User Modeling and User-Adapted Interaction (2018) 28:237–276
https://doi.org/10.1007/s11257-018-9205-x

Personalizing recommendation diversity based on user personality

Wen Wu[1] · Li Chen[1] · Yu Zhao[2]

Received: 21 December 2017 / Accepted in revised form: 18 July 2018 / Published online: 26 July 2018

Abstract

In recent years, diversity has attracted increasing attention in the field of recommender systems because of its ability of catching users' various interests by providing a set of dissimilar items. There are few endeavors to personalize the recommendation diversity being tailored to individual users' diversity needs. However, they mainly depend on users' behavior history such as ratings to customize diversity, which has two limitations: (1) They neglect taking into account a user's needs that are inherently caused by some personal factors such as personality; (2) they fail to work well for new users who have little behavior history. In order to address these issues, this paper proposes a generalized, dynamic personality-based greedy re-ranking approach to generating the recommendation list. On one hand, personality is used to estimate each user's diversity preference. On the other hand, personality is leveraged to alleviate the cold-start problem of collaborative filtering recommendations. The experimental results demonstrate that our approach significantly outperforms related methods (including both non-diversity-oriented and diversity-oriented methods) in terms of metrics measuring recommendation accuracy and personalized diversity degree, especially in the cold-start setting.

Keywords Recommender system · Diversity · Personality traits · User survey · Greedy re-ranking

✉ Wen Wu
cswenwu@comp.hkbu.edu.hk

Li Chen
lichen@comp.hkbu.edu.hk

Yu Zhao
jasonchao@gmail.com

[1] Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China

[2] Douban Inc., Beijing, China

1 Introduction

Recommender systems (RS) have become increasingly popular in many web applications for eliminating online information overload and making personalized suggestions to users. In recent years, diversity has been recognized as an important metric for evaluating the effectiveness of online recommendations, because diverse recommendations can stimulate users to explore potential areas of interest (Knijnenburg et al. 2012; McNee et al. 2006; Wu et al. 2012). The research attention in this area has mainly been devoted to increasing the diversity of recommendations and maintaining accuracy at the same time (Bradley and Smyth 2001; Carbonell and Goldstein 1998; Sha et al. 2016; Willemsen et al. 2016; Zhang and Hurley 2008; Ziegler et al. 2005). However, they commonly adopt a fixed strategy to adjust the degree of recommendation diversity for all the users, which ignores individual users' diversity needs. We illustrate the problem in the left hand side of Fig. 1: Providing a diverse recommendation list to a user with broad tastes would be a good idea, but it may be too overwhelming for a user who inherently has narrow interests. Although several researchers in the recommender community have recently attempted to adapt recommendations to users' propensity for diversity (Di et al. 2014; Eskandanian et al. 2017; Shi et al. 2012), most of them rely on users' behavior history like ratings. These approaches thus may not be well applicable to new users who have few behavior records. Moreover, the results may not be tailored to a user's needs that can be caused by some personal factors such as personality.

Indeed, personality refers to the enduring patterns of thought, feeling, motivation and behavior that are expressed in different circumstances (Roberts 2009). A popularly used personality model is the so-called Big-Five Factor model, which defines user personality as five traits (Digman 1990): *Openness to Experience (O)*, *Conscientiousness (C)*, *Extroversion (E)*, *Agreeableness (A)*, and *Neuroticism (N)*. Studies have shown that personality not only affects users' decision making process, but also correlates with their attitudes, tastes, and behavior (Ajzen 2005; John and Srivastava 1999). Drawing on these intrinsic inter-related patterns, personality has been increasingly incorporated into recommender systems in recent years. It is reasonable to believe that personality-based recommender systems can provide more personalized information or services, because they can better understand users from the psychological perspective and better explain why a user prefers one option to the other. For instance, as people with similar personality characteristics are more likely to have similar interests and preferences (Nunes and Hu 2012), personality has been adopted to enhance the nearest neighbor measure in collaborative filtering (CF) based recommender systems (Tkalcic et al. 2009; Tobias et al. 2016; Wu and Chen 2015). Some commercial web sites, such as Whattorent,[1] have also considered the relationship between personality and users' preference for item attributes such as movie genre to recommend items.

On the other hand, the impact of personality on users' diversity preference has also been studied in some recent papers (Chen et al. 2013, 2016; Tintarev et al. 2013; Wu et al. 2013). For example, Chen et al. (2013) explored the correlation between personality traits and users' diversity preference for movies through a user survey

[1] http://whattorent.com.

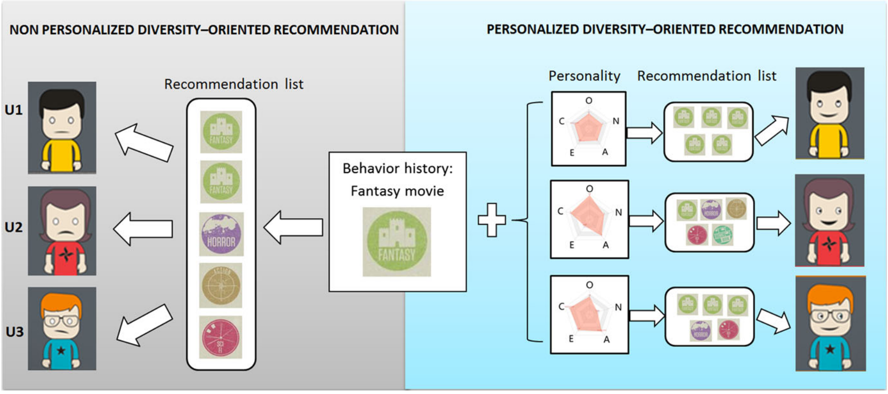


Fig. 1 Sketches of non-personalized and personalized diversity-oriented recommendations (*Note:* The example of the non-personalized diversity-oriented recommendation is shown on the left, where all of the three users receive recommendations (i.e., 2 fantasy movies, 1 horror movie, 1 action movie and 1 sci-fi movie) with the same diversity degree if they have similar behavior history. The example of the *personalized* diversity-oriented recommendation is shown on the right, where the recommendation diversity is adjusted according to both users' behavior history and personality)

(with 181 participants). They found that *Conscientiousness* can significantly affect users' diversity preference for some movie attributes such as release time and country. For example, users who are self-organized (with high C[2]) are more likely to prefer to watch movies with different "release time", whereas those who are disorganized (with low C) tend to prefer diversity in terms of "country". Motivated by the survey's findings (Chen et al. 2013), Wu et al. (2013) proposed a personality-based diversity-adjusting strategy for recommender systems based on a pre-defined set of rules. In this paper, we are interested in extending the previous work from two aspects: (1) Conducting a larger-scale user survey for validating the effect of personality on users' diversity needs in a broader domain that covers different kinds of items (e.g., interest groups that contain users' preferences for various topics such as entertainment, culture, technology, and life); (2) developing a more dynamic, personalized diversity approach based on users' personality, for addressing the cold-start issue.

The main idea of our approach is illustrated on the right hand side of Fig. 1, by which we aim to personalize recommendations' diversity based on both users' behavior history and their personality traits.

Specifically, we surveyed 1706 users on Douban Interest Group,[3] which collected their personality information and interaction behavior. Douban Interest Group is a popular Chinese online community where users can join in different types of interest groups (e.g., "Sports", "Music", "Health", "Academic"), leave comments, and recommend topics to their friends. The multiple linear regression results show that users' big-five personality traits have statistically significant impact ($p < 0.01$ with

[2] High C means that the user has high score on the personality trait *"Conscientiousness"*, which is also applied to the other abbreviations (i.e., O for *"Openness to Experience"*, E for *"Extroversion"*, A for *"Agreeableness"*, and N for *"Neuroticism"*).

[3] https://www.douban.com/group/explore.

Bonferroni-type adjustment) on their diversity preference. For instance, more creative (with high O) and more introverted (with low E) person is more inclined to join different types of groups (see more details in Sect. 3).

Inspired by these observations, we have further developed a generalized, dynamic diversity adjusting approach based on user personality. In particular, personality is incorporated into a greedy re-ranking process, by which we select the item that can best balance accuracy and personalized diversity at each step, and then produce the final recommendation list. The role of personality in our approach is twofold: (1) To estimate individual users' propensity for diversity; (2) to enhance the performance of collaborate filtering (CF) based recommendations in the cold-start setting. To the best of our knowledge, our approach is the first one that takes into account user personality to achieve personalization in recommendation diversity. Through experiments, we demonstrate that the approach can achieve better performance than related methods (including both non-diversity-oriented and diversity-oriented methods) in terms of metrics measuring recommendation accuracy and personalized diversity degree.

In the following sections, we first introduce related work respectively on diversity-oriented and personality-based recommender systems (Sect. 2). We then present details of our user survey, including the procedure and observations (Sect. 3). We further describe our personality-based greedy re-ranking approach in Sect. 4, followed by the experimental setup and results analysis (Sect. 5). Finally, we summarize the experimental findings and discuss our work's limitations in Sect. 6, and conclude the paper in Sect. 7.

2 Related work

2.1 Diversity-oriented recommender systems

To evaluate a recommender system's performance, many studies have focused on "recommendation accuracy", which has principally been measured by the calculated distance between the predicted rating/ranking of items and the target user's true preference (e.g., through metrics RMSE, MAE, nDCG, etc.) (Herlocker et al. 2004; Shani and Gunawardana 2011). To increase accuracy, some recommender systems have targeted to return items that are similar to those that users have previously liked (Adomavicius and Tuzhilin 2005). For example, Peter may receive fantasy movies as recommendations because he gave high ratings for this type of movie before. In recent years, researchers have recognized that accuracy alone is not sufficient to fully reflect a user's potential interests (Kaminskas and Bridge 2016; McNee et al. 2006). Diversity has been considered equally important. Concretely, diversity measures the average or aggregate dissimilarity of items in the recommendation list (Herlocker et al. 2004), for which the similarity is usually determined based on the item's content (e.g., movie genres) (Herlocker et al. 2004). Diversity is a desirable property for a recommender system because varied options can cover different aspects of user interests so as to increase their satisfaction (Hu and Pu 2011b). In addition, the existence of diversity has the capability of reducing users' decision-making effort by supporting them to compare different recommendations (Knijnenburg et al. 2012).

So far, diversity-oriented methods have mainly been aimed at achieving an optimal balance between accuracy and diversity from algorithm development. Re-ranking is one of classic approaches, which is the process of rearranging items that are most relevant to user preference (e.g., through user-based collaborative filtering algorithm) (Adomavicius and Kwon 2009). Re-ranking normally follows a greedy strategy, where the item that maximizes an objective function is selected into the recommendation list at each iteration. Maximal Marginal Relevance (MMR) as introduced by Carbonell and Goldstein (1998) was one of the earliest diversification methods via greedy re-ranking. They aimed to re-rank documents for achieving a trade-off between utility and diversity. Concretely, utility is measured as the relevance of item to the query and diversity is measured as the dissimilarity degree between the item being considered and those already selected. Through a user survey, they found that 80% of participants prefer MMR to the standard relevance-based ranking method. Ziegler et al. (2005) used a heuristic algorithm based on taxonomy similarity to increase the diversity within a recommendation list. They defined a weighting factor to control the contributions from two sets, one containing items that are similar to the user's attribute-based preference profile and the other with items ranked in the reversed order of their similarity to the user's profile. Through an online user study, they found that although this approach slightly sacrifices accuracy, users perceive it more positive w.r.t. diversity and coverage. Willemsen et al. (2016) proposed a diversification method according to the latent features of a matrix factorization model. They maximized the width of the item set (i.e., the distance perpendicular to the user vector) to achieve diversity, and minimized the height of the set (i.e., the distance parallel to the user vector) to maintain the recommendation quality (in terms of predicted rating). A greedy selection algorithm similar to Ziegler et al. (2005) was adopted to generate the recommendations. They tested their latent feature diversification method through two user studies and found that their proposed method increases users' perceived diversity and attractiveness of the item set, while reducing their choice difficulty. Smyth and McClave (2001) compared several optimization strategies, including bounded random selection, greedy selection, and bounded greedy selection. Their experimental results demonstrated that the bounded greedy selection strategy offers the best performance, as it not only ensures optimal balance between recommendation accuracy and diversity, but also reduces calculation complexity by decreasing the number of iteration cycles. Bradley and Smyth (2001) further proposed a diversity-preserving similarity-based retrieval algorithm based on the bounded greedy selection strategy. They reported that this method can deliver significant improvement in recommendation diversity without compromising accuracy. Besides the greedy re-ranking strategy, Zhang and Hurley (2008) relaxed a binary optimization problem with a trust region algorithm and produced the top-N recommendations to maximize the diversity of a retrieved recommendation list as well as its similarity to the target user's query. Experimental results showed that this method can increase the likelihood of recommending diverse items while maintaining accuracy. Sha et al. (2016) proposed a general framework to formalize item recommendation as a combinatorial optimization problem, which integrated item relevance, interest coverage, and recommendation diversity. The experiment indicated that this method outperforms the state-of-the-art techniques in terms of both precision and diversity.

Some of the other studies proposed to directly increase diversity when generating recommendations. For instance, Vargas and Castells (2013) modeled a user's sub-preference profiles for different features and combined sub-profile recommendations through the aspect-based diversification algorithm, which performs a rank-aware allocation of items by taking into account the relative importance of each sub-profile. Their experiment showed this method can achieve better diversity results than a probabilistic Latent Semantic Analysis Recommender (pLSA) (Hofmann 2004) and a List-wise Matrix Factorization Recommender (ListRank) (Shi et al. 2010). Zeng et al. (2010) developed an algorithm to increase recommendation diversity by considering the effects of both similar users and dissimilar users. Two users were regarded similar if they had rated items in common. An item's prediction score for the target user was computed by combining positive scores from similar users and negative scores from dissimilar users in a linear manner. Extensive analyses on real-life movie datasets showed that this approach outperforms the standard CF algorithm regarding both accuracy and diversity. Mourão et al. (2011) tried to increase the diversity in CF-based recommendations by rescuing forgotten items that were preferred by a user in the past, so that these items might be selected by the user at present. The experiment on a Last.fm dataset indicated that this approach can help to increase the diversity degree of returned recommendations.

However, the above approaches did not tailor recommendation diversity to individual users' intrinsic needs. Recently, personalization in diversity has attracted some attention. For instance, Di et al. (2014) proposed an adaptive attribute-based diversification approach which can customize the diversity degree of the top-N recommendation list by taking into account a user's needs for diversity. To be specific, they first modeled each user's inclination to diversity with respect to different item attributes (e.g., genre, director, actor in the movie domain) given her/his rating profile. They then assigned a weight to item attribute for defining the item–item similarity measure (e.g., higher weight will be given to the attribute for which the user shows stronger diversity preference). An experiment in the movie domain showed this approach achieves relatively high accuracy and diversity compared to the standard rating-based method without diversification strategy. Eskandanian et al. (2017) developed a pre-filtering approach to personalizing diversity under the assumption that each user's diversity preference can differ from others'. More specifically, they first automatically segmented users according to their diversity preference level, which was determined as the distribution across categories (e.g., movie genres) of the user's rated items. Then, they ran a standard collaborative recommendation algorithm on each segment separately. They proved that this pre-filtering approach can achieve satisfactory performance in terms of both accuracy and diversity on two datasets MovieLens and Yelp. Shi et al. (2012) proposed a novel recommendation framework by combining matrix factorization (MF) (Koren et al. 2009) with the portfolio theory of information retrieval (Wang and Zhu 2009), with the objective of adapting the recommendation diversification degree to individual users' needs. They concretely modeled the coverage of a user's preference based on the distribution of latent factors, and represented the uncertainty by using the variances of latent factors. The distribution and uncertainty of latent factors in a user profile were then used to determine the final diversity degree of a recommendation list. Experimental results showed this method can effectively adjust the trade-off between relevance and diversity of recommended items.

