.heroku.com/
[7] Sptotify Web APIs: https://developer.spotify.com/documentation/web-api/

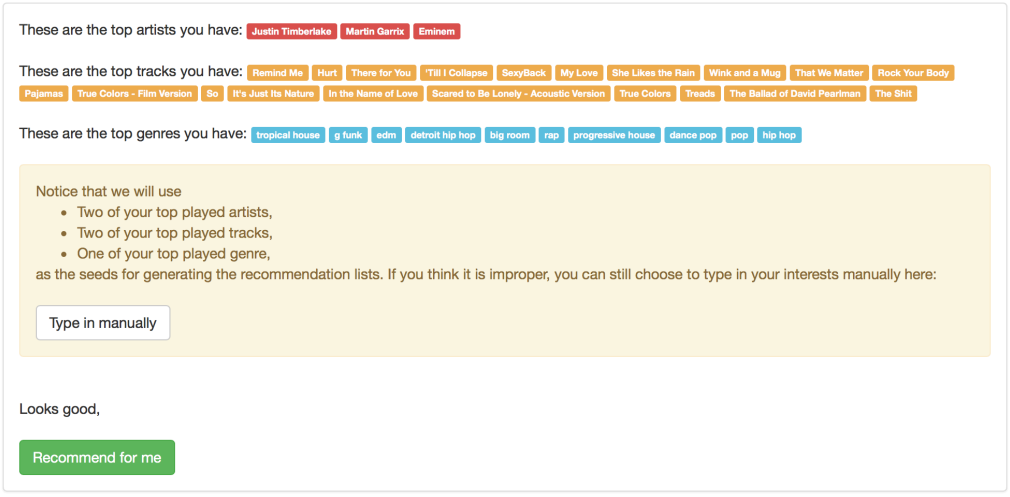



Figure 5.2: Interface for generating recommendation based on Spotify information.

- Personality Test: The second part is the TIPI personality test, 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). The whole list of the ten questions can be found in Chapter 2. Through analyzing the scores of these ten questions, we map them into the scores of five personality factors (Openness to Experience (O), Conscientiousness (C), Extroversion (E), Agreeableness (A), and Emotional Stability (ES)). For each personality factor, the score is further mapped into four different personality levels: Low, Medium Low, Medium High, and High.

- Recommendation: The third part requires users to provide their original music preference. Two channels are provided:

  **Log into their Spotify account** Users can choose to log into their Spotify account to let the system automatically extract their top-played artists, tracks, and genres information. Since the Spotify Recommendation has a restriction on the total number of seeds (5 seeds) used for recommendation, for users choosing to use Spotify, we use 2 top-played artists as the artist seeds, 2 top-played tracks as the track seeds, and the top-played genre as the genre seed for the final recommendation. In case some of the users do not use Spotify frequently or their top-played information is not enough for generating the recommendation, we provide the second channel to obtain their original music preference: type in interests manually.

  **Type in manually** Users can alternatively choose to type in their interests manually. For users choosing to type in interests manually, we request them to type in at least one artist seed, 1 track seed, and 1 genre seed. While the total number of seeds still cannot exceed 5. The interface for both channels can be checked in Figure 5.2 and 5.3.

  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. For each track, we first use the Spotify APIs to extract the corresponding metadata (such as genre information and audio features) associated with it. We then check this user's personality information and mapped all of them into the objective function parameters in our re-ranking algorithm. We applied the diversification algorithm to the whole input list (50 tracks). The result is the re-ranked list with 10 tracks.

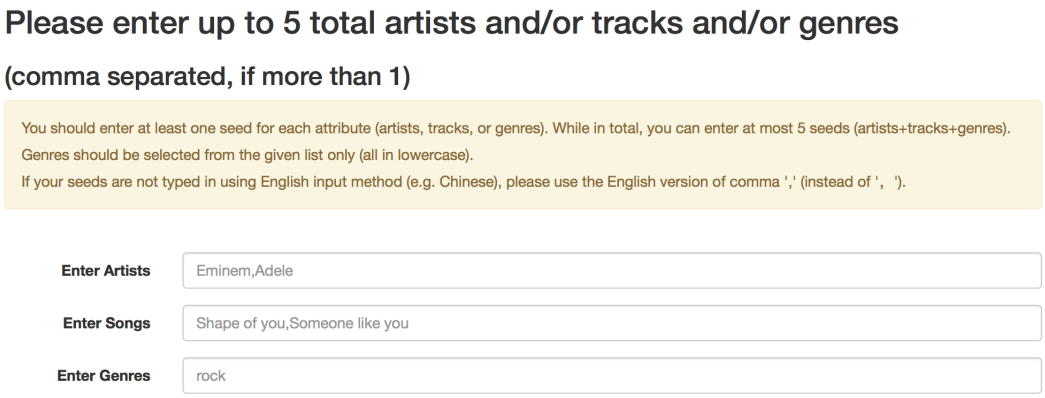



Figure 5.3: Interface for generating recommendation based on manual input.

To minimize any carryover effects, we show these two lists in random order to users (displayed as List A and List B, see Figure 5.1). 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. Questions we discussed in Section 5.3 are asked to users. In order to proceed to the next part, users need to finish all the ratings of the 20 tracks and all the feedback questions.

- User Suggestions: In this part, users can comment on the problems they have encountered during the evaluation or any additional suggestions. This part is in the form of free text, and users can choose to skip this part.

### 5.4.1. Ethical Clearance

As a part of our whole research project, our research (see Section 3.4.3) is approved by the Delft Human Research Ethics Committee (HREC) [8].

The consent form offered to users can be accessed via the first page of our survey website [9]. In the form, we describe the purpose/nature of this research and why participants are recruited. We declare that the participation of survey is voluntary and users can choose not to participate and withdraw at any time. We also describe the whole procedures users need to take during our survey. All users are informed that their information is confidential and anonymous (including IP addresses). The results of this study will only be used for the creation of scholarly publications. All users need to agree with the consent form before they can take the survey.

Similar to our pilot study, we only store the data related to our research purpose, which includes users' basic information, users' personality test results, users' ratings on the recommendation lists and scores for their feedback questions. Users' comments (free text) are also recorded. For users who use the Spotify to generate their recommendations, we do not store their account information.

