A comparative user study on rating vs. personality quiz based preference elicitation methods

Conference Paper · February 2009

DOI: 10.1145/1502650.1502702 · Source: DBLP

2 authors, including:

Pearl Pu
École Polytechnique Fédérale de Lausanne
176 PUBLICATIONS **6,054** CITATIONS

SEE PROFILE

A Comparative User Study on Rating vs. Personality Quiz based Preference Elicitation Methods

Rong Hu
Human Computer Interaction Group
Swiss Federal Institute of Technology (EPFL)
CH-1015, Lausanne, Switzerland
Tel: +41-21-6931327
E-mail: rong.hu@epfl.ch

Pearl Pu
Human Computer Interaction Group
Swiss Federal Institute of Technology (EPFL)
CH-1015, Lausanne, Switzerland
Tel: +41-21-6936081
E-mail: pearl.pu@epfl.ch

ABSTRACT
We conducted a user study evaluating two preference elicitation approaches based on ratings and personality quizzes respectively. Three criteria were used in this comparative study: perceived accuracy, user effort and user loyalty. Results from our study show that the perceived accuracy in two systems is not significantly different. However, users expended significantly less effort, both perceived cognitive effort and actual task time, to complete the preference profile establishing process in the personality quiz-based system than in the rating-based system. Additionally, users expressed stronger intention to reuse the personality quiz-based system and introduce it to their friends. After using these two systems, 53% of users preferred the personality quiz-based system vs. 13% of users preferred the rating-based system, since most users thought the former is easier to use.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces – evaluation/methodology, interaction style; H1.2 [User/Machine Systems]: Software Psychology.

General terms: Design, Experimentation, Human Factors.

Keywords: Preference elicitation, rating-based, personality quiz, recommender systems, user study

INTRODUCTION
Since the recommender was considered as an independent system research field in the mid-1990s, most researchers have been focusing on the recommendation technologies that explicitly rely on the rating's structure, such as the popularly used content-based and collaborative filtering approaches [1]. In these systems, users' ratings are used to build their preference profiles. Although it is a direct way to learn about users' preferences from their ratings, the increase of accuracy follows the increase of users' effort. That is, users need to rate sufficient items to accurately express their interests. Another drawback of rating-based approaches is that they don't take users' psychological characteristics into account.

Recently, some studies have confirmed that there is a strong correlation between human personality and tastes. Kemp revealed a coherent general personality portrait of musicians and preliminarily explored the relationship between personality and music preferences [7]. In the landmark work [12], Rentfrow and Gosling found that personality, self-views, and cognitive abilities could all have roles to play in the formation and maintenance of music preferences. Additionally, we found that several websites have began using personality quizzes to determine users' interest profiles and recommend movies or music based on these profiles, such as Whattorent[1] and Yobo[2]. Our informal survey showed that this elicitation method receives a rather positive acceptance level among users whom we surveyed. Some users even reported that they prefer this method of revealing their tastes, especially when the product domain involves entertainment products such as movies, music, etc. We became interested in assessing whether the personality test can be used as a new preference elicitation mechanism and how its performance is compared with other elicitation processes, like rating-based approaches.

In this paper, we conducted an in-depth within-subjects user study comparing the performance of rating-based and personality quiz-based preference elicitation approaches. We believe that by evaluating both methods, we could potentially improve some specific aspects of the preference elicitation process and bring human personality into recommender applications. After reviewing several currently available recommenders, we chose a commonly used rating-based movie recommender system, MovieLens[3], as a baseline to be compared with a personality quiz-based movie recommender, Whattorent. In this comparative study, we

IUI'09, February 8–11, 2009, Sanibel Island, Florida, USA.

[1] http://whattorent.com/
[2] http://www.yobo.com
[3] http://www.movielens.org

investigated the extent to which users feel the recommendations match their tastes, the cognitive and actual effort they consume in the preference initialization phase and their loyalty to these two systems.

THE TWO MOVIE SYSTEMS

MovieLens.org

MovieLens is a research site run by GroupLens Research in the computer science department at the University of Minnesota. It makes use of collaborative filtering technology to recommend movies. When a new user comes to MovieLens, he/she was required to rate at least 15 movies which he/she has seen before (this was required when we did this experiment). These ratings are used to build a user profile and find the "neighbors" who have the similar tastes as this user. Based on the ratings from these like-minded neighbors, MovieLens predicts how much the user possibly is going to like a movie. Then it presents all recommended movies in a descending order of the predicted scores. Users can refine their preference profiles by rating more movies or changing previous ratings in the future interactions with the system.