Unfortunately, these approaches primarily rely on users' history behavior records such as ratings, which limits their applicability to new users who have few behavior records. In addition, they did not consider the effect of some personal factors such as personality on users' diversity needs. In our previous work, we found personality can significantly affect a user's diversity preference for movies (Chen et al. 2013). Therefore, in this work, we are motivated to develop an approach for personalizing recommendation diversity based on user's personality, to address the cold-start issue in particular.

2.2 Personality-based recommender systems

As mentioned before, personality can affect users' attitudes, tastes and behavior (Ajzen 2005). In the field of recommender systems, Karumur et al. (2017) found that personality traits significantly correlate with newcomer retention, activity degree, and rating pattern. For example, introverted (with low E) or aggressive (with low A) users are more likely to return to use the system compared to those with high E or high A. In addition, disorganized users (with low C) behave more actively in a recommender system like MovieLens than self-organized users (with high C). As for the rating pattern, people with high O tend to provide more positive ratings.

In the past few years, some researchers have attempted to incorporate personality into the process of generating recommendations (Tkalcic et al. 2016). For instance, Tkalcic et al. (2009) used personality to improve the nearest neighbor measure in CF systems, and identified that the personality-based similarity measure is more accurate than the traditional rating-based measure in the cold-start situation. Hu and Pu (2011a) developed a cascade style hybrid CF, which adopts the pure personality-based algorithm to make initial predictions for unobserved items and then applies the classic CF method on the user-item matrix. Their experiment identified that the hybrid CF significantly outperforms the non-personality-based approach in sparse datasets. More recently, Fernandez-Tobias et al. (2016) developed three approaches to mitigating the new user problem respectively based on (a) personality-based matrix factorization (MF), which improves the recommendation prediction model by directly incorporating user personality into MF; (b) personality-based active learning, which regards personality as the additional useful preference for improving the output of recommendation process; (c) personality-based cross-domain recommendation, which exploits personality to enrich user profile as obtained from auxiliary domains with the aim of compensating for the lack of user preference data in the target domain. They found that all the three proposed personality-based methods achieve performance improvements in real-life datasets, among which the personality-based cross-domain recommendation performs the best.

Personality has also been incorporated into preference-based recommender systems. Hu and Pu (2010) established a personality-based interest profile for each user, which reflects the relationship between personality and the user's preference for music genres (Rentfrow and Gosling 2003). Items that best match the user's profile are then recommended. For example, energetic and rhythmic music will be recommended to extrovert people. They then conducted a user study that demonstrated that users in this system not only are willing to find songs for themselves, but also enjoy recommending songs to their friends. Some commercial web sites also used personality to produce preference-based recommendations. For instance, Whattorent (http://www.whattorent.com/) adopts the LaBarrie theory,[4] which states that a movie viewer interacts with a movie emotionally in the same manner that s/he interacts with human beings. The site first establishes a general model that describes how users react to the world and their average emotional state. A database also stores the correlations between personality and movies' attributes like genre. Movies are then recommended given the user's personality and the current mood. Hu and Pu (2009) performed a user study to compare Whattorent with MovieLens (a classical rating-based recommender system), and found that Whattorent is easier to use and leads to higher user loyalty.

Recently, there are few endeavors to identify the effect of personality on users' propensity for diversity within multiple recommendations. Tintarev et al. (2013) adopted a user-as-wizard method to explore how people diversify a list of books when they recommend them to their friends. In order to verify the assumption that people with higher *Openness to Experience* would be more willing to receive diverse recommendations, they conducted a user survey (with 120 users). The results showed, although *Openness to Experience* does not affect items' overall diversity, participants, who are more creative, prefer higher diversity in respect of book genre. As mentioned before, in our previous work (Chen et al. 2013), we did a user survey to identify whether people, with different personality traits, would have different needs for recommendation diversity. By analyzing 181 users' responses to our survey, we found that the five personality traits all, to a certain extent, influence users' diversity preference for movies. For instance, diversity preference w.r.t. "actor/actress" is positively correlated with *Openness to Experience* (imaginative and creative users), and that w.r.t. "country" is significantly negatively correlated with personality traits *Conscientiousness* and *Agreeableness*. We also found that the overall diversity across all attributes is significantly negatively correlated with *Conscientiousness*, which suggests that disorganized users are more likely to choose diverse movies. In a follow-up work (Wu et al. 2013), we developed a simple personality-based diversity-adjusting strategy rooted in the preference-based recommending process: (1) Building a user's preference profile that includes her/his personality and criteria for the item's attributes, (2) recommending items that match the user's profile. To be more specific, we first used the user's personality profile to locate n diverse items. We then retrieved m most relevant items according to her/his attribute preferences. The total number of recommendations is hence N, $N = m + n$.

However, there are several limitations of the above-described work. Firstly, the scale of each user survey was a bit small (at most 181 users in Chen et al. 2013). Indeed, larger samples might be useful to more accurately represent the characteristics of the population from which they are derived (Cronbach 1972). Secondly, little work has been done to identify the role of personality in personalizing recommendation diversity for new users. Thirdly, the existing personality-based diversity approach is

[4] www.whattorent.com/theory.php.

simply based on a set of pre-defined rules (Wu et al. 2013), which is not dynamic and generalized.

In the current work, we are interested in extending the previous work from two aspects. One is conducting a larger-scale user survey to validate the relationship between users' personality traits and their preference for diversity in a broader domain that covers different types of items (such as Douban's Interest Group where users can join groups with various topics including entertainment, culture, technology, life, and so on). The second is developing a more dynamic, generalized personality-based recommendation diversity approach and identifying its performance in the cold-start setting.

3 Preliminary study: user survey

3.1 Materials and participants

The user survey was performed on Douban Interest Group (https://www.douban.com/group/explore). Douban is a very popular social media site in China that has attracted millions of users to use since its launch in 2005. Being part of Douban, Douban Interest Group is an online community that allows users to form different groups to discuss various aspects of life such as entertainment, culture, technology, and life. Concretely, it consists of 19 types of groups, including "Sports", "Animation", "Music", "Health", "Life Style", "Academic" and so on.[5] Every member can create a topic within her/his joined group, and then any other members can comment, recommend the group/topic to their followers, or "Like" (see the snapshots in Fig. 2).

From July to August 2016, we sent survey invitation to 128,940 active users[6] in Douban via "Dou You" (a communication tool similar to private message), of whom 2076 users accepted. Before starting the survey, every participant was asked to read an instruction for the upcoming tasks and sign an informed consent form. S/he was then required to answer some questions about her/his personal background such as age, gender, education level and personality. After filtering out incomplete and invalid answers,[7] we remained 1706 users' responses (with 1291 females). Most of these users are Chinese (99.2%), with different educational backgrounds (66% with Bachelor degree, 18% with Master degree, 3% with PhD, and 13% miscellaneous) and age ranges (2% in the range of 10–19 years, 81% in the range of 20–29 years, 16% in the range of 30–39 years, and 1% aged over 40). As the incentive, we opened a lottery draw with 42 awards (a total cost of RMB 4200).

[5] The type classification is determined by Douban, and each group belongs to one type.

[6] *Active user* refers to the user who has at least one behavior record during the year 2015 (from Jan. 1 to Dec. 31), e.g., creating a group/topic, joining a group, leaving a comment, recommending or liking a group/topic.

[7] To clean the data, we first excluded 118 incomplete answers (5.7%). We then analyzed users' answers to the personality questionnaire and filtered out all of the contradictory records [e.g., a user rated 5 (out of 5) on two opposite statements "I think I am cold and aloof" and "I think I am considerate and kind to almost everyone"], by which we further removed 252 invalid answers.

3.2 Procedure and measurement

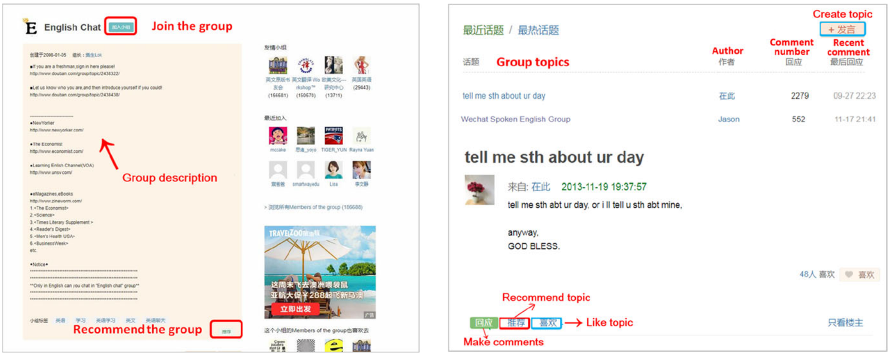
One widely used personality model is called Big-Five Factor model that defines five personality traits (Digman 1990). *Openness to Experience* can be used to judge whether a person is creative/open-minded (with high O) or reflective/conventional (with low O). *Conscientiousness* inherently leads a person to become self-disciplined/prudent (with high C) or careless/impulsive (with low C). *Extroversion* distinguishes people who are sociable/talkative (with high E) from those who are reserved/shy (with low E). *Agreeableness* reflects individual differences in concern with cooperation and social harmony. People with high A tend to be trusting and cooperative, while people with low A are likely to be aggressive and cold. The last factor is *Neuroticism*, which reflects an individual's tendency to experience psychological distress: People with high N are more emotionally unstable than those with low N.

In our work, each user's personality was assessed via the Big-Five Inventory (BFI) (John and Srivastava 1999), because it not only reaches adequate levels in terms of convergent validity and discriminant validity measurements, but also complies with the trend toward using shorter instrument for saving users' time and effort (John and Srivastava 1999). Concretely, the BFI is a 44-item instrument that yields a score for each of the big-five personality traits: *Openness to Experience* (10 items), *Conscientiousness* (9 items), *Extroversion* (8 items), *Agreeableness* (9 items), and *Neuroticism* (8 items).

Each item is a short statement, for participants to rate on a 5-point Likert scale (from 1 "strongly disagree" to 5 "strongly agree"). For example, one statement of assessing *Extroversion* is "I see myself as someone who is talkative." In order to place the five traits' scores in the same range, we rescaled the original metric of each BFI item (i.e., 1–5 scales) to a popularly-used metric known as percentage of maximum possible [POMP (Cohen et al. 1999), range 0–100],[8] and calculated the score for each trait by taking the average of the POMP scores on the items belonging to that trait (Helson and Soto 2005; Kaiseler et al. 2012; Srivastava et al. 2003; Wood and Wortman 2012).

The reliability test of our BFI instrument shows that the internal consistency coefficient (Cronbach's alpha) of the five traits *Openness to Experience (O)*, *Conscientiousness (C)*, *Extroversion (E)*, *Agreeableness (A)*, and *Neuroticism (N)* are 0.931, 0.897, 0.887, 0.898, and 0.853 respectively. These values are all above 0.70, suggesting that the assessing statements have satisfactory internal validity (Nunnally 1967). In addition, Fig. 3a presents the mean POMP value of each personality trait: O (mean $= 65.063$ out of 100, SD $= 14.588$), C (mean $= 54.719$, SD $= 13.142$), E (mean $=$ 49.280, SD $= 16.682$), A (mean $= 62.914$, SD $= 12.805$), and N (mean $= 52.158$, SD $= 13.902$). We compared our users' personality distribution with the existing BFI norms [i.e., Srivastava et al.'s work (2003) that covers 132,515 American samples]. Although the absolute value of each personality trait is not the same [probably due to culture differences (McCrae and Terracciano 2005)], we observe that our Chinese samples score the highest on O, followed by A and C (refer to Fig. 3a), which is similar to American samples (Srivastava et al. 2003). Additionally, Fig. 3b illustrates the

[8] A POMP score is a linear transformation of any raw metric into a 0–100 scale, where 0 represents the minimum possible score and 100 represents the maximum possible score (Cohen et al. 1999). In this case, $Score_{POMP} = (Score_{ORI} - 1) \times 25$, where $Score_{ORI}$ ranges from 1 to 5.



English Chat 加入小组

创建于2008-01-05 组长：晨生Lok

●If you are a freshman,sign in here please!
http://www.douban.com/group/topic/2438322/

●Let us know who you are,and then introduce yourself if you could!
http://www.douban.com/group/topic/2438438/

●NewYorker
http://www.newyorker.com/

●The Economist
http://www.economist.com/

●Learning Enlish Channel(VOA)
http://www.unsv.com/

●eMagazines,eBooks
http://www.zineworm.com/
1.<The Economist>
2.<Science>
3.<Times Literary Supplement >
4.<Reader's Digest>
5.<Men's Health USA>
6.<BusinessWeek>
etc.

●Notice●

Only in English can you chat in "English chat" group

小组标签 英语 学习 英语学习 英文 英语聊天

推荐

友情小组
英文原版书友会 (166681)
英文翻译 Workshop™ (150570)
欧美文化—研究中心 (13711)
英国英语 (29443)

最近加入
mccake
思迪_yoyo
TIGER_YUN
Rayna Yuan
雷爸爸
smartwayedu
Lisa
李文静

> 浏览所有Members of the group (186688)

这个小组的Members of the group也喜欢去

最近话题 / 最热话题

+ 发言

话题	作者	回应	最后回应
tell me sth about ur day	在此	2279	09-27 22:23
Wechat Spoken English Group	Jason	552	11-17 21:41

tell me sth about ur day

来自: 在此 2013-11-19 19:37:57

tell me sth abt ur day, or i ll tell u sth abt mine,

anyway,
GOD BLESS.