## 5.5. Hypotheses

Similarly, before the actual evaluation, we first propose our hypotheses beforehand:

- H1: Regarding the Recommendation Quality, the re-ranked recommendation list matches more of users' interests.
- H2: Regarding the Recommendation Diversity, users can perceive that the re-ranked recommendation list is more diverse than the initial recommendation list.
- H3: Regarding User satisfaction, users are more satisfied with the re-ranked recommendation list than the initial recommendation list.

[8]Delft Human Research Ethics Committee (HREC): https://www.tudelft.nl/over-tu-delft/strategie/strategiedocumenten-tu-delft/integriteitsbeleid/human-research-ethics/

[9]Consent form: http://music-rs-personality-online.herokuapp.com

Table 5.1: Demographic profiles of 25 participants for the online evaluation (numbers in the bracket stand for the total number of users for each case).

| | |
|---|---|
| Gender | Male (13); Female (8); Prefer Not to Answer (4) |
| Age | 21-30 (25) |
| Profession | Student (22); Engineer (3) |
| Education | College (4); Graduate School (21) |
| Frequency of using Music Service | ”Regularly (daily/almost daily)” (19);<br>”Moderately (1-3 times a week)” (3);<br>”Infrequently (a few times a month)” (3);<br>”Very infrequently (just a few times overall)” (0);<br>”Never” (0) |
| Frequency of using Music Recommendations | ”Regularly (daily/almost daily)” (6);<br>”Moderately (1-3 times a week)” (8);<br>”Infrequently (a few times a month)” (4);<br>”Very infrequently (just a few times overall)” (7);<br>”Never” (0) |
| Whether they like the recommendations or not | ”Definitely” (2);<br>”Very Probably” (8);<br>”Probably” (10);<br>”Probably not” (3);<br>”Very probably not” (0);<br>”I have never used the recommendations service before.” (2) |

## 5.6. Results

Our evaluation of the two lists are mainly based on the independent and dependent variables we discussed in Section 5.2 and 5.3. In this section, we show the results of our online evaluation.

### 5.6.1. Participants

We conducted our online evaluation mainly with the students in TU Delft. In total, 25 users participated to the evaluation. All of them are aged from 21-30 years old. Table 5.1 shows the general demographic properties of these 25 participants.

### 5.6.2. Precision & Diversity of the two lists

Following the metrics we defined in Section 5.3.1, we measured the precision and ILD for both recommendation lists. Table 5.2 shows the Precision@10 and ILD@10 results. Visualizations of the results can be referred to Figure 5.4. We adopted student t-test for the statistical significance computation. For the Precision@10, statistic = -2.06, $p < 0.05$. For ILD@10, statistic = -4.82, $p < 0.001$.

Table 5.2: Precision@10 and ILD@10 for the two lists. We used Student t-Test for computing the p-values. Student t-test is used. All p-values are smaller than 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) |

### 5.6.3. Recommendation Quality

Regarding the User Feedback questions, we separate all the questions into four aspects. The first aspect measures the Recommendation Quality of the two lists, which is related to Q1, Q2, and Q3 we defined in Section 5.3. Results for Q1 and Q2 are shown in Table 5.3. For Q1 and Q2, we used student


(a) Precision@10 for both lists

(b) ILD@10 for both lists

Figure 5.4: Visualization of the Precision@10 and ILD@10 for the two recommendation lists

t-test for the statistical significance computation (statistic=-3.00, p=0.004).

Table 5.3: Average ratings for accuracy of recommendations for both lists in users' perspective. The value is measured from 1 to 5, meaning from 'Disagree strongly' to 'Agree strongly'. Student t-test is used, p<0.01.

| | Initial List $L_1$ | Re-ranked List $L_2$ |
|---|---|---|
| Average ratings for Accuracy | 3.4 | 4.12 |
| Standard deviation (std) | 0.98 | 0.65 |

Q3 further compares the recommendation quality of the two lists with categorical answers. Results for Q3 are shown in Table 5.4. We adopted Chi-Squared Test for computing the statistical significance for Q3 (statistic=3.92, p=0.14). Noted that users do not know which list is $L_1$ or $L_2$. We used 'List A' and 'List B' in the feedback questions. While the system at the back-end knows the mapping between List A, B and $L_1$, $L_2$.

Table 5.4: Acceptance rate for the Recommendation Quality for the two recommendation lists. Chi-Squared Test is used, p>0.05.

| | Recommendation Quality (match more of the user interests) | | |
|---|---|---|---|
| Statement | "The Initial List $L_1$ is better" | "The Re-ranked List $L_2$ is better" | "Hard to tell" |
| Acceptance Rate | 8.0 % | 52.0 % | 42.0 % |

### 5.6.4. Recommendation Diversity

The second aspect of the feedback questions evaluates whether users can perceive the diversity changes in our re-ranked list. We first evaluate whether users can perceive the diversity in each list via Q4 and Q5 on a 5-point Likert scale. Then we further ask them which one is more diverse (Q6). Users can choose 'List A' or 'List B' or just 'Hard to tell' if they think it is really no big difference. For Q4 and Q5, results are shown in Table 5.5. Results for Q6 are shown in Table 5.6. Student t-test is also used for computing the statistical significance for Q4 and Q5 (statistic=–2.39, p=0.02). For Q6, we adopted Chi-Squared Test (statistic=3.92, p=0.14).

Table 5.5: Average ratings for diversity of recommendations for both lists in users' perspective on a 5-point Likert scale. Student t-test is used, p<0.05.

| | Initial List $L_1$ | Re-ranked List $L_2$ |
|---|---|---|
| Average ratings for Diversity | 3.28 | 3.92 |
| Standard deviation (std) | 0.96 | 0.89 |

Table 5.6: Acceptance rate for the Recommendation Diversity for the two recommendation lists. Chi-Squared Test is used, p<0.05.

| | Recommendation Diversity (which one is more diverse) | | |
|---|---|---|---|
| Statement | "The Initial List $L_1$ is more diverse" | "The Re-ranked List $L_2$ is more diverse" | "Hard to tell" |
| Acceptance Rate | 16.0 % | 48.0 % | 36.0 % |

### 5.6.5. User Satisfaction

The third aspect of the feedback questions evaluates the user satisfaction towards the two recommendation lists. For eahc list, we directly ask users whether they are satisfied with list on a 5-point Likert scale (Q7 & Q8). Results are shown in Table 5.7. Student t-test is used (statistic=-2.03, p<0.05)).