Whattorent.com

Whattorent provides the service of movie recommendation in terms of users' personality and mood. This site was built based on the "LaBarrie Theory", which states that "movie viewers emotionally interact with a film in the same manner that they interact with other human beings". According to this theory, the developers of Whattorent designed a personality quiz which includes 20 questions. These questions put users in situations that they have been in before or can easily imagine experiencing, such as,

What would you say if a friend said, "Let's go see a great new independent foreign movie with English subtitles that has some powerful yet difficult to handle scenes"?

The designers hope that by asking about seemingly unimportant life experiences they would get accurate responses from their visitors. By deciphering users' personality and calculating how they would respond to each movie in the database, Whattorent makes movie recommendations which maximize some desired criteria. Movies that have been seen, that have been recommended and that the user did not like are all tracked and are not considered in the future analysis.

EXPERIMENT

In order to understand whether the personality quiz-based preference elicitation process can achieve high subjective opinions from the users, as well as whether it can be an alternate way to help users to build their preference models, we conducted a user study that compared one personality quiz-based system with one rating-based system.

Evaluation Criteria

We first establish the evaluation criteria on which the comparison study was based.

Perceived Accuracy. It has been found that preference matching positively influences users' elaboration on and acceptance of the recommendation [11]. However, since the present study was concerned with the preference elicitation process rather than recommendation algorithms, we adopted the perceived accuracy as our criterion like the definition in [4]. The perceived accuracy measures how users feel the recommendations match their preferences. More concretely, it is measured by both participants' ratings to the recommended movies and their responses to the post-questionnaire on whether they feel the recommendations are tailored to their preferences.

User Effort. The related works have indicated that user effort is also a critical issue needed to be considered, since it can influence the user preference on product alternatives and the perceived satisfaction of recommenders [8, 9]. In our study, we adopted the method used in [3]. User effort is measured by the amount of the subjective effort that users feel they consume (called as perceived cognitive effort) and the actual time they spend in completing the preference initialization process.

User Loyalty. In the e-commerce environments, it is meaningful to appraise whether the system have the ability to convince its users to purchase the recommended product, stimulate them to reuse the system, or persuade them to introduce the system to their friends for increasing the number of users [3, 6, 10]. In our study, we measure all of them as users' loyalty to systems.

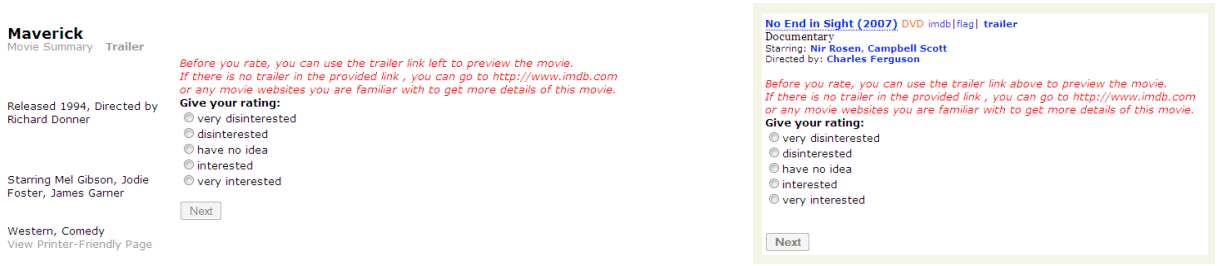
Materials

In our study, we adopted MovieLens as the representative of rating-based systems compared to Whattorent as the representative of personality quiz-based systems. However, their interfaces displaying recommendations are completely different. Whattorent presents one recommendation at a time, while MovieLens lists 15 movies in each page. This difference would affect users' perceived satisfaction with systems [2, 5]. In order to avoid such potential influences, we rearranged their interfaces by utilizing an open-source filtering HTTP proxy (PAW)[4]. Taking advantage of PAW, we modified the HTML responses from MovieLens server to show one recommended movie at a time, simulating Whattorent's original interface (See Figure 1). Additionally, due to the nature of movie products, a clip could provide more useful information which cannot be given by the text description, such as scenes or images of actors. Therefore, we designed a trailer link attached to each recommended movie using PAW, to help users to quickly and easily evaluate the recommended movies. We also appended a button group next to each recommended movie to facilitate users to give ratings on 5-scales ranging from very disinterested to very interested.