48人 喜欢 喜欢

回应 推荐 喜欢

只看楼主

Fig. 2 Snapshots of the Douban Interest Group community (https://www.douban.com/group/explore) (*Note*: The left hand side shows the snapshot of users' group-level behavior in the community. For instance, users can join the group and recommend it to their followers. The right hand side shows the snapshot of topic-level behavior. In the joined group, users can create topics, and like, recommend, or leave comments on any topics in their interest)

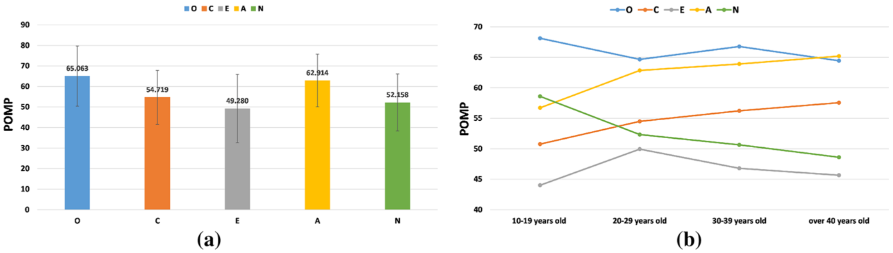


Fig. 3 Big-Five personality scores and distributions of the participants who joined in our user survey. **a** BFI scores of participants who joined in our user survey (POMP). **b** BFI scores of our participants broken down by age (POMP)

mean scores of our participants' five personality traits broken down by age, with trends consistent with those in Srivastava et al. (2003). For example, the values of *Conscientiousness* and *Agreeableness* increase throughout early and middle adulthood, while *Neuroticism* shows a downward trend with increasing age. The comparative results hence indicate that our samples are representative.

In addition to users' answers to our survey questions, we also have their behavioral logs (e.g., *like* the group topic, *recommend* the group or topic to her/his followers, and *leave comments*) that have been automatically recorded. Through statistical analysis, we found those 1706 participants joined a total of 2396 Douban groups during the year 2015 (from Jan. 1 to Dec. 31) and each user joined on average 20.11 groups (standard deviation=16.98, see Fig. 4). We used the proportion of each type c that appears in groups that a user u has joined, to represent her/his preference for that type:

$$Pref_c(u) = \frac{N_{u,c}}{N_u} \tag{1}$$

where N_u refers to the total number of groups joined by user u, and $N_{u,c}$ refers to the number of u's joined groups that belong to type c.

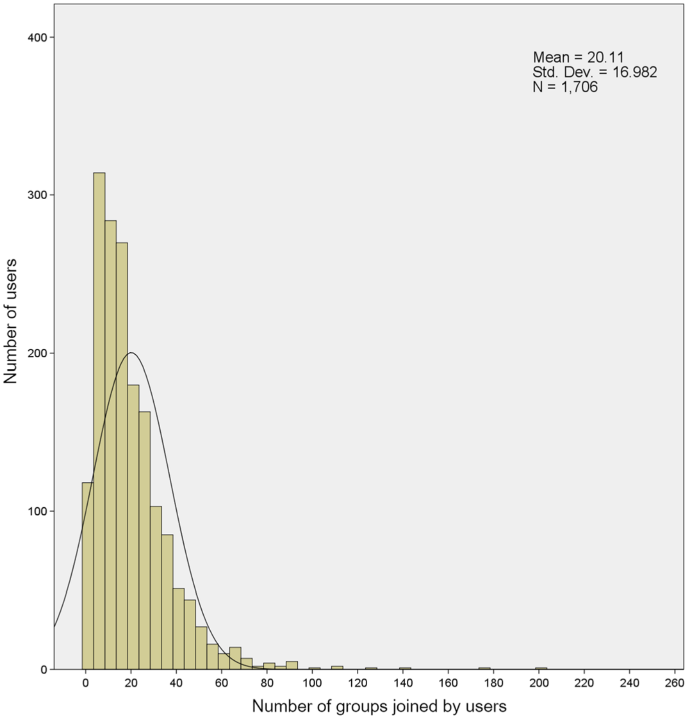


Fig. 4 Distribution of users' group joining records in Douban

We further calculated each user's diversity preference, for which the diversity is defined as the distribution over all types of the user's joined groups.[9] Specifically, we adopted *Shannon Entropy* (Rényi et al. 1961) to measure the diversity preference:

$$Div(u) = -\sum_{c \in C} P(c|u)\log_2(P(c|u)) \tag{2}$$

where $P(c|u)$ refers to the relative frequency of group type c among the groups user u has joined, which can be measured via Eq. 1, and C refers to the set of group types. A higher entropy value suggests that the user prefers to join in groups with different types (i.e., with high diversity preference).

3.3 The impact of personality on users' preference for diversity and group types

We ran multiple linear regression (Seber and Lee 2012) for analyzing the impact of personality on users' preference for diversity and group types, for which users' five personality traits are predictors and their diversity preference as well as group type

[9] Similar to other work that often uses genres of movies, cuisine types in restaurants, or topic categories of news stories, to calculate items' diversity (Eskandanian et al. 2017), we mainly considered the types of groups users have joined.

Table 1 Multiple linear regression of five personality traits on users' diversity preference and group type preference (*$p < 0.01$ with Bonferroni-type adjustment)

	Openness to Experience (O)	*Conscientious -ness (C)*	*Extroversion (E)*	*Agreeableness (A)*	*Neuroticism (N)*
Standardized coefficients (Beta)					
Diversity preference	**0.109***	−0.014	− **0.083***	0.014	0.038
Group type preference					
Sports	0.004	**0.078***	**0.209***	0.010	− **0.156***
Animation	**0.080***	− 0.041	− 0.041	0.013	− 0.031
Religion	0.037	− 0.022	− 0.032	0.038	0.012
Music	**0.090***	0.001	− 0.014	− 0.040	0.031
Health	− 0.020	**0.111***	**0.103***	− 0.009	0.036
Literature	**0.095***	0.041	0.022	− 0.038	− 0.048
Costume	0.025	− 0.005	0.001	0.020	**0.200***
Food	0.003	0.009	0.035	0.015	− 0.009
Love	− 0.005	0.022	0.001	− 0.005	− 0.006
Movie	**0.067***	0.039	− 0.029	−**0.239***	0.045
Fashion	0.024	− 0.028	**0.081***	0.024	0.062
Social	0.031	− 0.001	0.039	0.006	0.008
Lifestyle	0.022	0.021	**0.072***	0.010	0.054
Shopping	0.020	− 0.035	**0.171***	0.025	0.031
Art	**0.127***	0.019	0.039	0.002	0.046
Emotion	0.019	− 0.010	− 0.007	− 0.007	− 0.022
Chatting	0.027	0.007	0.007	− 0.031	0.022
Academic	**0.084***	0.007	− 0.025	− 0.006	0.051
Interest	**0.106***	0.027	− 0.013	− 0.021	0.052

preferences are dependent variables. This method enables us to see the relative effect of each personality trait. However, performing multiple testing may result in the inflation of Type I error (i.e., accepting "spurious" significance results as "real") (Perrett et al. 2006). To solve this issue, we used a Bonferroni-type adjustment (Armstrong 2014), which is one of the commonly used methods for adjusting the significant levels of individual tests when multiple tests are performed on the same data. To be specific, the adjusted level of significance, in general α/k for k tests, is used to conduct each of the k individual tests (Perrett et al. 2006). Table 1 shows the results of multiple linear regression analyses, where $p < 0.01$ (=0.05/5) indicates that changes in one predictor can be significantly associated with changes in the dependent variable.

Concretely, users' diversity preference is significantly positively influenced by *Openness to Experience (O)*, and negatively by *Extroversion (E)*, implying that users who are more creative and introverted are inclined to join different types of groups. Moreover, as for the effect of personality on users' group type preference, our results show that all of the five personality traits significantly affect users' preference for some particular types. For instance, groups about "Health" and "Sports" are more preferred by people who are more self-organized (with high C) and extroverted (with high E). Other groups related to aesthetics (e.g., "Art" and "Literature") and entertainment (e.g., "Animation", "Music" and "Movie") are more preferred by people who are more creative and aesthetic sensitive (with high O). These people also prefer the groups about "Academic" and "Interest". For the groups related to the daily life like "Fashion", "Lifestyle", and "Shopping", they are more preferred by more extroverted people (with high E). In addition, people who are more suspicious (with low A) tend to prefer "Movie" type groups, whereas those who are more emotionally unstable (with high N) are likely to prefer "Costume" type groups.

4 Methodology: personality-based greedy re-ranking for personalized recommendation diversity

Inspired by our user survey's observations that personality can significantly influence users' diversity preference as well as their group type preference, we have further developed a personality-based greedy re-ranking approach to improving recommendation diversity (shorten as PB Greedy). In our approach, personality takes two major roles. On one hand, it is used to predict a user's diversity preference, which enables the system to make personalized recommendations tailored to the user's inherent needs for diversity. On the other hand, personality is leveraged to enhance the measure of user-user similarity in collaborative filtering (CF) so as to improve recommendation in the cold-start setting. Given that different personality trait may have different impacts on users' behavior (as motivated by the results shown in Table 1), we automatically compute weights (relative importance) of the five personality traits.

The workflow of our approach can be seen in Fig. 5, which is mainly composed of two steps: (1) To predict a user's preference for un-experienced items, and (2) re-rank the items to meet the user's diversity preference. We concretely adopt the greedy re-ranking technique (Adomavicius and Kwon 2009), because it can not only be easily incorporated into the existing recommender algorithms but also explicitly control the level of diversification. To be specific, S denotes a candidate item set of size n for a user u, which is generated according to her/his predicted preference for items. T denotes the re-ranked list that user u will finally receive, which includes N items ($N < n$, because the recommendation list T is reproduced from the larger set of candidate items). At each iteration, we add an item that maximizes the objective function $Score_{final}$ with the aim of optimizing the trade-off between the user's preference for the item and her/his diversity preference for all items selected so far:

$$Score_{final}(u, i) = \beta * Score_{Pref}(u, i) + (1 - \beta) * Score_{PersonalizedDiv}(u, i) \tag{3}$$

where $Score_{Pref}(u, i)$ denotes the user u's preference for item i (see details under Step 1 below), $Score_{PersonalizedDiv}(u, i)$ represents the personalized diversity degree (see details under Step 2), and the parameter β is used to balance the two types of preference.

Step 1: Predicting user preference for items

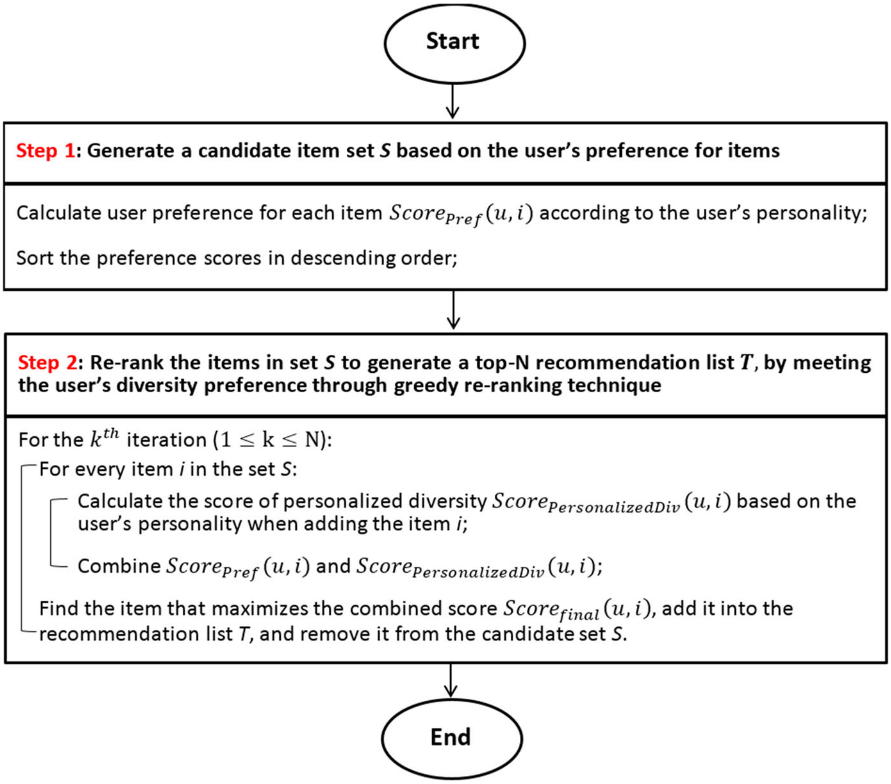

Start
Step 1: Generate a candidate item set S based on the user's preference for items
Calculate user preference for each item $Score_{Pref}(u,i)$ according to the user's personality;
Sort the preference scores in descending order;
Step 2: Re-rank the items in set S to generate a top-N recommendation list T, by meeting the user's diversity preference through greedy re-ranking technique
For the k^{th} iteration ($1 \leq \mathrm{k} \leq \mathrm{N}$):
For every item i in the set S:
Calculate the score of personalized diversity $Score_{PersonalizedDiv}(u,i)$ based on the user's personality when adding the item i;
Combine $Score_{Pref}(u,i)$ and $Score_{PersonalizedDiv}(u,i)$;
Find the item that maximizes the combined score $Score_{final}(u,i)$, add it into the recommendation list T, and remove it from the candidate set S.
End

Fig. 5 Personality-based greedy re-ranking approach for achieving personalization in recommendation diversity

User-based CF has been widely used to produce recommendation, which mainly depends on users' behavior data (such as ratings) to find k most similar neighbors by computing user-user similarity and then receive the top-N most relevant items (Su and Khoshgoftaar 2009). However, in practice, this method will encounter the cold-start problem when a new user uses the system because s/he has few behavior records. As people with similar personalty characteristics tend to have similar interests and preference (Nunes and Hu 2012), in our approach we attempt to incorporate users' personality values to alleviate this issue, by which a new user u's preference for an item i can be predicted based on the rating profiles of her/his neighbors who have similar personality values.

Specifically, we first adopt the 5-dimension vector $ps_u = (ps_u^1, ps_u^2, ps_u^3, ps_u^4, ps_u^5)^T$ to define a user u's big-five personality traits. We then compute the personality-based similarity between two users u and v via *Euclidean distance* measure[10]:

[10] The reason we did not use Cosine similarity measure (Qian et al. 2004) is because it may produce the deviation when two compared vectors are along with the same direction. For example, given three users' personality vectors ($ps_a = (1, 1, 1, 1, 1)^T$, $ps_b = (2, 2, 2, 2, 2)^T$, and $ps_c = (5, 5, 5, 5, 5)^T$), we can obtain the Cosine similarity results: $Sim_{Cosine}(ps_a, ps_b) = Sim_{Cosine}(ps_a, ps_c) = 1$, but in fact, user

$$Simp(u, v) = \frac{1}{1 + \left(\sqrt{\sum_{k=1}^{5} w_k^2 (ps_u^k - ps_v^k)^2}\right)^2} \tag{4}$$

where w_k represents the weight of the kth personality trait (i.e., relative importance of influencing users' preference). To derive w_k, for each pair of two users in the training set (e.g., users a and b; see more details of our dataset splitting strategy in Sect. 5.2), we try to minimize the distance between $Simp(a, b)$ (i.e., user-user similarity based on personality; see Eq. 4) and $Simr(a, b)$ (i.e., user-user similarity based on rating behavior; see Eq. 5). The rationale behind this process is that users' personality might be reflected in their rating behavior (Ajzen 2005).

$$Simr(a, b) = \frac{\sum_{i \in I} (r_{a,i} - \overline{r_a})(r_{b,i} - \overline{r_b})}{\sqrt{\sum_{i \in I} (r_{a,i} - \overline{r_a})^2} \sqrt{\sum_{i \in I} (r_{b,i} - \overline{r_b})^2}} \tag{5}$$

In the above formula, I is the set of items rated by both users a and b, $r_{a,i}$ is the rating given to item i by user a, and $\overline{r_a}$ is the average rating given by user a.