Table 5.7: Average ratings for user satisfaction on both recommendation lists on a 5-point Likert scale. Student t-test is used, p<0.05.

| | Initial List $L_1$ | Re-ranked List $L_2$ |
|---|---|---|
| Average ratings for Satisfaction | 3.36 | 3.92 |
| Standard deviation (std) | 0.93 | 0.97 |

**Full comparison of Recommendation Quality, Diversity and User Satisfaction** The visualization of the full comparison for user feedback (average level of ratings) on Recommendation Quality, Diversity and User Satisfaction is shown in Figure 5.5.

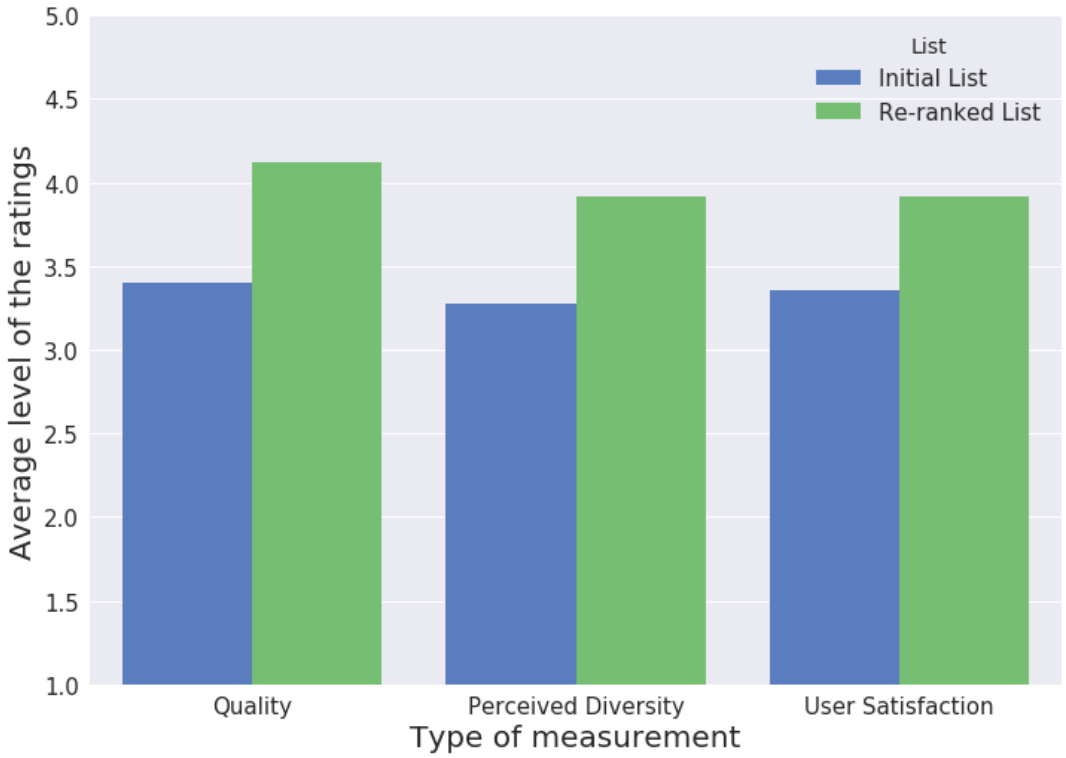



Figure 5.5: Full comparison for Recommendation Quality (Accuracy), Diversity and User Satisfaction

### 5.6.6. Order of Tracks

The last aspect of the feedback questions evaluates the impact of the order of tracks in the two recommendation list (Q9). Our algorithm is based on greedy heuristics. We want to know whether the order of tracks generated in such way would be preferred by users. This question is evaluated with

categorical answers. Results are shown in Table 5.8. Chi-Squared Test is adopted for computing the the statistical significance (statistic=3.44, p=0.18).

Table 5.8: Acceptance rate regarding the order of tracks for the two recommendation lists. Chi-Squared Test is used, p>0.05.

| | Order of tracks | | |
|---|---|---|---|
| Statement | "The Initial List $L_1$'s track order is better" | "The Re-ranked List $L_2$'s track order is better" | "Hard to tell" |
| Acceptance Rate | 16.0 % | 44.0 % | 40.0 % |

## 5.7. Discussion

Overall, looking at Figure 5.4 and 5.5, we can clearly see that our re-ranked recommendation list outperforms the initial recommendation list in all aspects. Although the difference is not so large, considering that we also raise the diversity level of the recommendation at the same time, we 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.

**Recommendation Precision and ILD** When we look at the results in Table 5.2 a little bit further, we see that, for Precision@10, our re-ranked list is a little bit higher than the initial list (mean($L_2$) = 0.67 against mean($L_1$) =0.58 ), which means that on average the re-ranked list recommends more relevant items to users. The difference is not big (around one relevant track difference), while our algorithm has raised the average diversity degrees in general.

**Recommendation Quality** For Recommendation Quality, results in Table 5.3 show that most of users agreed that the recommendations made in the re-ranked list match their interests. And the average ratings for the re-ranked list is much higher than the initial list. The average ratings here do not directly represent the actual ratings for the tracks in the lists, but they represent the levels of acceptance on whether the current recommendation list matches their interests. Thus, we find support for hypothesis **H1** in our evaluation. This hypothesis is further verified by the results shown in Table 5.4. From Table 5.4, we see that, for recommendation quality, around 52% of users agreed that our re-ranked list is better. Only 8% of users showed their preference for the initial list. Since we have seen that the difference for precision@10 for the two lists is very subtle, not surprisingly, still a larger proportion (42%) of users indicated that it is hard to tell which one is better.

**Recommendation Diversity** The main function of our diversification methodology is to add diversity to the recommendation list. From Table 5.2, we have already seen that the average diversity degrees for our re-ranked list are much higher than the initial list. However, that does not mean that users can perceive such difference on diversity degrees, especially in the case of our personality-based diversification algorithm. Since our diversification algorithm is personality-based, for different users with different personalities, diversity degrees are set differently. This means that for some users in extreme cases, the difference on the diversity degrees for the two lists may be very tiny. Thus, we want to know whether on average users can perceive such difference on the change of diversity degrees. From the Table 5.5, we see that users can perceive that the re-rank list is more diverse than the initial list. While, from Table 5.6, although we see that around half of the users consider that the re-ranked list is more diverse, 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. Thus, our hypothesis **H2** is only partially supported here.