Participants

A total of 30 participants (11 females) were recruited in our user study. Every user was given a cash voucher for a

[4] http://paw-project.sourceforge.net



(a) Modified interface of Whattorent

(b) Modified interface of MovieLens

Figure 1: Snapshots of the transformed recommendation interfaces. The interface of MovieLens is modified to simulate Whattorent's original interface. A trailer link and a rating button group are added in both systems.

movie ticket valued at 20 CHF. These participants are from 11 different countries, have different professions (student, research assistant, electronic engineer, sales manager, human resource, store leader, housewife and dentist assistance) and are distributed in four age groups. Table 1 shows some details of their demographic characteristics.

Gender	Male		Female	
	19(63.3%)		11(36.7%)	
Nationality	11 different countries (Germany, France, Spain, England, Switzerland, China, Italy, etc.)			
Age	16-25	26-35	36-45	>45
	18(60%)	9(30%)	2(6.7%)	1(3.3%)
Frequency of watching movies (per month)		never	≤3 times	>3 time
		0	17	13
Frequency of using movie recommender systems (per month)		never	≤3 times	>3 time
		28	1	1

Table 1: Demographic characteristics of participants.

Experiment Design and Procedure

Our user study was conducted in a within-subjects design. All participants used both systems and then were asked which one they preferred. To minimize any carryover effects inherent to the within-subjects design, the total of 30 participants were randomly assigned to two experiments conditions. Each condition has a different order of using recommender systems. That is, 15 users in one condition evaluated the rating-based recommender system (MovieLens) first and then the personality quiz-based system (Whattorent). The other condition has the reverse sequence.

The user study was run at the office of one administrator, who supervised the experiment and helped participants complete all tasks successfully, with the help of a provided desktop computer. An online procedure containing instructions, evaluated systems and questionnaires was implemented so that participants could easily follow, and we could also record all of their responses in log files automatically. In the beginning, participants were first debriefed on the objective of the experiment and the upcoming tasks, and then start evaluating these two systems one by one. The main user tasks in every system are similar:

1. Register a new account in order to avoid any influences from the use history.
2. Rate 15 movies which has been seen before (in the rating-based system) or answer 20 personality questions followed by 2 more questions on users' current mood (in the personality quiz-based system).
3. Rate 6 recommended movies on 5-scales ranging from 1 (very disinterested) to 5 (very interested).
4. Complete an online evaluation questionnaire. Each question is responded on a 5-point Liker scale ranging from 1 (strongly disagree) to 5 (strongly agree).

Finally, participants were asked to fill in a final online post-questionnaire about their preference on these two systems and reasons.

RESULTS AND ANALYSIS

Perceived Accuracy

In our user study, the perceived accuracy was assessed by two questions (listed in Figure 2), which are asked positively and then negatively. In order to measure how well the two questions are related, we calculated the 2-tail Pearson Correlation between them. The result shows they are significantly related (correlation coefficient = -0.806, $p<0.001$). The averages of users' responses are shown in Figure 2. For easily comparing, we reversed the responses from the negative question so that they have the same scale meaning as those from the positive question. It means the higher rating level represents the more the participants satisfied with the recommendations. Although the personality quiz-based system obtained higher scores on the perceived accuracy, the difference between the two methods is not significant by paired t-test (mean = 3.52, SD = 1.09 for the personality quiz-based system vs. mean = 3.22, SD = 1.26 for the rating-based system, $t = 1$, $p = 0.326$).

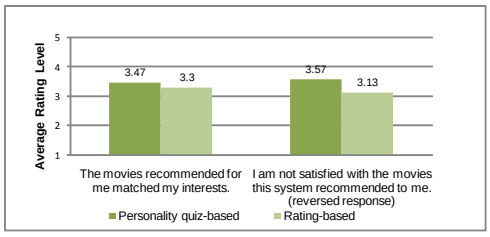


Figure 2: The perceived accuracy. The related evaluation questions are listed below.

The statistical results of the users' ratings on recommended movies are shown in Figure 3. They show that the system using the personality quiz-based approach had more movies rated as interested (18.9% + 35% = 54.9%) compared to the system based on ratings (23.3% + 19.4% = 42.7%). Similarly, the distribution difference between the two systems is not significant by chi-test (chi square = 7.274, p = 0.122).

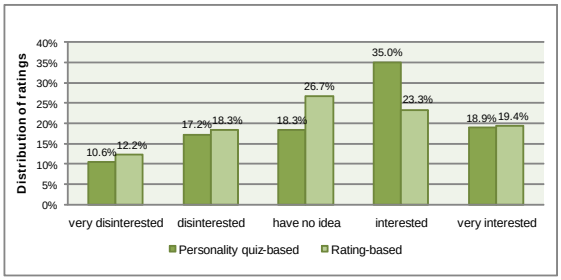


Figure 3: The distribution of the ratings of the recommended movies.