More formally, the process of obtaining each personality trait's weight w_k is illustrated as follows:
argmin $f(w_k)$ = **argmin** $|Simp(a, b) - Simr(a, b)|$
By substituting Eq. 4, we obtain $\sum_{k=1}^{5} w_k^2 (ps_a^k - ps_b^k)^2 = \frac{1}{Simr(a,b)} - 1$.
We then define three notations x, A, and y:

$$x = \begin{bmatrix} w_1^2 \\ w_2^2 \\ \vdots \\ w_5^2 \end{bmatrix} \quad A^T = \begin{bmatrix} (ps_a^1 - ps_b^1)^2 \\ (ps_a^2 - ps_b^2)^2 \\ \vdots \\ (ps_a^5 - ps_b^5)^2 \end{bmatrix} \quad y = \frac{1}{Simr(a, b)} - 1$$

It hence becomes to find a nonnegative vector x to minimize $f(x) = \frac{1}{2}||Ax - y||^2$, subject to $x \geq 0$. We adopt the MatLab "lsqnonneg" function that has commonly been used to solve the Nonnegative Least Squares problem (NNLS) (Chen and Plemmons 2009) to find the vector, with the default parameter set (Lawson and Hanson 1995).

Therefore, suppose w_k can be derived through the above process and the similarity between two users u and v $Simp(u, v)$ can be subsequently calculated via Eq. 4. We then use an intuitive way to combine $Simp(u, v)$ with rating-based similarity $Simr(u, v)$ for computing the final user-user similarity score as follows:

$$Simpr(u, v) = \alpha \times Simr(u, v) + (1 - \alpha) \times Simp(u, v) \tag{6}$$

Footnote 10 continued
b should be more similar to user a than user c, which can be more accurately identified by the Euclidean distance measure.

where α manipulates the relative weights of the two similarity measures, i.e., $Simr(u, v)$ and $Simp(u, v)$. In our experiment, it is formally set as $0.8 * \frac{|I_u \cap I_v|}{min\{|I_u|, |I_v|\} + 0.5 * \frac{1}{min\{|I_u|, |I_v|\}}}$, where I_u and I_v respectively denote the set of items rated by users u and v. If two users have high proportion of common ratings relative to the size of their rated item sets, $Simr(u, v)$ will be assigned higher weight. Otherwise, $Simp(u, v)$'s weight will be higher. In addition, if the target user has no behavior records ($|I_u| = 0$), the pure personality-based similarity $Simp(u, v)$ will be used (i.e., $\alpha = 0$).

The rating predicted for an unknown item i for the target user u is computed by Eq. 7, which considers the average ratings with the aim of compensating for some variations in users' ratings[11] (Schafer et al. 2007):

$$Score_{Pref}(u, i) = \overline{r_u} + k \sum_{v \in \Omega_u} Simpr(u, v) \times (r_{v,i} - \overline{r_v}) \tag{7}$$

where $r_{v,i}$ denotes user v's rating for item i, $\overline{r_u}$ and $\overline{r_v}$ respectively represent user u's and user v's average ratings, k is equal to $\frac{1}{\sum_{v \in \Omega_u} |Simpr(u,v)|}$, and Ω_u refers to the set of u's neighbors who have rated item i.

Step 2: Adjusting diversity degree within the recommendation list

In the following, we explain our personality-based diversity adjusting strategy. Given the significant results shown in our survey's findings (Table 1), we assume users' personality traits can infer their inherent preference for recommendation diversity. Our purpose is thus to generate a recommendation list whose degree of diversity is closest to the target user's personality-estimated diversity preference, so as to achieve the optimal personalization in diversity:

$$Score_{PersonalizedDiversity}(u, i) = -|Div_{ps}(u) - Div_{R \cup \{i\}}(u)| \tag{8}$$

To be specific, we use a linear weighted summation to predict the user u's personality-based diversity preference $Div_{ps}(u)$, as inspired by the multiple linear regression results in Table 1.

$$Div_{ps}(u) = \theta_0 + \theta_1 ps_u^1 + \theta_2 ps_u^2 + \theta_3 ps_u^3 + \theta_4 ps_u^4 + \theta_5 ps_u^5 \tag{9}$$

where ps_u^k denotes the user's value of one personality trait, θ_k ($1 \le k \le 5$) indicates the strength of the relationship between this personality trait and the user's preference for diversity, and θ_0 is the constant term (i.e., the intercept). $Div_{R \cup \{i\}}(u)$ is the diversity degree within the current recommendation list after a new item is added, by means of calculating Shannon Entropy over the categories (e.g., genres of movies, types of groups) these recommended items belong to:

$$Div_{Rec}(u) = -\sum_{k \in K} P(k|u) \log_2(P(k|u)) \tag{10}$$

[11] Users vary in their use of a rating scale (Schafer et al. 2007). For instance, one optimistic user may consistently rate items 4 out of 5 stars, while a pessimistic user may often give 3 starts even though s/he likes the item.

where K is the set of categories existed in the current recommendation list, and $P(k|u)$ is the proportion of the entire items recommended to user u made up of the items belonging to the category k.

The parameters θ_k in Eq. 9 are obtained during the training process in the following way. Formally, we first simplify the formula by letting $\theta^T = [\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5]$ and $x = [1, ps_c^1, ps_c^2, ps_c^3, ps_c^4, ps_c^5]$, where user c belongs to the training set. Therefore, $Div_{ps}(c) = \theta^T x = h_\theta(x)$. We then define the cost function as $J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$, where m is the number of training examples, $h_\theta(x^{(i)})$ denotes the estimated diversity preference for the ith training example using the parameters θ, and $y^{(i)}$ denotes the actual diversity preference derived from the ith training example's behavior records via Shannon Entropy (see Eq. 2). We can then get the vector θ by minimizing the cost function $J(\theta)$. Specifically, we repeat updating θ_j until convergence: $\theta_j = \theta_j - a\frac{\partial}{\partial\theta_j}J(\theta)$, where the parameter a denotes the learning rate, and the partial derivation is calculated as:

$$\begin{aligned}\frac{\partial}{\partial\theta_j}J(\theta) &= \frac{\partial}{\partial\theta_j}\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 = \frac{1}{2m}\frac{\partial}{\partial\theta_j}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \\ &= 2 \times \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\frac{\partial}{\partial\theta_j}(h_\theta(x^{(i)}) - y^{(i)}) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\end{aligned}$$

We postulate that our approach as described above can achieve personalization in diversity. For instance, an item equipped with the capability of diversifying the recommendation list might be added if the user is expected to have strong diversity preference according to her/his personality values. Otherwise, more similar items will be included in the recommendation list. Moreover, our personality-based diversity adjusting strategy can be applicable in the cold-start setting because personality is used to measure user-user similarity in collaborative filtering.

5 Experiment

5.1 Data set

The experiment was carried out on Douban Interest Group dataset collected in our user survey (see Sect. 3). To the best of our knowledge, relative to the other datasets that also contain users' personality values, our dataset is the largest [1706 users, vs. 250 in Celli et al. (2013) and 181 in Chen et al. (2013)]. It also includes users' preference data, with 34,311 group joining records from those 1706 users in 2396 groups (e.g., user u joins group i). As explicit ratings are not available in this dataset, we quantified the user u's preference for group i based on her/his implicit behavior feedback: $t_{u,i} = \langle u, i, \overline{r_{u,i}}\rangle$, where $\overline{r_{u,i}}$ is the normalized activity degree in the range [0,1] (treated as a virtual rating). Formally, the activity degree $r_{u,i}$ is calculated as:

$$r_{u,i} = Num_{likes}(u, i) + Num_{recommendations}(u, i) + Num_{comments}(u, i) \qquad (11)$$

where $Num_{likes}(u,i)$, $Num_{recommendations}(u,i)$, and $Num_{comments}(u,i)$ respectively denote the number of likes, the number of recommendations, and the number of comments provided by the user u in group i. In social networking sites, “Like”, “Recommend”, and “Comment” have been commonly regarded as typical activities users have often done to express their opinions and emotions (Vries et al. 2012). According to the statistics in Douban Interest Group dataset, the average numbers of likes, recommendations, and comments per user in respect of a group are 15.7 (SD = 26.8), 13.2 (SD = 5.2), and 29.8 (SD = 93.4) respectively. Therefore, we combined these activities linearly with equal weights, not only because their mean values have similar orders of magnitude, but also because all of the activities can reflect users’ preference (Nadkarni and Hofmann 2012; Thackeray et al. 2012). The activity degree is then normalized into [0,1] via the logarithmic form of normalization: $\overline{r_{u,i}} = \frac{\log_{10} r_{u,i}}{\log_{10} max}$, where max gives the maximum value among all of the samples.

5.2 Evaluation procedure and metrics

We randomly selected 80% of the 1706 users to train the model and then tested on the remaining 20% users. During the training phase, the parameters w_k in Eq. 4 and θ_k in Eq. 9 were derived with training users’ behavior records and personality values. During the prediction phase, for each test user, we randomly selected a subset of her/his behavior records to calculate the hybrid user-user similarity via Eq. 6. In the cold-start setting, the amount of each test user’s behavior records varies in the range from 0 to 5: “0” refers to the pure cold-start scenario, and “5” indicates that each test user has five behavior records used for training.

We performed tenfold cross validation, and evaluated the recommendation performance in terms of both accuracy and diversity.

Specifically, we measured accuracy via the commonly used metrics including *precision*, *recall*, *F1-score*, and *nDCG*:

- **Precision** (Powers 2011) is a measure of exactness: $precision = \frac{|R \cap T|}{|R|}$, where R is the set of recommended groups, and T is the set of groups actually joined by the user.
- **Recall** (Powers 2011) is a measure of completeness: $recall = \frac{|R \cap T|}{|T|}$, which refers to the proportion of joined groups that are included in the recommendation list.
- **F1-score** (Powers 2011) is a combination of precision and recall: $F1\text{-}score = \frac{2 \times precision \times recall}{precision + recall}$.
- **nDCG** (Järvelin and Kekäläinen 2002) refers to the normalized discounted cumulative gain, which is a measure of ranking accuracy where positions are discounted logarithmically. Formally, the discounted cumulative gain (DCG) accumulated at a particular rank position p is defined as $DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)}$, where $rel_i = 1$ if the item is contained in real data; otherwise $rel_i = 0$.

  On the other hand, we measured diversity via three metrics including *α-nDCG*, *Adaptive α-nDCG* and our proposed *Diversity Fitness (DivFit)*, where *α-nDCG* is a standard metric used to calculate the diversity degree within recommendation list, while *Adaptive α-nDCG* and *DivFit* are used to measure personalization in diversity.

- α-**nDCG** is a popularly-used metric to evaluate the diversity considering the ranking position (Clarke et al. 2008). It scores a ranking by rewarding newly-found items and penalizing the "nuggets already seen". In our experiment, we used group types as "nuggets". Formally, the gain vector of α-*nDCG* is defined as (Clarke et al. 2008):
$$G[k] = \sum_{c \in C} P(i_k^u \in c) \times (1 - \alpha)^{q_{c,k-1}^u} \tag{12}$$
where $q_{c,k-1}^u = \sum_{j=1}^{k-1} P(i_j^u \in c)$, which denotes the number of groups ranked up to position $k-1$ that contain type c in the recommendation list for user u ($q_{c,0}^u = 0$). In addition, $P(i_k^u \in c) = 1$ if the kth group in the recommendation list for user u contains type c; otherwise, $P(i_k^u \in c) = 0$. α is a constant set to control the magnitude of penalty for redundant recommendations with the aim of balancing relevance and diversity (higher value of α indicates the larger penalty). We set $\alpha = 0.5$ in our experiment to give equal importance to relevance and diversity, which is also a default value used in the diversity task of TREC 2009 Web track.[12]
- **Adaptive** α-**nDCG** is an advanced version of α-*nDCG*, which measures the degree to which diversity is personalized for each user by incorporating a personalization factor $P(c|u)$. Concretely, the gain vector of *Adaptive* α-*nDCG* is defined as (Eskandanian et al. 2017):
$$G'[k] = \sum_{c \in C} P(i_k^u \in c) \times P(c|u) \times (1 - \alpha')^{q_{c,k-1}^u} \tag{13}$$
where $P(c|u)$ refers to the relative appearance of group type c among all the groups user u has joined previously. Similar to Eq. 12, α' is a factor to penalize the redundancy of items in a rank list, which is determined based on pre-trials. Large value of α' diminishes the influence of $P(c|u)$.
- **Diversity Fitness (DivFit)** is our newly proposed metric to measure the personalization in diversity, which concretely calculates the fitness between the diversity degree $Div_{Rec}(u)$ within the top-N recommendation list and the user's actual diversity preference $Div_{Act}(u)$:
$$DivFit = \frac{1}{k} \sum_{u=1}^{k} |Div_{Act}(u) - Div_{Rec}(u)| \tag{14}$$
where k is the number of test users. $Div_{Rec}(u)$ is calculated by means of Shannon Entropy over the group types these recommended items belong to (refer to Eq. 10), and $Div_{Act}(u)$ is calculated based on user u's actual behavior records via Shannon Entropy (refer to Eq. 2). A smaller $DivFit$ means that the diversity of the recommendation list has a better fit to the user's actual diversity preference.

[12] https://trec.nist.gov/data/web09.html.

Table 2 Summary of compared methods

	Algorithm	Diversity oriented	Personalized diversity oriented
Our method	Personality-based greedy re-ranking (PB Greedy)	√	√
Variations of our method	Step 1 of PB Greedy: Personality-based collaborative filtering (PB_Step1)		
	Step 2 of PB Greedy: Personality-based diversity adjusting (PB_Step2)	√	√
Related methods	Baseline: Rating-based collaborative filtering (RB) (Su and Khoshgoftaar 2009)		
	Basic greedy re-ranking (GreedyRR) (Bradley and Smyth 2001)	√	
	Maximal Marginal Relevance (MMR) (Carbonell and Goldstein 1998)	√	
	Adaptive Maximal Marginal Relevance (AdaMMR) (Di et al. 2014)	√	√
	Clustering-based collaborative filtering (Clustering) (Eskandanian et al. 2017)	√	√

5.3 Compared methods

We compared our approach with five related methods, as well as two variations of our approach (i.e., PB_Step1 and PB_Step2). These algorithms can be classified into three categories: non-diversity-oriented (*RB* and *PB _Step1*), diversity-oriented (*GreedyRR* and *MMR*), and personalized diversity-oriented (*AdaMMR*, *Clustering*, and *PB_Step2*) (see summary in Table 2). The detailed description of each method is given below.

5.3.1 Non-diversity-oriented methods

- **RB** (Su and Khoshgoftaar 2009): RB is the standard rating-based CF. Referring to Eq. 5, the user-user similarity is calculated only on the basis of users' behavior records, where $r_{u,i}$ denotes user u's normalized activity degree in group i (see Eq. 11). The user u's preference for an unknown group can then be obtained by replacing $Simpr(u, v)$ with $Simr(u, v)$ in Eq. 7.
- **PB_Step1**: This method only includes the first step of our proposed personality-based greedy re-ranking approach (see Fig. 5). That is, personality is only used to measure users' preference for items, but not for adjusting recommendations' diversity.