**User Satisfaction & Order of Tracks** When we look at the overall user satisfaction towards the two lists, results in Table 5.7 show that users are more satisfied with our re-ranked list, which is also in line with our hypothesis **H3**. Table 5.8 also shows that more users consider that the order of tracks in our re-ranked list in better than the order of tracks in the initial list. While the p-value for Chi-Square Test is also larger than 0.05, which means that this conclusion is not supported.

Thus, concluded from all of the user feedback, we can say that, on average, our re-ranked list not only matches more of users' music interests, but it also has increased the diversity degrees of the

recommendation list. Users may perceive the diversity changes we made to the recommendations. At the same time, they are satisfied with our personality-based diversification algorithm which adds more diversity to their recommendations.

## 5.8. Limitation

One of the limitation of our current online evaluation lies in its limited size of samples. In our current evaluation, only 25 users have participated, and they are all in the same age group (21-30). We cannot tell whether different age groups will result in different feedback on the recommendations.

From results in Table 5.4 and 5.6, we still see that a large number of users cannot clearly tell which list is better. On the one hand, this phenomenon may imply that the parameters we set in our objective function should be further adjusted to enlarge this difference. On the other hand, this situation could also be resulted from the quality of the initial recommendation list generated by the Spotify RecSys. If the quality of the initial recommendation list is not that good, our re-ranked list will also result in relatively bad recommendations. Since Spotify does not make its recommendation algorithm open, we cannot tell whether their recommendation algorithm is good or not. For future work, we suggest to replace the Spotify RecSys with better state-of-the-art RecSys (with a reasonable computation time).

Another limitation is also linked with the Spotify RecSys. In this online evaluation, because of the restriction from Spotify, users can only provide 5 seeds information for recommendations, which are probably not enough to generate an initial recommendation list with a high quality (match most of users' interests). If users can provide more information for the initial recommendation (such as listening history), the re-ranked results could be better.

## 5.9. Conclusion

In this chapter, we have conducted an online evaluation to further verify the effectiveness of our personality-based diversification algorithm with real user feedback. Results show that re-ranked list generated based on our diversification method outperforms the initial recommendation list both in recommendation quality and diversity. Users are also more satisfied with our re-ranked list with a relatively higher diversity. Recalling our second research question:

- RQ2: What is the effect (on diversity and accuracy) of adjusting the diversity degrees in Music Recommender Systems based on users' personality information?

We have proposed a re-ranking method to incorporate users' personality information into the diversity adjusting strategy in music recommendation. Through both of our offline and online evaluation, we show that this personality-based diversification method can enhance the diversity adjusting strategy in music recommendations.

# 6

# Discussion and Future Work

In this chapter, we conclude our research on using personality information to adaptively adjust the diversity degrees for users in music recommendations. We first take a recap of the whole research work. Then we will discuss our findings both in our Pilot Study and Evaluations. Conclusions on each research question then follow. At last, we give our expectations and suggestions on future work.

## 6.1. Research Recap

In this section, we give a summary on our research work. We first start with our research motivation. Then we give a recap on our two research steps: the Pilot Study and the proposal of the Diversity Adjusting Strategy. Procedures and results of the evaluations will also be briefly shown.

### 6.1.1. Purpose

Our research is trying to reduce the research gap between the research on diversity-based recommender systems and personality-based recommender systems by combining these two branches of research together. To accomplish this, we have first conducted a pilot study to explore the relation between users' personality information and their diversity needs on music recommendations. Then, we proposed a personality-based diversification algorithm based on the relation model we defined in the pilot study. Both offline and online evaluations are conducted to verify the efficiency and effectiveness of the proposed diversification method.

### 6.1.2. Pilot Study

Our pilot study is designed to explore the correlation between users' personality factors and their diversity needs on music recommendations. The two materials we studied here are users' personality profiles and their music preference (at least 20 preferred tracks). We adopted the Big-Five Factor Model (FFM) as our personality model and used the Ten Items Personality Inventory (TIPI) personality test to extract the corresponding personality factors. To collect users' music preference, we designed a website [1] in which users need to provide at least 20 songs they normally listened to and can best describe their music taste (via Spotify) besides the personality test (TIPI).

After we collected these information, on the one hand, for each user, we mapped their TIPI question scores into the five personality factor scores defined in FFM. On the other hand, we selected six attributes of the track (*Release Times, Artists, Number of Artists, Genres, Tempo and Key*) to further compute their diversity degrees within the preference list for each user. We adopted the Intra List Diversity (ILD) and the Shannon Entropy as our diversity metrics. Spearman's rank correlation coefficient was used for investigating the correlation between users' personality factors and their diversity needs.

We spread the survey via Crowdflower and in several universities. At last, we filtered 148 participants' data for correlation analysis. Via the pilot study, we found several important correlations between users' personality factors and their corresponding diversity needs. We summarize these correlations in two aspects: the single attribute diversity and the overall diversity. For single attribute

[1] Survey address: https://music-rs-personality.herokuapp.com

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 found in our pilot study are important for the proposal of our later diversity adjusting strategy.

### 6.1.3. Diversity Adjusting Strategy

Based on the results found in our pilot study, we then proposed our personality-based diversification algorithm, which incorporates users' personality factors into a re-ranking function. Specifically, greedy heuristics are used in our work. The 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 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.

We incorporate users' personality factors as the balancing parameters $\lambda$, $\theta_1$, $\theta_2$, and $\theta_3$ into the objective function so that the balance between the similarity function and the diversity function can be controlled by users' specific personality factors.

### 6.1.4. Evaluation

To evaluate the effectiveness of our proposed personality-based diversification algorithm, we conducted both offline and online evaluation.

**Offline Evaluation.** For offline evaluation, we combined our pilot study dataset with a complementary dataset with much larger user data: The Echo Nest Taste Profile Subset (TPS) [2] [93]. 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. For each user in the pilot study dataset, we filtered out the top-5 rated tracks (ratings all $\geq 4$)) to form our testing set. The remaining user data of the pilot study dataset is combined with the TPS subset to form our whole training set.