User Effort

To compare users' perceived cognitive effort, we also asked two interrelated questions (correlation coefficient = -0.55, $p < 0.001$, in 2-tail Pearson Correlation). The questions and statistics of user responses are presented in Figure 4. Similarly, we reversed the responses from the negative question. The high rating level means it is easy to use. The results show that the overall cognitive effort is perceived significantly lower in the personality quiz-based system than rating-based system (mean = 3.63, SD = 1.00 for personality quiz- based process vs. mean = 2.4, SD = 1.26 for rating-based process, $p < 0.001$).

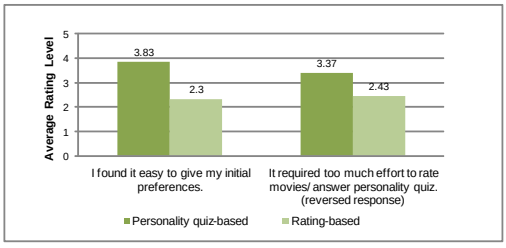


Figure 4: Perceived cognitive effort. The related evaluation questions are listed below.

Additionally, we recorded the actual completion time which is defined as the amount of time one user spend from the beginning of rating movies or answering the personality quiz to obtaining the first recommendation. When using the personality quiz-based system, participants spent 6.8 minutes (SD = 2.11) on average in the preference initialization phase while they spent highly significantly more time (18.7 minutes on average, SD = 6.32) with the rating-based system ($t = -9.46$, $p < 0.001$). Furthermore, 90% (27 out of 30) of the participants went through over 10 pages (15 movies in each page)when giving ratings on the rating-based system. And there was one user who didn't finish the rating task, even navigating directly to the last rating page. Movielens provides 2000 movies in total for users to rate. That is, the user flicked through 2000 movies but still could not complete the sign-up process. In order to continue the experiment, this user was asked to go back to the previous rating pages and rate up to 15 movies depending on the information provided by imdb.com. We removed this outlier when calculating the average completion time.

User Loyalty

Figure 5 shows the results on users' loyalty to the two systems. As we can see, participants expressed a slightly higher intention to rent or buy the DVDs of the recommended movies in the personality quiz-based system than the rating-based system ($t = 0.108$, $p = 0.915$). However, they stated a significant stronger intention to reuse the personality quiz-based system in the future compared to the rating-based system ($t = 2.163$, $p < 0.05$). Likewise, more participants would like to introduce the personality quiz-based system to their friends, even though the difference is moderately significant ($t = 2.02$, $p = 0.052 < 0.1$). In particular, no one strongly agreed with the assertion on the rating-based system (That is, there is no "5" in their responses).

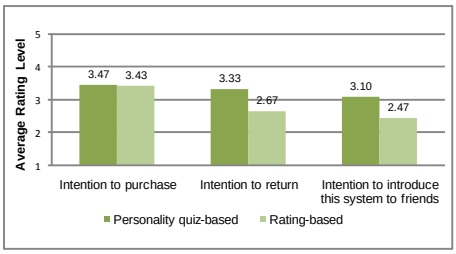


Figure 5: User loyalty to the two recommenders.

Preference of Elicitation Methods

In the end, all participants were asked to fill in a final questionnaire about their preference on these two systems. The result shows that 53.33% of users (16 out of 30) preferred the personality quiz-based system, while 13.33% of users (4 out of 30) preferred the rating-based system. This phenomenon exhibits a significant distribution ($p < 0.001$ with a chi-test). Furthermore, we measured the reasons why they have such preference along two dimensions: ease-to-use and accuracy. From participants' responses, we can conclude that they preferred the personality quiz-based system mainly because it costs them less effort to obtain recommendations, while the high recommendation quality is the main reason for the participants who preferred the other system. Besides, 23.33% of participants preferred the personality quiz-based system because of both ease-to-use and accurate recommendations, while 3.33% of participants preferred the rating-based system due to the both reasons.

Discussion

So far, the personality quiz-based system performs significant better in all evaluation criteria except the perceived accuracy. However, this helped us establish a baseline for the comparison of other factors. That is, given the same perceived accuracy, users experienced less effort, felt more inclined to recommend the system to their friends, and more likely to reuse it. The response to the final post-questionnaire shows that most participants (53.33%) sub-

jectively preferred the personality quiz-based system to the rating-based system (13.33%). The positive arguments are that the situations assumed in the personality quiz are those which have been experienced before or can be easily imagined experiencing. Participants therefore did not feel this task requires extra memory effort.