5.3.2 Diversity-oriented methods

- **GreedyRR** (Bradley and Smyth 2001): The framework of the basic Greedy re-ranking approach is similar to ours. The main difference lies in the definition of the objective function. In GreedyRR, the primary objective is to optimize the trade-off between similarity and diversity (see Eq. 15), but individual users' diversity preference is not taken into consideration:

$$Score_{final}(u, i) = \lambda * Score_{Pref}(u, i) + (1 - \lambda) * Div_{R\cup\{i\}}(u) \tag{15}$$

where $Score_{Pref}(u, i)$ refers to the user u's preference for group i, which can be estimated by the rating-based CF (i.e., RB). $Div_{R\cup\{i\}}(u)$ represents the diversity degree within the recommendation list for u after adding group i, which can be calculated via Shannon Entropy (Eq. 10). λ denotes the adjusting parameter, which can be determined through experimental trials (the same as adjusting parameters β in Eq. 3, and γ in Eq. 16 and Eq. 17).

- **MMR** (Carbonell and Goldstein 1998): The Maximal Marginal Relevance is a measure for quantifying the extent of dissimilarity between the item currently being considered and those already selected. Higher MMR means the considered item is not only relevant to the user's preference but also contains minimal similarity to previously selected items. With MMR, the objective function (Eq. 3), is redefined as:

$$Score_{final}(u, i) = \gamma * Score_{Pref}(u, i) - (1 - \gamma) * max_{j\in T} Sim(i, j) \tag{16}$$

where Sim denotes a similarity score between the "item under consideration" and "an item that has already been selected" with respect to the group types they belong to, which is calculated by Jaccard Index (Di et al. 2014), and γ is used to control the similarity-diversity balance (higher γ leads to higher similarity, whereas lower γ results in higher diversity).

5.3.3 Personalized diversity-oriented methods

- **AdaMMR** (Di et al. 2014): As mentioned in Sect. 2 "Related Work", Di *et al.* proposed an Adaptive Maximal Marginal Relevance to customize the degree of recommendation diversity for the Top-N recommendation list. Their main assumption is that users who have explored items with different properties in the past might be more willing to accept diverse recommendations at present. To be specific, they considered a user's diversity inclination over different item attributes, such as the movie's genre, actor, and director. For instance, given an attribute a, they interpreted a high entropy as the user's inclination to choose items with different values of a. Conversely, a low entropy value was interpreted as the user's willingness to choose similar items in terms of that attribute. Therefore, they defined the similarity measure $Sim(i, j)$ in Eq. 16 as $Sim(i, j) = \frac{\sum_{a\in A} \omega_a \cdot Sim_a(i,j)}{max\{\omega_a\}\cdot|A|}$, where $Sim_a(i, j)$ is the similarity score between i and j with respect to attribute a. For weights ω, if a user has high propensity for diversity w.r.t. an attribute a, high value will be assigned to ω_a. On the contrary, for a user with low propensity for diversity w.r.t. attribute a, low ω_a will be assigned.
For the comparison, in our experiment that involves only one attribute (i.e., group type), we revised this method as:

$$Score_{final}(u,i) = \gamma * Score_{Pref}(u,i) \\ -(1-\gamma) * max_{j \in T}(Entropy_{normalize} * Sim(i,j)) \tag{17}$$

We used the normalized entropy (in range [0,1]) as weight to adjust the item–item similarity, for which the entropy is calculated with the user's behavior records (refer to Eq. 2).

- **Clustering** (Eskandanian et al. 2017): The core idea of this method is, because different users' propensities for diversity may be different, a pre-filtering clustering approach was proposed to group users with similar levels of diversity preference. Specifically, a user's diversity preference was computed as the distribution over categories (i.e., group types in our case) in her/his user profile. For the target user, the closest cluster was assigned through a partition-based K-Medoids clustering algorithm (Kaufman and Rousseeuw 1987). Then, the standard rating-based CF (i.e., RB) was performed on the cluster of users that the target user belongs to.
- **PB_Step2**: As the other variation of our approach, personality is only used to adjust recommendation diversity (i.e., Step 2 of PB Greedy), but users' preference for items are purely predicted with their behavior records without considering their personality values.

5.4 Results and analysis

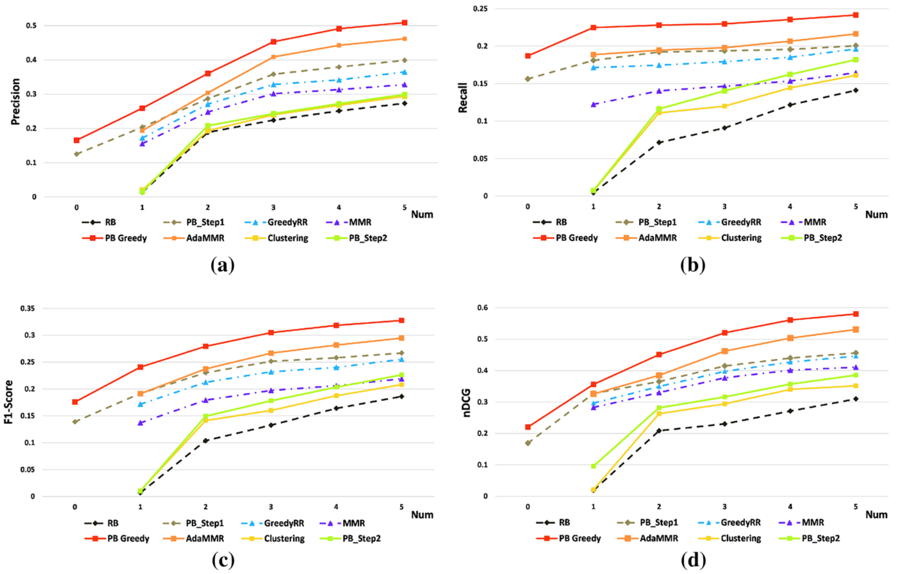
In our experiment, we set N as 10 (the number of items in a recommendation list). The optimal values of parameters in all the methods were identified through experimental trials. Specifically, the neighborhood size for CF is set as 300, the learning rate a of the gradient descent in Eq. 9 is set as 0.005, the penalty parameter α' in the metric Adaptive α-nDCG (Eq. 13) is set as 0.1, the adjusting parameter β in Eq. 3 is 0.9, λ in basic greedy re-ranking (Eq. 15) is 0.5, and γ in the MMR-based approaches (Eqs. 16 and 17) is 0.5.

5.4.1 Cold-start scenarios

First of all, we compared those methods in cold-start scenarios (the number of training records used for each test user ranges from 0 to 5).

In terms of the recommendation accuracy (see Fig. 6), we can see only our proposed personality-based greedy re-ranking approach (PB Greedy) and the variation PB_Step1 are able to produce recommendations in the pure cold-start condition (i.e., the number of training records is 0), which is because they take into account user personality in the recommendation process. The comparison between these two methods further shows that PB Greedy performs better, probably because it additionally achieves personalization in diversity which is tailored to individual users' intrinsic needs based on their personality.

Then, when test users' behavior records (used for training) increase from 1 to 5 ($1 \leq Num \leq 5$), we find that all of the methods obtain significant improvements against the baseline rating-based CF (RB) ($p < 0.05$ via two-tailed paired t test).



(a) (b) (c) (d)

Fig. 6 Comparison of different algorithms' recommendation accuracy in cold start scenarios (Num refers to the number of behavior records from each test user used in the training phase). Dashed lines with diamond markers indicate the non-diversity-oriented methods; Dash-dot lines with triangle markers indicate the diversity-oriented methods; Solid lines with square markers indicate the personalized diversity-oriented methods. **a** Precision. **b** Recall. **c** F1-score. **d** nDCG

Among them, our approach performs the best regarding the metrics precision, recall, F1-score, and nDCG, followed by AdaMMR that is also personalized diversity oriented. PB_Step1, which does not consider using personality to adjust recommendation diversity, is the third best. Relatively, the accuracy of PB_Step2 and Clustering is lower. As for PB_Step2, it is possibly because of its difficulty in predicting users' preference with few behavior records. For Clustering, it is probably due to the lack of sufficient training data to cluster users, which may lead to biases in locating neighbors.

On the other hand, Fig. 7 illustrates the diversity performance of these compared recommendation approaches. Regarding the diversity metric α-*nDCG* (see Fig. 7a), the diversity-oriented approaches perform best, where MMR performs slightly better than GreedyRR ($p > 0.05$). The possible reason is that their main idea of maximizing the recommendation diversity enables them to reduce the penalty of recommending redundant items, and consequently increase the value of α-*nDCG* (Clarke et al. 2008). In comparison, the personalized diversity-oriented approaches (e.g., PB Greedy, AdaMMR, PB_Step2, and Clustering) achieve relatively worse diversity performance, probably because they tend to produce more similar items within a recommendation list when a user has low diversity preference, which discounts the gains in α-*nDCG* (refer to Eq. 12). However, both of these two types of diversity-oriented approaches perform significantly ($p < 0.05$) better than PB_Step1 and the baseline RB which generate the recommendations without taking diversity into consideration.

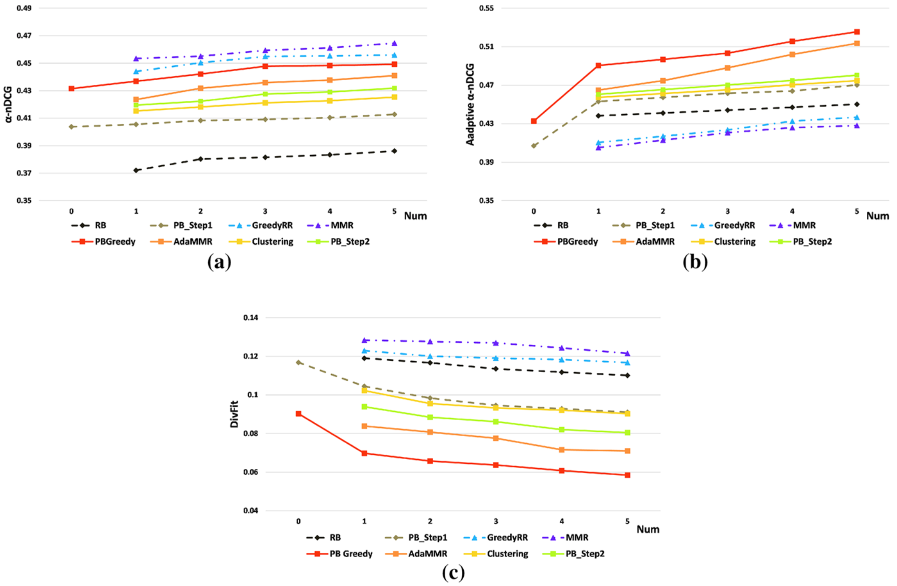


Fig. 7 Comparison of different algorithms' recommendation diversity in cold start scenarios. **a** α-nDCG. **b** Adaptive α-nDCG. **c** DivFit

With respect to the personalized diversity metrics $Adaptive\ \alpha\text{-}nDCG$ (see Fig. 7b, the higher, the better) and $DivFit$ (see Fig. 7c, the lower, the better), we obtain consistent results. That is, compared with the non-diversity-oriented approaches (PB_Step1 and baseline RB) and diversity-oriented methods (MMR and GreedyRR), the methods that consider individual users' diversity preference perform significantly better ($p < 0.05$), among which our approach PB Greedy achieves the best result, followed by AdaMMR, PB_Step2, and Clustering. As for GreedyRR and MMR, they even perform worse than the non-diversity-oriented methods, which is probably because presenting diverse recommendations to those who actually have low diversity preference leads to great deviation in achieving personalization in diversity. Combined with the results of α-$nDCG$ (refer to Fig. 7a), it is reasonable to see that the "one-size-fits-all" solutions that attempt to maximize recommendation diversity for all users may not be suitable for matching individual users' inherent propensity for diversity.

5.4.2 Overall comparisons involving non cold-start scenarios

In addition to the cold-start setting, we also evaluated the effectiveness of our approach in non-cold-start scenarios, where each test user is equipped with more than 5 behavior records for training (up to 20 because it is the average number of groups joined by our users, see Fig. 4).

The overall comparison results in respect of accuracy metrics are shown in Table 3, where we can see when test users have more behavior records for training, all of the methods' accuracy values are improved. This observation is reasonable, because incor-

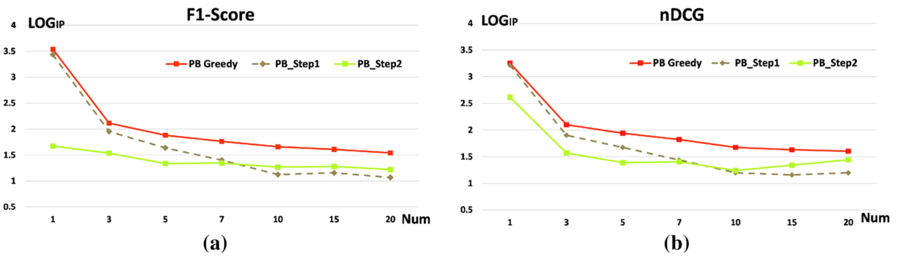


Fig. 8 Accuracy improvement percentage of our approach PB Greedy and two variations (i.e., PB_Step1 and PB_Step2) with the increase of *Num* (the number of the test user's behavior records used for training) (*Note:* LOG_{IP} is the logarithm scale measured as $\log_{10}(Improvementpercentage*100)$). **a** Improvement percentage in terms of F1-score. **b** Improvement percentage in terms of nDCG

porating more behavior records may allow these methods to better understand users' preference. Moreover, we find our proposed personality-based greedy re-ranking method (PB Greedy) still significantly ($p < 0.05$) outperforms the other seven methods in the non cold-start scenarios, but its "improvement percentage"[13] against the baseline RB is decreasing (see Fig. 8a, b), which may be because the rating-based similarity will account for higher weight than that based on personality when calculating the final user-user similarity (refer to Eq. 6). Relatively, the rating-based CF achieves better accuracy performance when more behavior data are available for locating nearest neighbors.

In order to identify the actual role of personality in each step of our proposed approach, we compared the accuracy improvement percentage of our approach PB Greedy and two variations (i.e., PB_Step1 and PB_Step2) (see Fig. 8). Our results demonstrate that when users have at least 10 training records, PB_Step2 can be superior to PB_Step1 (which only uses personality to estimate users' preference for items) regarding accuracy metrics, implying that utilizing personality to adjust recommendation diversity might be more valuable to improve the recommendation accuracy when the behavior records are sufficient for training.