We adopted a factorization machine as the basic recommender system and applied our diversification algorithm on the recommendation list generated by this system. Since our diversification algorithm is built upon a re-ranking algorithm, 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 size of the input list (LS), the size of the unrated items used for recommendation (K), and the personality related parameters. We tested all their influences both on accuracy (hit rate) and diversity (ILD). Results show that our personality-based re-ranking diversification algorithm helps to enhance both the recommendation accuracy and the diversity degrees compared with the original recommendation list. Our algorithm also outperforms the re-ranked recommendation with incorrect personality information in accuracy.

**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 further conducted an online evaluation. The online evaluation consists of a user study similar to our pilot study, in which we first collect users' personality information and then present them two recommendation list (in random order) at the same time. One of the recommendation list is generated

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

directly by the Spotify Recommender Systems. The other one is generated using our diversification algorithm considering users' personality information. Each list consists of 10 tracks. Users need to show whether they like each track or not. They also need to provide feedback on these two lists via a post-task questionnaire. The feedback questions are mainly regarding the following three main aspects: Recommendation Quality, Recommendation Diversity, and User Satisfaction.

We recruited 25 participants at a university. Participants' ages ranged from 21-30 years old. Results show that our re-ranked recommendation list outperforms the initial recommendation list in all three aspects (recommendation quality, diversity, and user satisfaction).

## 6.2. Discussion

The main contribution of our work is the proposal of the personality-based re-ranking diversification algorithm, which can adjust the recommendation diversity degrees flexibly for people with different personalities. The algorithm is built upon the findings we found in our pilot study, in which we found several important correlations between people's personalities and their diversity needs. Our pilot study is well-designed and the personality information collected from the 148 participants is also in line with the personality distribution from the general population [69]. The key limitation of our pilot study lies in its limited sample size. If more participants are recruited in our pilot study, the correlation between personality factors and diversity needs may be stronger. Another limitation of our pilot study is that we did not include more features (e.g. more audio features like loudness) for correlation research. Chances are that other audio feathers will also have a significant correlation with people's personality factors.

Our diversity adjusting strategy is based on a re-ranking function, which is simple but effective. Chances are that other diversification models (e.g. optimization based diversification) would yield better results compared with our current strategy. We leave that for future exploration. In addition, in our algorithm, the mapping function from the personality factors to the personality related parameters ($\lambda$, $\theta_1$, $\theta_2$, $\theta_3$) is still a little bit fixed (linearly). A better idea is to learn the mapping function from a much larger user dataset with personality information. To the best of our knowledge, currently, there is no such dataset. We encourage later researchers would be able to construct a larger user dataset with corresponding personality information for further research.

In offline evaluation, we found that the re-ranked recommendation list generated based on our diversification algorithm outperforms the initial recommendation list both in accuracy and diversity. However, at the same time, we also found that almost all kinds of re-ranking methods in our experiment help to raise the recommendation accuracy and diversity compared with the initial recommendation list. Although our personality-based re-ranking algorithm still outperforms the other re-ranking methods (with random parameter combinations) a little bit, we say that, in general, re-ranking methods help to enhance the personalization in music recommendation. Later researchers are encouraged to repeat our experiments with different user datasets with personality information to check if there are similar results.

Our online evaluation results also show that our diversification algorithm yields better recommendation quality. Users can also subtly perceive that the recommendation diversity is increased and they are more satisfied with the change. The key limitation of our online user survey still lies in its limited sample size. Again, later researchers are suggested to repeat our research with more participants.

## 6.3. Conclusion

To draw the conclusions, let us first recap our research questions

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

### 6.3.1. Main Research Question

The main research question is stated as follows:

- How does personality information affect how diversity degrees should be applied in Music Recommender Systems?

To answer this question, we first conducted a pilot study to investigate the relation between users' personality information and their diversity needs in music recommendation. Based on the findings, we proposed a personality-based diversification algorithm to flexibly adjust diversity degrees for people with different personalities in music recommender systems. We demonstrated that such strategy helps to enhance the recommendation quality and diversity. Users are also more satisfied with the re-ranked list generated by our algorithm.

### 6.3.2. First Sub-question

Furthermore, we discuss our first sub-question:

- Is there an underlying relationship between people's personality and their needs for recommendation diversity in Music domain?

The simple answer is yes. Through our pilot study, we show that there exist several important correlations between people's personality and their needs for recommendation diversity, especially when we consider the personality factor *Emotional Stability*. For single attribute diversity, we found that they are mostly positively correlated with personality factors *Extraversion, Agreeableness, and Emotional Stability*. For the overall diversity, we found that it is mostly positively correlated with *Emotional Stability*. Specific summary on those correlations can be found in Section 6.1.2 in this Chapter.

### 6.3.3. Second Sub-question

Our second sub-question is stated as follows:

- What is the effect (on diversity and accuracy) of adjusting the diversity degrees in Music Recommender Systems based on users' personality information?

Via our offline and online evaluation on our diversity adjusting strategy, we demonstrated that our personality-based diversification algorithm can generate recommendation lists with better quality and higher diversity degrees. Users also show more satisfaction towards our re-ranked list. Thus, we conclude that, using people's personality information to adjust the diversity degrees in music recommendations can yield better recommendation quality and higher diversity degrees.

## 6.4. Future Work

In this section, we discuss about the future work.

### 6.4.1. Pilot Study with Larger Samples

One of the limitations of our research is the limited sample size (148 participants) in our pilot study. With more user data collected, the correlation between users' personality factors and their diversity needs may be more stronger. We did not find a large user dataset with personality information before conducting our pilot study. Thus, we constructed this pilot study dataset on our own. For further research, later researchers are encouraged to construct a larger user dataset if they want to study the influence of personality information on their diversity needs of recommendations. Such dataset is not only beneficial for the research on diversity problems, but is also beneficial for the general personalized recommendation research.

### 6.4.2. Personality Extraction

**More accurate explicit personality test.** In our research, considering the time limitation for users, we used a very short personality test (TIPI) to extract users' personality information, which contains only 10 self-assessment questions. This personality test is simple, efficient and adopted by many researchers. While the drawback of the this test lies in its accuracy. For further research, we would adopt more accurate explicit personality tests such as the original 44-item Big-Five Inventory (BFI) [61].