According to participants' comments, we obtained several reasons why most participants felt they spend more effort on the rating-based system. Firstly, it seems difficult for people to keep a movie (name or plots) in their memories for a long time unless the movie is special for them. Participants felt hard to give an accurate evaluation to a movie, which resulted in a large amount of average ratings (3 stars, "It is Ok"). Secondly, most movies are translated into different names when they released in different languages. Some participants had to expend great effort to recognize whether the movie is the one they have seen before, especially for the participants from non-English speaking countries. (MovieLens is an American website.)

Based on the similar perceived recommendation accuracy, more participants chose to reuse the personality quiz-based system and recommend it to their friends since they consumed less time and mental energy there. However, the difference in the intention to purchase is not significant. The reason might be that quality is a more important factor users would consider when they plan to buy a product. Some related research has indicated that when users put forth more effort into a system, they would expect to obtain more benefits from the system and be more likely to be loyal users [10]. We think the reason why our results are different from theirs is that in our case the participants felt that the recommendation qualities of both systems are similar, they therefore would express more loyal to the one which requires less time and energy.

Moreover, after interviewing the participants who favored the rating-based system, we can also see some advantages of it. They thought they could have more control on the recommendations by directly changing their rating profiles. In addition, the transparency of its recommendation mechanism allows the participants to have more confidence on the recommendations, while some participants thought there is a magic in the personality quiz-based system.

CONCLUSION AND FUTURE WORK

We conducted a within-subject user study to evaluate the performance of two preference elicitation methods: the personality quiz-based and the commonly used rating-based approaches, in terms of perceived accuracy, user effort and user loyalty. The positive results show that personality quiz-based preference elicitation technology has a high potential to be a powerful tool as an alternative or supplement to the existing preference elicitation methods.

Since personality reflects people essential characteristics, we believe that personality-based preference elicitation approaches can also assist users to reveal the hidden preferences and thereby increase the recommendation accuracy. However, since the two systems used in our study depend on different databases, we cannot conclude that the personality quiz-based preference elicitation methods better understand users' tastes. This important issue will be investigated in the follow-up works.

ACKNOWLEDGMENTS

We thank the EPFL & Chinese Government for sponsoring the reported research work. We are grateful to the participants of our user studies for their patience and time.

REFERENCES

1. Adomavicius, G. Toward the Next Generation of Recommender Systems: A Survey of The State-of-the-Art and Possible Extensions. *IEEE Trans. Knowledge and Data Eng.*, 17, 6(2005), 734-749.
2. Bilgic, M. and Mooney, R.J. Explaining Recommendations: Satisfaction vs. Promotion. *Beyond Personalization Workshop*, IUI, 2005.
3. Chen, L. and Pu, P. Hybrid Critiquing-based Recommender Systems. In *Proc. of International Conference on Intelligent User Interfaces (IUI2007)*, 2007, 22-31.
4. Gretzel, U. and Fesenmaier, D.R. Persuasion in Recommender Systems. *International Journal of Electronic Commerce*, 11, 2(2006): 81-100.
5. Häubl, G. and Trifts, V. Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science*, 19, 1(2000), 4-21.
6. Jone, N. and Pu, P. User Technology Adoption Issues in Recommender Systems. In *Proc. of Networking and Electronic Commerce Research Conference (NAEC2007)*, 2007.
7. Kemp, A.E. The Music Temperament: Psychology and Personality of Music1ians. Oxford University Press, New York, 1996.
8. Kivetz, R. and Dimonson, I. Earning the right to indulge: Effort as a determination of customer preferences toward frequency program rewards. *Journal of Marketing Research*, 39(2002), 155-170.
9. Kruger J., Wirtz D., Van Boven L. and Altermatt T.W. The effort heuristic. *Journal of Experimental Social Psychology*, 40, 1(2004), 91-98.
10. McNee, S.M., Lam, S.K., Konstan, J.A., Riedal, J. Interfaces for eliciting new user preferences in recommender systems. *User Modeling*, Johnstown, PA, USA, Springer Verlag(2003) 178-187.
11. Tam, K.Y. and Ho, S.Y. Web personalization as a persuasion strategy: An elaboration likelihood model perspective. *Information Systems Research*, 16, 3(2005), 271-291.
12. Rentfrow, P.J. and Gosling, S.D. The do re mi's of everyday life: The Structure and Personality Correlates of Music Preferences. *Journal of Personality and Social Psychology* 84, 6(2003), 1236-1256.