Table 4 shows the overall comparison results w.r.t. diversity metrics in both cold-start and non-cold-start scenarios. Similar to the accuracy results presented in Table 3, we observe that all the methods' diversity performances are enhanced when incorporating more behavior records of test users for training. However, the absolute value of the diversity improvement percentage is lower relative to that of accuracy. For instance, the maximum average improvement percentages[14] of diversity (w.r.t. the metric *α-nDCG*) and accuracy (w.r.t. $F1\text{-}score$) are 18.39% (SD = 2.34%, via MMR) and 546.55% (SD = 1181.95%, via PB Greedy) respectively. The results suggest that accuracy may be more sensitive to the size of training data, compared to diversity.

[13] $Improvementpercentage(IP) = \frac{Value_{testmodel} - Value_{Baseline}}{Value_{Baseline}}$, where $Value_{Baseline}$ and $Value_{testmodel}$ respectively denote the performance of the baseline RB approach and the test model such as PB Greedy.

[14] *Average improvement percentage (Average IP)*$=\frac{\sum_{n \in N} IP_{Num=n}}{|N|}$, where N refers to the set of training data size ($N = \{1, 3, 5, 7, 10, 15, 20\}$).

Table 3 Overall comparison results in respect of accuracy metrics (*note*: *Num* refers to the number of the test user's behavior records used for training

	Non-diversity-oriented methods		Diversity-oriented methods		Personalized diversity-oriented methods			
	Baseline: RB^{1}	PB_Step1^{2}	GreedyRR^{3}	MMR^{4}	Our method: PB Greedy^{5}	AdaMMR^{6}	Clustering^{7}	PB_Step2^{8}
Accuracy Metric: Precision								
Num = 1	0.0139	0.2038 (1363.8%)^{1}	0.1722 (1136.4%)^{1}	0.1557 (1018.2%)^{1}	**0.2590** **(1760.4%)**^{1,2,3,4,6,7,8}	0.1937 (1290.9%)^{1}	0.0203 (45.45%)^{1}	0.0150 (7.71%)^{1}
Num = 3	0.2243	0.3584 (59.82%)^{1}	0.3282 (46.35%)^{1}	0.3015 (34.44%)^{1}	**0.4531** **(102.01%)**^{1,2,3,4,6,7,8}	0.4089 (82.31%)^{1}	0.2399 (6.96%)^{1}	0.2430 (8.34%)^{1}
Num = 5	0.2734	0.3989 (45.91%)^{1}	0.3649 (33.47%)^{1}	0.3278 (19.91%)^{1}	**0.5091** **(86.18%)**^{1,2,3,4,6,7,8}	0.4620 (68.98%)^{1}	0.2939 (7.5%)^{1}	0.2991 (9.39%^{1})
Num = 7	0.3076	0.4129 (34.25%)^{1}	0.3763 (22.35%)^{1}	0.3443 (11.93%)^{1}	**0.5244** **(70.49%)**^{1,2,3,4,6,7,8}	0.4797 (55.97%)^{1}	0.3359 (9.22%)^{1}	0.3449 (12.11%^{1})
Num = 10	0.3306	0.4332 (31.03%)^{1}	0.3881 (17.38%)^{1}	0.3513 (6.24%)^{1}	**0.5343** **(61.61%)**^{1,2,3,4,6,7,8}	0.5038 (52.37%)^{1}	0.3919 (18.53%)^{1}	0.3983 (20.46%)^{1}
Num = 15	0.3532	0.4464 (26.41%)^{1}	0.3944 (11.68%)^{1}	0.3810 (7.89%)^{1}	**0.5476** **(55.05%)**^{1,2,3,4,6,7,8}	0.5165 (46.24%)^{1}	0.4320 (22.33%)^{1}	0.4380 (24.02%)^{1}
Num = 20	0.3657	0.4473 (22.32%)^{1}	0.4103 (12.18%)^{1}	0.4051 (10.76%)^{1}	**0.5566** **(52.21%)**^{1,2,3,4,6,7,8}	0.5215 (42.61%)^{1}	0.4535 (24.02%)^{1}	0.4654 (27.25%)^{1}
Accuracy Metric: Recall								
Num = 1	0.0045	0.1811 (3926.8%)^{1}	0.1715 (3713.5%)^{1}	0.1224 (2620.9%)^{1}	**0.2249** **(4899.4%)**^{1,2,3,4,6,7,8}	0.1887 (4095.4%)^{1}	0.0069 (53.70%)^{1}	0.0075 (66.55%)^{1}
Num = 3	0.0909	0.1938 (113.12%)^{1}	0.1794 (97.23%)^{1}	0.1466 (61.22%)^{1}	**0.2298** **(152.69%)**^{1,2,3,4,6,7,8}	0.1980 (117.73%)^{1}	0.1200 (31.98%)^{1}	0.1404 (54.43%)^{1}
Num = 5	0.1411	0.2007 (42.22%)^{1}	0.1963 (39.14%)^{1}	0.1645 (16.56%)^{1}	**0.2416** **(71.22%)**^{1,2,3,4,6,7,8}	0.2164 (53.38%)^{1}	0.1611 (14.18%)^{1}	0.1820 (28.99%)^{1}

Table 3 continued

	Non-diversity-oriented methods		Diversity-oriented methods		Personalized diversity-oriented methods			
	Baseline: RB^{1}	PB_Step1^{2}	GreedyRR^{3}	MMR^{4}	Our method: PB Greedy^{5}	AdaMMR^{6}	Clustering^{7}	PB_Step2^{8}
Num = 7	0.1714	0.2066	0.2076	0.1886	**0.2599**	0.2307	0.1963	0.2202
		(20.55%)^{1}	(21.16%)^{1}	(10.04%)^{1}	**(51.64%)**^{1,2,3,4,6,7,8}	(53.38%)^{1}	(14.57%)^{1}	(28.49%)^{1}
Num = 10	0.1953	0.2049	0.2212	0.2127	**0.2689**	0.2413	0.2282	0.2292
		(4.89%)^{1}	(13.27%)^{1}	(8.89%)^{1}	**(37.05%)**^{1,2,3,4,6,7,8}	(23.56%)^{1}	(16.84%)^{1}	(17.37%)^{1}
Num = 15	0.2092	0.2264	0.2348	0.2277	**0.2789**	0.2521	0.2430	0.2430
		(8.22%)^{1}	(12.25%)^{1}	(8.86%)^{1}	**(33.34%)**^{1,2,3,4,6,7,8}	(20.50%)^{1}	(16.18%)^{1}	(16.18%)^{1}
Num = 20	0.2455	0.2591	0.2577	0.2579	**0.3071**	0.2821	0.2695	0.2712
		(5.55%)^{1}	(4.98%)^{1}	(5.04%)^{1}	**(25.11%)**^{1,2,3,4,6,7,8}	(14.90%)^{1}	(9.79%)^{1}	(10.47%)^{1}
Accuracy Metric: F1-score								
Num = 1	0.0068	0.1918	0.1783	0.1370	**0.2407**	0.1912	0.0103	0.0099
		(2720.9%)^{1}	(2427.3%)^{1}	(1915.6%)^{1}	**(3440.8%)**^{1,2,3,4,6,7,8}	(2711.4%)^{1}	(51.60%)^{1}	(46.95%)^{1}
Num = 3	0.1325	0.2516	0.2320	0.1973	**0.3049**	0.2668	0.1599	0.1780
		(89.82%)^{1}	(75.02%)^{1}	(48.86%)^{1}	**(130.07%)**^{1,2,3,4,6,7,8}	(101.30%)^{1}	(20.72%)^{1}	(34.30%)^{1}
Num = 5	0.1862	0.2670	0.2553	0.2191	**0.3277**	0.2948	0.2081	0.2263
		(43.46%)^{1}	(37.16%)^{1}	(17.68%)^{1}	**(76.03%)**^{1,2,3,4,6,7,8}	(58.36%)^{1}	(11.81%)^{1}	(21.58%)^{1}
Num = 7	0.2201	0.2754	0.2676	0.2437	**0.3475**	0.3116	0.2478	0.2688
		(25.12%)^{1}	(21.58%)^{1}	(10.71%)^{1}	**(57.88%)**^{1,2,3,4,6,7,8}	(41.55%)^{1}	(12.60%)^{1}	(22.11%)^{1}
Num = 10	0.2456	0.2782	0.2818	0.2650	**0.3577**	0.3263	0.2885	0.2910
		(13.28%)^{1}	(14.76%)^{1}	(7.89%)^{1}	**(45.67%)**^{1,2,3,4,6,7,8}	(32.89%)^{1}	(17.46%)^{1}	(18.50%)^{1}
Num = 15	0.2628	0.3004	0.2944	0.2851	**0.3696**	0.3388	0.3111	0.3126
		(14.34%)^{1}	(12.04%)^{1}	(8.49%)^{1}	**(40.66%)**^{1,2,3,4,6,7,8}	(28.94%)^{1}	(18.39%)^{1}	(18.98%)^{1}

Table 3 continued

	Non-diversity-oriented methods		Diversity-oriented methods		Personalized diversity-oriented methods			
	Baseline: RB^1	PB_Step1^2	GreedyRR^3	MMR^4	Our method: PB Greedy^5	AdaMMR^6	Clustering^7	PB_Step2^8
Num = 20	0.2938	0.3281	0.3166	0.3151	**0.3958**	0.3661	0.3381	0.3427
		(11.70%)^1	(7.76%)^1	(7.26%)^1	**(34.75%)**^{1,2,3,4,6,7,8}	(24.62%)^1	(15.09%)^1	(16.65%)^1
Accuracy Metric: nDCG								
Num = 1	0.0188	0.3272	0.2964	0.2829	**0.3558**	0.3257	0.0209	0.0957
		(1642.9%)^1	(1478.7%)^1	(1405.4%)^1	**(1795.1%)**^{1,2,3,4,6,7,8}	(1634.8%)^1	(11.26%)^1	(409.71%)^1
Num = 3	0.2302	0.4145	0.3794	0.3769	**0.5202**	0.4618	0.2939	0.3160
		(80.03%)^1	(72.61%)^1	(63.67%)^1	**(125.92%)**^{1,2,3,4,6,7,8}	(100.59%)^1	(27.65%)^1	(37.26%)^1
Num = 5	0.3098	0.4561	0.4460	0.4109	**0.5801**	0.5308	0.3514	0.3853
		(47.23%)^1	(43.97%)^1	(32.63%)^1	**(87.24%)**^{1,2,3,4,6,7,8}	(71.33%)^1	(13.44%)^1	(24.39%)^1
Num = 7	0.3590	0.4574	0.4549	0.4233	**0.5975**	0.5474	0.3995	0.4497
		(27.42%)^1	(26.71%)^1	(17.93%)^1	**(66.45%)**^{1,2,3,4,5,7,8}	(52.48%)^1	(11.30%)^1	(25.28%)^1
Num = 10	0.4092	0.4739	0.4643	0.4309	**0.6029**	0.5642	0.4532	0.4805
		(15.82%)^1	(13.48%)^1	(5.32%)^1	**(47.34%)**^{1,2,3,4,6,7,8}	(37.90%)^1	(10.75%)^1	(17.45%)^1
Num = 15	0.4237	0.4847	0.4727	0.4458	**0.6047**	0.5708	0.4912	0.5163
		(14.39%)^1	(11.56%)^1	(5.20%)^1	**(42.71%)**^{1,2,3,4,6,7,8}	(34.72%)^1	(15.92%)^1	(21.85%)^1
Num = 20	0.4348	0.5035	0.4938	0.4619	**0.6091**	0.5811	0.5416	0.5549
		(15.80%)^1	(13.58%)^1	(6.25%)^1	**(40.10%)**^{1,2,3,4,6,7,8}	(33.65%)^1	(24.58%)^1	(27.63%)^1

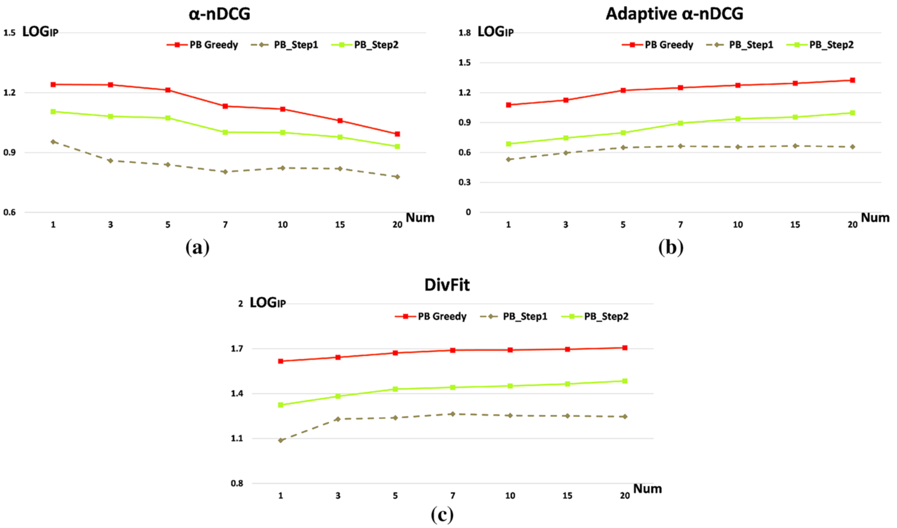
The value inside the parenthesis indicates the improvement percentage against the baseline *RB* approach, the number in superscript indicates that the method significantly ($p < 0.05$) outperforms the referred one in terms of the corresponding metric, and the best result in each line is in bold face)

Table 4 Overall comparison results in respect of diversity metrics

	Non-diversity-oriented methods		Diversity-oriented methods		Personalized diversity-oriented methods			
	Baseline: RB^{1}	PB_Step1^{2}	GreedyRR^{3}	MMR^{4}	Our method: PB Greedy^{5}	AdaMMR^{6}	Clustering^{7}	PB_Step2^{8}
Diversity Metric: α-nDCG								
Num = 1	0.3721	0.4055	0.4440	**0.4534**	0.4369	0.4236	0.4153	0.4195
		(8.99%)	(19.33%)^{1}	**(21.85%)**^{1,2,6,7,8}	(17.42%)^{1}	(13.84%)^{1}	(11.61%)^{1}	(12.75%)^{1}
Num = 3	0.3815	0.4091	0.4549	**0.4593**	0.4477	0.4359	0.4212	0.4275
		(7.22%)	(19.23%)^{1}	**(20.39%)**^{1,2,6,7,8}	(17.35%)^{1}	(14.25%)^{1}	(10.40%)^{1}	(12.06%)^{1}
Num = 5	0.3862	0.4128	0.4560	**0.4646**	0.4492	0.4410	0.4254	0.4319
		(6.90%)	(18.08%)^{1}	**(20.31%)**^{1,2,6,7,8}	(16.32%)^{1}	(14.19%)^{1}	(10.15%)^{1}	(11.84%)^{1}
Num = 7	0.3981	0.4234	0.4580	**0.4658**	0.4521	0.4449	0.4330	0.4381
		(6.36%)	(15.04%)^{1}	**(16.99%)**^{1,2,6,7,8}	(13.56%)^{1}	(11.75%)^{1}	(8.75%)^{1}	(10.03%)^{1}
Num = 10	0.4005	0.4270	0.4594	**0.4711**	0.4530	0.4499	0.4381	0.4406
		(6.64%)	(14.71%)^{1}	**(17.63%)**^{1,2,6,7,8}	(13.11%)^{1}	(12.35%)^{1}	(9.41%)^{1}	(10.01%)^{1}
Num = 15	0.4068	0.4336	0.4616	**0.4756**	0.4534	0.4541	0.4425	0.4454
		(6.59%)	(13.48%)^{1}	**(16.93%)**^{1,2,6,7,8}	(11.47%)^{1}	(11.63%)^{1}	(8.78%)^{1}	(9.49%)^{1}
Num = 20	0.4189	0.4440	0.4687	**0.4802**	0.4600	0.4586	0.4520	0.4545
		(6.00%)	(11.90%)^{1}	**(14.65%)**^{1,2,6,7,8}	(9.83%)^{1}	(9.50%)^{1}	(7.92%)^{1}	(8.52%)^{1}
Diversity Metric: Adaptive α-nDCG								
Num = 1	0.4383	0.4531	0.4105	0.4051	**0.4905**	0.4649	0.4573	0.4606
		(3.39%)	(− 6.34%)	(−7.55%)	**(11.92%)**^{1,2,3,4,7,8}	(6.08%)^{1}	(4.36%)	(4.85%)
Num = 3	0.4441	0.4616	0.4236	0.4207	**0.5033**	0.4881	0.4653	0.4702
		(3.94%)	(− 4.62%)	(− 5.27%)	**(13.34%)**^{1,2,3,4,7,8}	(9.91%)^{1}	(4.78%)	(5.57%)
Num = 5	0.4502	0.4703	0.4367	0.4281	**0.5255**	0.5136	0.4748	0.4804
		(4.46%)	(− 3.00%)	(− 4.92%)	**(16.71%)**^{1,2,3,4,7,8}	(14.07%)^{1}	(5.45%)	(6.28%)