**Implicit personality extraction** Besides the explicit personality extraction methods, we are also very interested in the implicit personality extraction methods such as using the social media to infer users' personality information. Such implicit methods have the advantage of not disturbing people with annoying questionnaires. While the measurement accuracy of such methods are still questionable. We would like to research on such implicit methods in future work.

### 6.4.3. Other Diversification Methods

The main diversification method we used in our research is the re-ranking mehtod, which is simple but quite effective. 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. Recently, Wu and Chen [98] also proposed a similar re-ranking method to adjust the individual diversity degrees considering people's personality information in recommendation problems. They compared their method with some other diversification methods like AdaMMR [89] or Clustering [99] and found that such re-ranking method outperforms the other personalized diversity-oriented methods both in accuracy and diversity. Inspired by such works, we also plan to try other diversification strategies and compare the performance with our current strategy.

### 6.4.4. Exploring Diversity Needs on other Track Attributes

In our study, we only explored the diversity degrees on limited number of track attributes such as *Artists* and two audio features *Tempo* and *Key*. While it is still possible that users' personality factors may have stronger correlations with other attributes or audio features like *Albums* and *Loudness*. Thus, in future work, more (audio) features with a larger participant pool will be studied.

### 6.4.5. Emotions

In our pilot study, we found that the personality factor '*Emotional Stability*' has a especially important correlation with users' diversity needs. Considering that the *emotion* is also an important context feature that may have impact on users' diversity needs in music recommendations, we suggest to conduct more research on the impact of emotions in the diversity problem in music recommendations.

# Bibliography


[1] S. M. McNee, J. Riedl, and J. A. Konstan, *Being accurate is not enough: how accuracy metrics have hurt recommender systems,* in *CHI'06 extended abstracts on Human factors in computing systems* (ACM, 2006) pp. 1097–1101.

[2] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan, *Exploring the filter bubble: the effect of using recommender systems on content diversity,* in *Proceedings of the 23rd international conference on World wide web* (ACM, 2014) pp. 677–686.

[3] E. Pariser, *The filter bubble: How the new personalized web is changing what we read and how we think* (Penguin, 2011).

[4] B. Smyth and P. McClave, *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, UK, 2001) pp. 347–361.

[5] N. Hurley and M. Zhang, *Novelty and diversity in top-n recommendation–analysis and evaluation,* ACM Transactions on Internet Technology (TOIT) **10**, 14 (2011).

[6] D. C. Funder, *Personality,* Annual Review of Psychology **52**, 197 (2001), pMID: 11148304, https://doi.org/10.1146/annurev.psych.52.1.197 .

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

[8] A. E. Kemp, *The musical temperament: Psychology and personality of musicians.* (Oxford University Press, 1996).

[9] M. Tkalcic, M. Kunaver, J. Tasic, and A. Košir, *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* (2009) pp. 30–37.

[10] R. Hu and P. Pu, *Enhancing collaborative filtering systems with personality information,* in *Proceedings of the fifth ACM conference on Recommender systems* (ACM, 2011) pp. 197–204.

[11] E. Perik, B. De Ruyter, P. Markopoulos, and B. Eggen, *The sensitivities of user profile information in music recommender systems,* Proceedings of Private, Security, Trust , 137 (2004).

[12] R. Hu and P. Pu, *A study on user perception of personality-based recommender systems,* User Modeling, Adaptation, and Personalization , 291 (2010).

[13] L. Chen, W. Wu, and L. He, *How personality influences users' needs for recommendation diversity?* in *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (ACM, 2013) pp. 829–834.

[14] N. Tintarev, M. Dennis, and J. Masthoff, *Adapting recommendation diversity to openness to experience: A study of human behaviour,* in *International Conference on User Modeling, Adaptation, and Personalization* (Springer, 2013) pp. 190–202.

[15] L. Chen, W. Wu, and L. He, *Personality and recommendation diversity,* in *Emotions and Personality in Personalized Services* (Springer, 2016) pp. 201–225.

[16] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, *Improving recommendation lists through topic diversification,* in *Proceedings of the 14th international conference on World Wide Web* (ACM, 2005) pp. 22–32.

[17] Y.-C. Ho, Y.-T. Chiang, and J. Y.-J. Hsu, *Who likes it more?: Mining worth-recommending items from long tails by modeling relative preference,* in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14 (ACM, New York, NY, USA, 2014) pp. 253–262.

[18] W. Premchaiswadi, P. Poompuang, N. Jongswat, and N. Premchaiswadi, *Enhancing diversity-accuracy technique on user-based top-n recommendation algorithms,* in *Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual* (IEEE, 2013) pp. 403–408.

[19] R. Hu and P. Pu, *Acceptance issues of personality-based recommender systems,* in *Proceedings of the third ACM conference on Recommender systems* (ACM, 2009) pp. 221–224.

[20] R. Hu and P. Pu, *Enhancing collaborative filtering systems with personality information,* in *Proceedings of the fifth ACM conference on Recommender systems* (ACM, 2011) pp. 197–204.

[21] P. Pu, L. Chen, and R. Hu, *Evaluating recommender systems from the user's perspective: survey of the state of the art,* User Modeling and User-Adapted Interaction **22**, 317 (2012).

[22] W. Wu, L. Chen, and L. He, *Using personality to adjust diversity in recommender systems,* in *Proceedings of the 24th ACM Conference on Hypertext and Social Media* (ACM, 2013) pp. 225–229.

[23] A. Koene, E. Perez, C. J. Carter, R. Statache, S. Adolphs, C. O'Malley, T. Rodden, and D. McAuley, *Ethics of personalized information filtering,* in *International Conference on Internet Science* (Springer, 2015) pp. 123–132.

[24] S. Nagulendra and J. Vassileva, *Understanding and controlling the filter bubble through interactive visualization: A user study,* in *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14 (ACM, New York, NY, USA, 2014) pp. 107–115.

[25] A. Caspi, B. W. Roberts, L. Pervin, and O. John, *Personality continuity and change across the life course,* Handbook of personality: Theory and research **2**, 300 (1990).

[26] P. T. Costa and R. R. McCrae, *The revised neo personality inventory (neo-pi-r),* The SAGE handbook of personality theory and assessment **2**, 179 (2008).

[27] C. G. Jung, *Psychological Types: Or, The Psychology of Individuation. Translated by H. Godwyn Baynes* (Harcourt, Brace, 1923).

[28] B. W. Roberts, K. E. Walton, and W. Viechtbauer, *Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies.* Psychological bulletin **132**, 1 (2006).

[29] M. B. Donnellan, *Personality stability and change,* .

[30] P. E. Tetlock, *Expert political judgment: How good is it? How can we know?* (Princeton University Press, 2017).

[31] R. Foster, *News plurality in a digital world* (Reuters Institute for the Study of Journalism Oxford, 2012).

[32] P. Resnick, R. K. Garrett, T. Kriplean, S. A. Munson, and N. J. Stroud, *Bursting your (filter) bubble: strategies for promoting diverse exposure,* in *Proceedings of the 2013 conference on Computer supported cooperative work companion* (ACM, 2013) pp. 95–100.

[33] N. Helberger, K. Karppinen, and L. D'Acunto, *Exposure diversity as a design principle for recommender systems,* Information, Communication & Society **21**, 191 (2018), https://doi.org/10.1080/1369118X.2016.1271900 .

[34] N. Tintarev, *Presenting diversity aware recommendations: Making challenging news acceptable,* (2017).

[35] P. Pu, L. Chen, and R. Hu, *A user-centric evaluation framework for recommender systems,* in *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11 (ACM, New York, NY, USA, 2011) pp. 157–164.

[36] G. Adomavicius and A. Tuzhilin, *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**, 734 (2005).

[37] P. Castells, N. J. Hurley, and S. Vargas, *Novelty and diversity in recommender systems,* in *Recommender Systems Handbook* (Springer, 2015) pp. 881–918.

[38] L. McAlister and E. Pessemier, *Variety seeking behavior: An interdisciplinary review,* Journal of Consumer research **9**, 311 (1982).

[39] B. E. Kahn, *Consumer variety-seeking among goods and services: An integrative review,* Journal of Retailing and Consumer Services **2**, 139 (1995).

[40] I. Cantador, I. Fernández-Tobías, and A. Bellogín, *Relating personality types with user preferences in multiple entertainment domains,* in *CEUR Workshop Proceedings* (Shlomo Berkovsky, 2013).

[41] R. P. Karumur, T. T. Nguyen, and J. A. Konstan, *Exploring the value of personality in predicting rating behaviors: a study of category preferences on movielens,* in *Proceedings of the 10th ACM Conference on Recommender Systems* (ACM, 2016) pp. 139–142.

[42] F. Ricci, L. Rokach, and B. Shapira, *Introduction to recommender systems handbook,* in *Recommender systems handbook* (Springer, 2011) pp. 1–35.

[43] M. Kunaver and T. Požrl, *Diversity in recommender systems–a survey,* Knowledge-Based Systems **123**, 154 (2017).

[44] K. Bradley and B. Smyth, *Improving recommendation diversity,* in *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland* (2001) pp. 85–94.

[45] D. M. Fleder and K. Hosanagar, *Recommender systems and their impact on sales diversity,* in *Proceedings of the 8th ACM conference on Electronic commerce* (ACM, 2007) pp. 192–199.

[46] S. Vargas, *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 (2011).

[47] S. Castagnos, A. Brun, and A. Boyer, *When diversity is needed. but not expected!* IMMM , 44 (2013), cited By 1.

[48] S. Vargas, L. Baltrunas, A. Karatzoglou, and P. Castells, *Coverage, redundancy and size-awareness in genre diversity for recommender systems,* in *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14 (ACM, New York, NY, USA, 2014) pp. 209–216.

[49] N. Lathia, S. Hailes, L. Capra, and X. Amatriain, *Temporal diversity in recommender systems,* in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (ACM, 2010) pp. 210–217.

[50] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang, *Solving the apparent diversity-accuracy dilemma of recommender systems,* Proceedings of the National Academy of Sciences **107**, 4511 (2010).

[51] X. Amatriain and J. M. Pujol, *Data mining methods for recommender systems,* in *Recommender systems handbook* (Springer, 2015) pp. 227–262.

[52] C. Yu, L. V. Lakshmanan, and S. Amer-Yahia, *Recommendation diversification using explanations,* in *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on* (IEEE, 2009) pp. 1299–1302.

[53] M. Zhang and N. Hurley, *Avoiding monotony: Improving the diversity of recommendation lists,* in *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08 (ACM, New York, NY, USA, 2008) pp. 123–130.

[54] S. A. Puthiya Parambath, N. Usunier, and Y. Grandvalet, *A coverage-based approach to recommendation diversity on similarity graph,* in *Proceedings of the 10th ACM Conference on Recommender Systems* (ACM, 2016) pp. 15–22.

[55] M. Jugovac, D. Jannach, and L. Lerche, *Efficient optimization of multiple recommendation quality factors according to individual user tendencies,* Expert Syst. Appl. **81**, 321 (2017).

[56] W. Zeng, M.-S. Shang, Q.-M. Zhang, L. Lü, and T. Zhou, *Can dissimilar users contribute to accuracy and diversity of personalized recommendation?* International Journal of Modern Physics C **21**, 1217 (2010).

[57] F. Mourão, C. Fonseca, C. S. Araujo, and W. Meira Jr, *The oblivion problem: Exploiting forgotten items to improve recommendation diversity.* (2011).

[58] R. Boim, T. Milo, and S. Novgorodov, *Diversification and refinement in collaborative filtering recommender,* in *Proceedings of the 20th ACM international conference on Information and knowledge management* (ACM, 2011) pp. 739–744.

[59] Y. Shi, X. Zhao, J. Wang, M. Larson, and A. Hanjalic, *Adaptive diversification of recommendation results via latent factor portfolio,* in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (ACM, 2012) pp. 175–184.

[60] R. Su, L. Yin, K. Chen, and Y. Yu, *Set-oriented personalized ranking for diversified top-n recommendation,* in *Proceedings of the 7th ACM conference on Recommender systems* (ACM, 2013) pp. 415–418.

[61] O. P. John and S. Srivastava, *The big five trait taxonomy: History, measurement, and theoretical perspectives,* Handbook of personality: Theory and research **2**, 102 (1999).