Table 4 continued

	Non-diversity-oriented methods		Diversity-oriented methods		Personalized diversity-oriented methods			
	Baseline: RB^{1}	PB_Step1^{2}	GreedyRR^{3}	MMR^{4}	Our method: PB Greedy^{5}	AdaMMR^{6}	Clustering^{7}	PB_Step2^{8}
Num = 7	0.4534	0.4743	0.4397	0.4297	**0.5338**	0.5221	0.4792	0.4919
		(4.61%)	(− 3.01%)	(− 5.24%)	**(17.72%)**^{1,2,3,4,7,8}	(15.15%)^{1}	(5.68%)	(7.82%)
Num = 10	0.4576	0.4783	0.4410	0.4354	**0.5436**	0.5320	0.4884	0.5011
		(4.52%)	(− 3.64%)	(− 4.85%)	**(18.80%)**^{1,2,3,4,7,8}	(16.25%)^{1}	(6.73%)	(8.67%)
Num = 15	0.4605	0.4818	0.4460	0.4401	**0.5508**	0.5456	0.5013	0.5060
		(4.63%)	(− 3.13%)	(− 4.42%)	**(19.62%)**^{1,2,3,4,7,8}	(18.49%)^{1}	(8.86%)	(9.01%)
Num = 20	0.4643	0.4854	0.4523	0.4473	**0.5624**	0.5561	0.5115)	0.5156
		(4.54%)	(− 2.60%)	(− 3.67%)	**(21.12%)**^{1,2,3,4,7,8}9	(19.93%)^{1}	(10.16%)	(9.95%)
Diversity Metric: DivFit								
Num = 1	0.1190	0.1045	0.1230	0.1283	**0.0698**	0.0839	0.1022	0.0939
		(12.23%)^{1}	(− 3.32%)	(− 7.83%)	**(41.37%)**^{1,2,3,4,6,7,8}	(29.53%)^{1}	(14.10%)^{1}	(21.20%)^{1}
Num = 3	0.1135	0.0946	0.1191	0.1270	**0.0637**	0.0775	0.0933	0.0862
		(16.69%)^{1}	(− 4.86%)	(− 11.81%)	**(43.90%)**^{1,2,3,4,6,7,8}	(31.71%)^{1}	(17.84%)^{1}	(24.11%)^{1}
Num = 5	0.1101	0.0911	0.1168	0.1215	**0.0584**	0.0710	0.0903	0.0805
		(17.32%)^{1}	(− 6.02%)	(− 10.35%)	**(46.94%)**^{1,2,3,4,6,7,8}	(35.56%)^{1}	(18.01%)^{1}	(26.93%)^{1}
Num = 7	0.1099	0.0898	0.1114	0.1201	**0.0562**	0.0701	0.0890	0.0796
		(18.37%)^{1}	(− 1.33%)	(− 9.21%)	**(48.93%)**^{1,2,3,4,6,7,8}	(36.23%)^{1}	(19.07%)^{1}	(27.65%)^{1}
Num = 10	0.1060	0.0870	0.0940	0.1032	**0.0539**	0.0688	0.0876	0.0761
		(17.92%)^{1}	(11.38%)	(2.64%)	**(49.16%)**^{1,2,3,4,6,7,8}	(35.12%)^{1}	(17.38%)^{1}	(28.26%)^{1}
Num = 15	0.1020	0.0839	0.0888	0.0977	**0.0514**	0.0670	0.0835	0.0723
		(17.82%)^{1}	(13.01%)	(4.22%)	**(49.67%)**^{1,2,3,4,6,7,8}	(34.35%)^{1}	(18.23%)^{1}	(29.14%)^{1}
Num = 20	0.0973	0.0802	0.0876	0.0954	**0.0478**	0.0553	0.0807	0.0677
		(17.64%)^{1}	(10.02%)	(2.01%)	**(50.90%)**^{1,2,3,4,6,7,8}	(43.14%)^{1}	(17.06%)^{1}	(30.50%)^{1}



(a)

(b)

(c)

Fig. 9 Diversity improvement percentage of our approach PB Greedy and two variations (i.e., PB_Step1 and PB_Step2) with the increase of *Num* (the number of the test user's behavior records used for training). **a** Improvement percentage in terms of α-nDCG. **b** Improvement percentage in terms of Adaptive α-nDCG. **c** Improvement percentage in terms of DivFit

In addition, we find MMR still performs the best in terms of the diversity metric α-nDCG in the non cold-start condition. As for the metrics of measuring personalization of diversity (i.e., Adaptive α-nDCG and DivFit), our proposed personality-based greedy re-ranking method (PB Greedy) always significantly outperforms the other seven methods ($p < 0.05$) when the number of training data increases.

Moreover, as it can be seen in Fig. 9a, when richer behavior records are included in training, it shows a downward trend of the improvement percentage w.r.t. the diversity metric α-nDCG for all of the three personality-based methods PB Greedy, PB_Step1, and PB_Step2. One possible reason is that we give equal weights to relevance and diversity in α-nDCG (i.e., $\alpha = 0.5$, see Eq. 12), which means the gains will be discounted based on both recommendation position and redundancy. Specifically, in the cold-start setting, our methods that incorporate personality information can perform better than the baseline RB in terms of both accuracy and diversity (refer to Figs. 6 and 7a). Relatively, in the non-cold-start scenarios, the gap of the accuracy between our personality-based methods and the baseline RB may be reduced, which leads to the lower improvement percentage in terms of α-nDCG. On the other hand, regarding the two metrics that measure the personalization in diversity [i.e., Adaptive α-nDCG (Fig. 9b) and DivFit (Fig. 9c)], the improvement percentages of our personalized diversity-oriented methods PB Greedy and PB_Step2 steadily go up with the increase of the training data size. It is likely because larger amount of behavior records enables these methods to better reveal users' preference for diversity and hence more accurately adjust the diversity degree.

6 Discussion

6.1 Summary of our experimental findings

To sum up, the experimental results show that our personality-based greedy re-ranking approach outperforms the other compared methods (including both non-diversity-oriented and diversity-oriented methods) in terms of both accuracy and personalized diversity metrics, in both cold-start and non-cold-start scenarios. Our experiment also demonstrates the role of personality in the recommendation process. In particular, personality proves to be more valuable in improving recommendation accuracy in the cold-start setting, because it can help compensate for the lack of behavior data for training so as to locate the nearest neighbors more precisely. On the other hand, personality also helps achieve the personalization in diversity even when users have richer behavior history. Although the diversity-oriented methods such as the Maximal Marginal Relevance approach and the basic Greedy Re-ranking approach can generally obtain good performance in respect of the traditional diversity metric, they are unlikely to meet individual users' inherent propensity for diversity.

In our view, this research brings several practical implications to recommender systems. First of all, users' diversity preference can be inferred from their personality traits to a certain extent. Compared with related personalized diversity-oriented approaches that are based on behavior data (Di et al. 2014; Eskandanian et al. 2017), our personality-based approach allows the system to dynamically adjust recommendation diversity in both cold-start and non-cold-start scenarios. Therefore, we believe personality can be more useful in recommender systems to assist the system in better understanding users' inherent preference from the psychological aspect, and hence providing more personalized services. The drawback is that it might be a bit time consuming by explicitly acquiring users' personality through questionnaire [though the 44-item BFI adopted in our work can actually be completed within 5 min (John and Srivastava 1999)]. Thus, in the future, we will consider how to implicitly elicit users' personality from their generated data. For example, we may infer their personality from behavior data collected in one or multiple source domains, and then use it to address the cold-start issue in the target domain.

Secondly, our approach can be effective to improve both recommendation ranking accuracy (i.e., the metrics Precision, Recall, F1-score, and nDCG) and personalized diversity (the metrics Adaptive α-nDCG and DivFit). In other words, our system can not only determine the correct order of a set of items for each user, but also adjust the degree of recommendation diversity that may better match the user's diversity preference.

Thirdly, as the traditional diversity metrics such as α-nDCG do not take into account the degree of personalization in recommendation diversity, in this work we proposed a new metric called Diversity Fitness (DivFit). Through calculating the fitness between the diversity degree within the top-N recommendation list and the user's actual diversity preference, DivFit enables to measure whether a method could deliver proper amount of diversity within their recommendations, being tailored to individual users' needs.

6.2 Limitations

In designing our approach for personalized recommendation diversity, we used the neighborhood-based collaborative filtering due to their simple interpretation, where the user-item ratings stored in the system (i.e., the virtual scores transferred from users' implicit behavior feedback) are directly used to predict ratings for new items (Desrosiers and Karypis 2011). However, this kind of method may be sensitive to sparse data because two users are unlikely to have common ratings when few ratings are available. Moreover, the user-user similarity computation turns out to be the performance bottleneck, which in turn may make the whole process unsuitable for real-time recommendation generation (Sarwar et al. 2001). Therefore, in the future, it might be interesting to consider the model-based approach, whose main idea is to model user-item interactions as factors representing latent characteristics of the users and items in the system and then predict ratings of users for new items with the latent factors (Breese et al. 1998). In addition, besides the basic greedy re-ranking technique explored in our work, we may use other optimization strategies such as bounded greedy selection because they may help reduce calculation complexity by decreasing the number of iteration cycles (Bradley and Smyth 2001; Smyth and McClave 2001). Another potential limitation of our current work is that the proposed approach may suffer from the filter bubble problem (Nguyen et al. 2014). For instance, similar items may be recommended to an individual user if s/he is estimated to have low diversity preference, which may stop her/him from exploring a broader space of options. On the other hand, the global coverage of our recommender system may be influenced, since coverage is an important indicator of recommendation quality by measuring the degree to which recommendations cover the set of available items (Ge and Delgado-Battenfeld 2010).

As for the dataset we used in this work, due to the lack of explicit ratings, we estimate a user's preference for a group based on her/his activity degree (i.e., linearly combining the number of likes, comments, and recommendations s/he performed in the group). On one hand, we assign equal weights to these activities, which, however, could be enhanced as they may reflect different preferences of the user. On the other hand, more implicit measures such as user's viewing time or clicking frequency could be considered as well. We may also consider the other approaches to transform the user's implicit feedback to the explicit rating. For example, Hu et al. (2008) transferred the implicit behavior into two separate magnitudes with distinct interpretations: preferences and confidence levels, which has been proven to be useful in generating recommendations. In addition, the effectiveness of our proposed personalized diversity metric (i.e., Diversity Fitness) should be validated with different datasets containing a larger-scale of samples' behavior records.

7 Conclusions and future work

In recent years, although there are many attempts to optimize the diversity degree within a recommendation list, they commonly adopted a fixed strategy to adjust the diversity for all users. In this paper, we are interested in tackling the prob-

Table 5 Summary of our research questions, methodologies, and key results

RQ1: Whether and how would users' personality influence their preference for diversity and group types?	
Method	A user survey conducted on Douban Interest Group with 1706 users
Key results	1. Users' preference for diversity is significantly positively influenced by *Openness to Experience* and negatively influenced by *Extroversion*
	2. Users' preference for some particular group types is significantly influenced by all of the five personality traits
RQ2: How could we achieve personalized diversity tailored to individual users' intrinsic needs based on their personality?	
Method	A personality-based greedy re-ranking approach
Key results	1. Our approach performs the best in terms of both recommendation accuracy and personalized diversity metrics
	2. Personality can be more useful in enhancing recommendation accuracy when users have few behavior records
	3 Personality can better help achieve the personalization in recommendation diversity when users have richer behavior history

lem of how to achieve personalized diversity tailored to individual users' intrinsic needs. Table 5 briefly summarizes our research questions, adopted methods, and key results.

Specifically, we first conducted a large-scale user survey on Douban Interest Group (that contains user preference data for various groups) and found that some personality traits (such as *Openness to Experience* and *Extroversion*) can significantly influence users' preference for diversity. For instance, people who are more creative (with high *Openness to Experience*) and more introverted (with low *Extroversion*) are more inclined to join different types of groups. In addition, each personality trait has its own impact on users' preference for group type. For example, people who are suspicious (with low *Agreeableness*) tend to prefer "Movie" type groups, whereas those who are emotionally unstable (with high *Neuroticism*) are likely to prefer "Costume" type groups.

Inspired by the survey's findings, we have further developed a personality-based greedy re-ranking approach (PB Greedy) with the goal of achieving personalized recommendation diversity. In this method, personality is used to estimate users' diversity preference as well as to enhance user-user similarity measure in collaborative filtering. The items that obtain the best balance between accuracy and personalized diversity are selected into the recommendation list.

In the experiment, we compared our method with seven approaches including the non-diversity-oriented rating-based CF (i.e., RB), two diversity-oriented methods (GreedyRR and MMR), two personalized diversity oriented methods (AdaMMR and Clustering), and two variations of our approach (PB_Step1 and PB_Step2). The experimental results show: (1) Our method PB Greedy is significantly better at improving recommendation accuracy in terms of the metrics Precision, Recall, F1-score, and nDCG; (2) our approach also outperforms the others w.r.t. the personalized diversity metrics including Adaptive α-nDCG and DivFit, although MMR achieves the best performance in respect of the traditional diversity metric α-nDCG; (3) when more behavior records are used for training, our approach's accuracy improvement percentage is decreased, suggesting that personality can be more valuable in enhancing recommendation accuracy in cold-start scenarios. However, personalization degree of recommendation diversity is still improved and balanced well with recommendation accuracy when there is richer behavior history.

In the future, on one hand, we will try to address the limitations of our current approach as discussed in Sect. 6.2. For instance, we may compare the performance of memory-based and model-based collaborative filtering methods that are all incorporated with personality, in terms of both predictive ability and run-time complexity. Some other optimization strategies may also be adopted for the purpose of achieving better performance. For solving the potential filter bubble problem, one of the reasonable solutions might be using epsilon-greedy strategy (Auer et al. 2002). For solving the potential filter bubble problem, one of the reasonable solutions might be using epsilon-greedy strategy (Auer et al. 2002). That is, we may generate recommendations according to our personality-based greedy re-ranking approach with 1-epsilon probability (i.e., exploitation) and add random items with epsilon probability (i.e., exploration). The exploration part is used to step outside the bubble, and to capture users' potential preference shift. In this way, even if the user has low diversity preference, s/he may still have chance to approach some diverse recommendations. In addition, the measurement of user preference for a specific item (e.g., Douban interest group in our experiment) could be enhanced. We may also consider more implicit behavior (e.g., reviewing time spent on the interest group) to infer users' preference.

On the other hand, we will conduct user evaluations to empirically validate the practical benefits of our approach to online users. We will also try to improve the recommender interface design being adaptive to users' diversity preference. For instance, we may consider using a diversity-oriented interface (e.g., the organization based interface (Chen and Pu 2007)) to display recommendations for those users who have strong diversity preference (as this interface organizes items into different categories and the diversity across categories is maximized), while for users with low diversity preference, the standard ranked list interface that ranks all items by their similarity score would be used. We will recruit users to evaluate such system for identifying the impact of personality on their perceptions of and satisfaction with the interface.

Acknowledgements We thank all participants who took part in our user survey. We also thank reviewers for their suggestions and comments. In addition, we thank Hong Kong Research Grants Council (RGC) for sponsoring the research work (under Project RGC/HKBU12200415).

References

Adomavicius, G., Kwon, Y.: Toward more diverse recommendations: item re-ranking methods for recommender systems. In: Workshop on Information Technologies and Systems (WITS 2009), pp. 417–440. Citeseer (2009)

Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)

Ajzen, I.: Attitudes, Personality, and Behavior. McGraw-Hill Education, London (2005)

Armstrong, R.A.: When to use the bonferroni correction. Ophthalmic Physiol. Opt. **34**(5), 502–508 (2014)

Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Mach. Learn. **47**(2–3), 235–256 (2002)

Bradley, K., Smyth, B.: Improving recommendation diversity. In: Proceedings of the 12th Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2001), pp. 85–94 (2001)

Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI 1998), pp. 43–52. Morgan Kaufmann Publishers Inc. (1998)

Carbonell, J., Goldstein, J.: 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 (SIGIR 1998), pp. 335–336. ACM (1998)

Celli, F., Pianesi, F., Stillwell, D., Kosinski, M.: Workshop on computational personality recognition (shared task). In: Proceedings of the Workshop on Computational Personality Recognition (2013)

Chen, D., Plemmons, R.J.: Nonnegativity constraints in numerical analysis. Birth Numer. Anal. **10**, 109–140 (2009)

Chen, L., Pu, P.: Preference-based organization interfaces: aiding user critiques in recommender systems. User Model. **2007**, 77–86 (2007)

Chen, L., Wu, W., He, L.: How personality influences users' needs for recommendation diversity? In: Proceedings of the 31st ACM Conference on Human Factors in Computing Systems (CHI 2013 Extended Abstracts), pp. 829–834. ACM (2013)

Chen, L., Wu, W., He, L.: Personality and recommendation diversity. In: Emotions and Personality in Personalized Services, vol. 3, pp. pp–201. Springer International Publishing (2016)

Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008), pp. 659–666. ACM (2008)

Cohen, P., Cohen, J., Aiken, L.S., West, S.G.: The problem of units and the circumstance for pomp. Multivar. Behav. Res. **34**(3), 315–346 (1999)

Cronbach, L.J.: Theory of generalizability for scores and profiles. The Dependability of Behavioral Measurements pp. 161–188 (1972)

De Vries, L., Gensler, S., Leeflang, P.S.: Popularity of brand posts on brand fan pages: an investigation of the effects of social media marketing. J. Interact. Mark. **26**(2), 83–91 (2012)

Desrosiers, C., Karypis, G.: A comprehensive survey of neighborhood-based recommendation methods. In: Recommender Systems Handbook, pp. 107–144. Springer, Berlin (2011)

Di Noia, T., Ostuni, V.C., Rosati, J., Tomeo, P., Di Sciascio, E.: An analysis of users' propensity toward diversity in recommendations. In: Proceedings of the 8th ACM Conference on Recommender Systems (RecSys 2014), pp. 285–288. ACM (2014)

Digman, J.M.: Personality structure: emergence of the five-factor model. Annu. Rev. Psychol. **41**(1), 417–440 (1990)

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

Ge, M., Delgado-Battenfeld, C., Jannach, D.: Beyond accuracy: evaluating recommender systems by coverage and serendipity. In: Proceedings of the 4th ACM Conference on Recommender Systems (RecSys 2010), pp. 257–260. ACM (2010)

Helson, R., Soto, C.J.: Up and down in middle age: monotonic and nonmonotonic changes in roles, status, and personality. J. Pers. Soc. Psychol. **89**(2), 194 (2005)

Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. **22**(1), 5–53 (2004)

Hofmann, T.: Latent semantic models for collaborative filtering. ACM Trans. Inf. Syst. **22**(1), 89–115 (2004)

Hu, R., Pu, P.: Acceptance issues of personality-based recommender systems. In: Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys 2009), pp. 221–224. ACM (2009)

Hu, R., Pu, P.: A study on user perception of personality-based recommender systems. User Modeling, Adaptation, and Personalization (UMAP 2010), pp. 291–302 (2010)

Hu, R., Pu, P.: Enhancing collaborative filtering systems with personality information. In: Proceedings of the 5th ACM Conference on Recommender Systems (RecSys 2011), pp. 197–204. ACM (2011)

Hu, R., Pu, P.: Helping users perceive recommendation diversity. In: DiveRS@ RecSys, pp. 43–50 (2011)

Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: Proceedings of the 8th International Conference on Data Mining (ICDM 2008), pp. 263–272. IEEE (2008)

Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (2002)

John, O.P., Srivastava, S.: The big five trait taxonomy: history, measurement, and theoretical perspectives. Handb. Pers. Theory Res. **2**(1999), 102–138 (1999)

Kaiseler, M., Polman, R.C., Nicholls, A.R.: Effects of the big five personality dimensions on appraisal coping, and coping effectiveness in sport. Eur. J. Sport Sci. **12**(1), 62–72 (2012)

Kaminskas, M., Bridge, D.: Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Trans. Interact. Intell. Syst. **7**(1), 2 (2016)

Karumur, R.P., Nguyen, T.T., Konstan, J.A.: Personality, user preferences and behavior in recommender systems. Inf. Syst. Front. **6**, 1–25 (2017)

Kaufman, L., Rousseeuw, P.: Clustering by Means of Medoids. North-Holland, Amsterdam (1987)

Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C.: Explaining the user experience of recommender systems. User Model. User-Adap. Interact. **22**(4–5), 441–504 (2012)

Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (2009)

Lawson, C.L., Hanson, R.J.: Solving Least Squares Problems. SIAM, Philadelphia (1995)

McCrae, R.R., Terracciano, A.: Personality profiles of cultures: aggregate personality traits. J. Pers. Soc. Psychol. **89**(3), 407 (2005)

McNee, S.M., Riedl, J., Konstan, J.A.: Being accurate is not enough: how accuracy metrics have hurt recommender systems. In: Proceedings of the 24th ACM Conference on Human Factors in Computing Systems (CHI 2006 Extended Abstracts), pp. 1097–1101. ACM (2006)

Mourão, F., Fonseca, C., Araujo, C.S., Meira Jr, W.: The oblivion problem: exploiting forgotten items to improve recommendation diversity. In: DiveRS@ RecSys, pp. 27–34 (2011)

Nadkarni, A., Hofmann, S.G.: Why do people use facebook? Pers. Individ. Differ. **52**(3), 243–249 (2012)

Nguyen, T.T., Hui, P.M., Harper, F.M., Terveen, L., Konstan, J.A.: Exploring the filter bubble: the effect of using recommender systems on content diversity. In: Proceedings of the 23rd International Conference on World Wide Web (WWW 2014), pp. 677–686. ACM (2014)

Nunes, M.A.S., Hu, R.: Personality-based recommender systems: an overview. In: Proceedings of the 6th ACM Conference on Recommender Systems (RecSys 2012), pp. 5–6. ACM (2012)

Nunnally, J.C., Bernstein, I.H., Berge, J.M.T.: Psychometric Theory, vol. 226. McGraw-Hill, New York (1967)

Perrett, D., Schaffer, J., Piccone, A., Roozeboom, M., et al.: Bonferroni adjustments in tests for regression coefficients. Mult. Linear Regres. Viewp. **32**, 1–6 (2006)

Powers, D.M.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. J. Mach. Learn. Technol. **2**, 37–63 (2011)

Qian, G., Sural, S., Gu, Y., Pramanik, S.: Similarity between euclidean and cosine angle distance for nearest neighbor queries. In: Proceedings of the 19th ACM Symposium on Applied Computing (SAC 2004), pp. 1232–1237. ACM (2004)

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

Rényi, A., et al.: On measures of entropy and information. In: Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. The Regents of the University of California (1961)

Roberts, B.W.: Back to the future: personality and assessment and personality development. J. Res. Pers. **43**(2), 137–145 (2009)

Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web (WWW 2001), pp. 285–295. ACM (2001)

Schafer, J.B., Frankowski, D., Herlocker, J., Sen, S.: Collaborative filtering recommender systems. In: The Aaptive Web, pp. 291–324. Springer, Berlin (2007)

Seber, G.A., Lee, A.J.: Linear Regression Analysis, vol. 329. Wiley, New York (2012)

Sha, C., Wu, X., Niu, J.: A framework for recommending relevant and diverse items. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016), pp. 3868–3874 (2016)

Shani, G., Gunawardana, A.: Evaluating recommendation systems. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P. (eds.) Recommender Systems Handbook, pp. 257–297. Springer, Boston (2011)

Shi, Y., Larson, M., Hanjalic, A.: List-wise learning to rank with matrix factorization for collaborative filtering. In: Proceedings of the 4th ACM Conference on Recommender Systems (RecSys 2010), pp. 269–272. ACM (2010)

Shi, Y., Zhao, X., Wang, J., Larson, M., Hanjalic, A.: Adaptive diversification of recommendation results via latent factor portfolio. In: Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012), pp. 175–184. ACM (2012)

Smyth, B., McClave, P.: Similarity vs. diversity. In: Aha, D.W., Watson, I. (eds.) Case-Based Reasoning Research and Development, pp. 347–361. Springer, Berlin (2001)

Srivastava, S., John, O.P., Gosling, S.D., Potter, J.: Development of personality in early and middle adulthood: set like plaster or persistent change? J. Pers. Soc. Psychol. **84**(5), 1041 (2003)

Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artif. Intell. **2009**, 4 (2009)

Thackeray, R., Neiger, B.L., Smith, A.K., Van Wagenen, S.B.: Adoption and use of social media among public health departments. BMC Pub. Health **12**(1), 242 (2012)

Tintarev, N., Dennis, M., Masthoff, J.: Adapting recommendation diversity to openness to experience: a study of human behaviour. In: International Conference on User Modeling, Adaptation, and Personalization (UMAP 2013), pp. 190–202. Springer, Berlin (2013)

Tkalcic, M., Kunaver, M., Tasic, J., Košir, A.: 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, pp. 30–37 (2009)

Tkalcic, M., Quercia, D., Graf, S.: Preface to the special issue on personality in personalized systems. User Model. User-Adap. Interact. **26**(2–3), 103 (2016)

Tobias, I.F., Braunhofer, M., Elahi, M., Ricci, F., Ivan, C.: Alleviating the new user problem in collaborative filtering by exploiting personality information. User Model. User-Adapt. Interact. **26**, 221–255 (2016)

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

Wang, J., Zhu, J.: Portfolio theory of information retrieval. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009), pp. 115–122. ACM (2009)

Willemsen, M.C., Graus, M.P., Knijnenburg, B.P.: Understanding the role of latent feature diversification on choice difficulty and satisfaction. User Model. User-Adap. Interact. **26**(4), 347–389 (2016)

Wood, D., Wortman, J.: Trait means and desirabilities as artifactual and real sources of differential stability of personality traits. J. Pers. **80**(3), 665–701 (2012)

Wu, W., Chen, L.: Implicit acquisition of user personality for augmenting movie recommendations. In: International Conference on User Modeling, Adaptation, and Personalization (UMAP 2015), pp. 302–314. Springer, Berlin (2015)

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

Wu, W., He, L., Yang, J.: Evaluating recommender systems. In: Proceedings of the 7th International Conference on Digital Information Management (ICDIM 2012), pp. 56–61. IEEE (2012)

Zeng, W., Shang, M.S., Zhang, Q.M., Lü, L., Zhou, T.: Can dissimilar users contribute to accuracy and diversity of personalized recommendation? Int. J. Mod. Phys. C **21**(10), 1217–1227 (2010)

Zhang, M., Hurley, N.: Avoiding monotony: improving the diversity of recommendation lists. In: Proceedings of the 2nd ACM Conference on Recommender Systems (RecSys 2008), pp. 123–130. ACM (2008)

Ziegler, C.N., McNee, S.M., Konstan, J.A., Lausen, G.: Improving recommendation lists through topic diversification. In: Proceedings of the 14th International Conference on World Wide Web (WWW 2005), pp. 22–32. ACM (2005)

Wen Wu is a Ph.D. candidate in the Department of Computer Science at Hong Kong Baptist University, under the supervision of Dr. Li Chen. She received her bachelor and master degrees in Computer Science from East China Normal University, China. Her primary interests lie in the areas of user modeling, personality-based recommender systems, and human computer interaction. She has been co-author of several papers that appear in reputable conferences. One of the papers was awarded Best Student Paper in the International Conference on User Modeling, Adaptation and Personalization (UMAP'15).

Li Chen is an Assistant Professor in the Department of Computer Science at Hong Kong Baptist University. Her current research focus is mainly on data-driven Web personalization systems, which integrate researches in artificial intelligence, recommender systems, user modeling, and user behavior analytics for the application in various domains including social media, e-commerce, and online education. She has authored and co-authored over 80 publications, most of which appear in high-impact journals and key conferences in the areas of recommender systems, user modeling, and intelligent user interfaces. She is now an ACM senior member, and an editorial board member of User Modeling and User-Adapted Interaction Journal (UMUAI).

Yu Zhao is a senior machine learning engineer at Douban Inc and working on machine learning and recommender systems. He received the Ph.D. degree in Control Science and Engineering from Tsinghua University, China in 2006. From 2006 to 2011, he was a researcher in NEC Laboratories China, where his research focused on web mining technologies.