[62] M. Tkalcic and L. Chen, *Personality and recommender systems,* Recommender Systems Handbook (2015), 10.1007/978-1-4899-7637-6-21.

[63] M. A. S. N. Nunes, *Recommender systems based on personality traits*, Ph.D. thesis, Université Montpellier II-Sciences et Techniques du Languedoc (2008).

[64] R. R. McCrae and O. P. John, *An introduction to the five-factor model and its applications,* Journal of personality **60**, 175 (1992).

[65] O. P. John and S. Srivastava, *The big five trait taxonomy: History, measurement, and theoretical perspectives,* Handbook of personality: Theory and research **2**, 102 (1999).

[66] J. L. Holland, *Making vocational choices: A theory of vocational personalities and work environments* (Psychological Assessment Resources, 1997).

[67] K. W. Thomas, *Conflict and conflict management: Reflections and update,* Journal of organizational behavior **13**, 265 (1992).

[68] J. A. Johnson, *Web-based personality assessment,* in *71st annual meeting of the eastern psychological association, Baltimore, MD* (2000).

[69] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, *A very brief measure of the big-five personality domains,* Journal of Research in personality **37**, 504 (2003).

[70] J. A. Johnson, *Ascertaining the validity of individual protocols from web-based personality inventories,* Journal of research in personality **39**, 103 (2005).

[71] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough, *The international personality item pool and the future of public-domain personality measures,* Journal of Research in personality **40**, 84 (2006).

[72] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, *Our twitter profiles, our selves: Predicting personality with twitter,* in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (IEEE, 2011) pp. 180–185.

[73] G. van Lankveld, P. Spronck, J. Van den Herik, and A. Arntz, *Games as personality profiling tools,* in *Computational Intelligence and Games (CIG), 2011 IEEE Conference on* (IEEE, 2011) pp. 197–202.

[74] W. Wu and L. Chen, *Implicit acquisition of user personality for augmenting movie recommendations,* in *International Conference on User Modeling, Adaptation, and Personalization* (Springer, 2015) pp. 302–314.

[75] J. Golbeck, C. Robles, and K. Turner, *Predicting personality with social media,* in *CHI'11 extended abstracts on human factors in computing systems* (ACM, 2011) pp. 253–262.

[76] R. Hu and P. Pu, *Using personality information in collaborative filtering for new users,* Recommender Systems and the Social Web **17** (2010).

[77] M. Tkalcic, M. Kunaver, A. Košir, and J. Tasic, *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)* (2011) p. 106.

[78] S. Rendle, *Factorization machines,* in *Data Mining (ICDM), 2010 IEEE 10th International Conference on* (IEEE, 2010) pp. 995–1000.

[79] S. Rendle, *Factorization machines with libfm,* ACM Transactions on Intelligent Systems and Technology (TIST) **3**, 57 (2012).

[80] B. Ferwerda, M. Schedl, and M. Tkalcic, *Personality & emotional states: Understanding users' music listening needs.* in *UMAP Workshops* (2015).

[81] A. Langmeyer, A. Guglhör-Rudan, and C. Tarnai, *What do music preferences reveal about personality?* Journal of Individual Differences (2012).

[82] B. Ferwerda, M. Tkalcic, and M. Schedl, *Personality traits and music genre preferences: How music taste varies over age groups,* in *Proceedings of the 1st Workshop on Temporal Reasoning in Recommender Systems (RecTemp) at the 11th ACM Conference on Recommender Systems, Como, August 31, 2017.* (2017).

[83] B. Ferwerda, M. Tkalcic, and M. Schedl, *Personality traits and music genres: What do people prefer to listen to?* in *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (ACM, 2017) pp. 285–288.

[84] B. Ferwerda, M. P. Graus, A. Vall, M. Tkalcic, and M. Schedl, *The influence of users' personality traits on satisfaction and attractiveness of diversified recommendation lists.* in *EMPIRE@ RecSys* (2016) pp. 43–47.

[85] J. A. Johnson, *Descriptions used in ipip-neo narrative report,* Retrieved December (2009).

[86] J. Jia, G. W. Fischer, and J. S. Dyer, *Attribute weighting methods and decision quality in the presence of response error: a simulation study,* Journal of Behavioral Decision Making **11**, 85 (1998).

[87] J. Johnson, *Descriptions used in ipip-neo narrative report,* (2009).

[88] J. Carbonell and J. Goldstein, *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, 1998) pp. 335–336.

[89] T. Di Noia, V. C. Ostuni, J. Rosati, P. Tomeo, and E. Di Sciascio, *An analysis of users' propensity toward diversity in recommendations,* in *Proceedings of the 8th ACM Conference on Recommender systems* (ACM, 2014) pp. 285–288.

[90] S. Vargas and P. Castells, *Exploiting the diversity of user preferences for recommendation,* in *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13 (LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, Paris, France, France, 2013) pp. 129–136.

[91] C. Freudenthaler, L. Schmidt-thieme, and S. Rendle, *Factorization machines factorized polynomial regression models,* (2009).

[92] I. Bayer, *fastfm: A library for factorization machines,* Journal of Machine Learning Research **17**, 1 (2016).

[93] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, *The million song dataset,* in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)* (2011).

[94] Ò. Celma Herrada, *Music recommendation and discovery in the long tail,* (2009).

[95] A. Bellogin, P. Castells, and I. Cantador, *Precision-oriented evaluation of recommender systems: an algorithmic comparison,* in *Proceedings of the fifth ACM conference on Recommender systems* (ACM, 2011) pp. 333–336.

[96] P. Cremonesi, Y. Koren, and R. Turrin, *Performance of recommender algorithms on top-n recommendation tasks,* in *Proceedings of the fourth ACM conference on Recommender systems* (ACM, 2010) pp. 39–46.

[97] P. Pu, L. Chen, and R. Hu, *A user-centric evaluation framework for recommender systems,* in *Proceedings of the fifth ACM conference on Recommender systems* (ACM, 2011) pp. 157–164.

[98] W. Wu, L. Chen, and Y. Zhao, *Personalizing recommendation diversity based on user personality,* User Modeling and User-Adapted Interaction (2018), 10.1007/s11257-018-9205-x.

[99] F. Eskandanian, B. Mobasher, and R. Burke, *A clustering approach for personalizing diversity in collaborative recommender systems,* in *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (ACM, 2017) pp. 280–284.