FULL PAPER

Proceedings of the ACM RecSys 2010 Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces (UCERSTI),
Barcelona, Spain, Sep 30, 2010
Published by CEUR-WS.org, ISSN 1613-0073, online ceur-ws.org/Vol-612/paper3.pdf

A User-Centric Evaluation Framework of Recommender Systems

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

Li Chen
Department of Computer Science
Hong Kong Baptist University
224 Waterloo Road, Hong Kong
Tel: +852-34117090
lichen@comp.hkbu.edu.hk

ABSTRACT

User experience research is increasingly attracting researchers' attention in the recommender system community. Existing works in this area have suggested a set of criteria detailing the characteristics that constitute an effective and satisfying recommender system from the user's point of view. To combine these criteria into a more comprehensive framework which can be used to evaluate the perceived qualities of recommender systems, we have developed a model called ***ResQue*** (***Re***commender ***s***ystems' ***Q***uality of ***u***ser ***e***xperience). ResQue consists of 13 constructs and a total of 60 question items, and it aims to assess the perceived qualities of recommenders such as their usability, usefulness, interface and interaction qualities, users' satisfaction of the systems, and the influence of these qualities on users' behavioral intentions, including their intention to purchase the products recommended to them, return to the system in the future, and tell their friend about the system. This model thus identifies the essential qualities of an effective and satisfying recommender system and the essential determinants that motivate users to adopt this technology. The related questionnaire can be further adapted for a custom-made user evaluation or combined with objective performance measures. We also propose a simplified version of the model with 15 questions which can be employed as a usability questionnaire for recommender systems.

Categories and Subject Descriptors

H1.2 [**User/Machine Systems**]: *Human factors;* H5.2 [**User Interfaces**]: *evaluation/methodology, user-centered design.*

General Terms

Measurement, Experimentation, Human Factors.

Keywords

Quality measurement, usability evaluation, recommender systems, quality of user experience, e-Commerce recommender, post-study questionnaire, evaluation of decision support.

1. INTRODUCTION

A recommender system is a web technology that proactively suggests items of interest to users based on their objective behavior or their explicitly stated preferences. It is no longer a fanciful website add-on, but a necessary component. According to the 2007 ChoiceStream survey,[1] 45% of users are more likely to shop at a website that employs recommender technology. Furthermore, a higher percentage (69%) of users in the highest spending category are more likely to desire the support of recommendation technology.

Characterizing and evaluating the quality of user experience and users' subjective attitudes toward the acceptance of recommender technology is an important issue which merits attention from researchers and practitioners in both web technology and human factor fields. This is because recommender technology is becoming widely accepted as an important component that provides both user benefits and enhances the website's revenue. For users, the benefits include more efficiency in finding preferential items, more confidence in making a purchase decision, and a potential chance to discover something new. For the marketer, this technology can significantly enhance user likelihood to buy the items recommended to them, their overall satisfaction and loyalty, increasing users' likelihood to return to the site and recommend the site to their friends. Thus, evaluating user's perception of a recommender system can help developers and marketers understand more precisely if users actually experience and appreciate the intended benefits. This will, in turn, help improve the various aspects of the system and more accurately predict the adoption of a particular recommender.

So far, previous research work on recommender system evaluation has mainly focused on algorithm accuracy [9,1], especially objective prediction accuracy [25,26]. More recently, researchers began examining issues related to users' subjective opinions [30, 13] and developing additional criteria to evaluate recommender systems [18, 33]. In particular, they suggest that user satisfaction does not always correlate with high recommender accuracy. Increasingly, researchers are investigating user experience issues such as identifying determinants that influence users' perception of recommender systems [30], effective preference elicitation methods [19], techniques that motivate users to rate items that they have experienced [2], methods that generate diverse and more satisfying recommendation lists [43], explanation interfaces [31], trust formation with recommenders [6], and design guidelines for enhancing a recommender's interface layout [22]. However, the field lacks a general definition and evaluation framework of what constitutes an effective and satisfying recommender system from the user's perspective.

UCERSTI Workshop of RecSys'10, Sept. 26-30, 2010, Barcelona, Spain.

[1] 2007 ChoiceStream Personalization Survey, ChoiceStream, Inc.

Proceedings of the ACM RecSys 2010 Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces (UCERSTI),
Barcelona, Spain, Sep 30, 2010
Published by CEUR-WS.org, ISSN 1613-0073, online ceur-ws.org/Vol-612/paper3.pdf

Our present work aims to review existing usability-oriented evaluation research in the field of recommender systems to identify essential determinants that motivate users to adopt this technology. We then apply well-known usability evaluation models, including TAM [7] and SUMI [15], in order to develop a more balanced framework. The final model, which we call ResQue, consists of 13 constructs and a total of 60 question items categorized into four main dimensions: the perceived system qualities, users' beliefs as a result of these qualities, their subjective attitudes, and their behavioral intentions. The structure and criteria of our framework is derived on the basis of three essential characteristics of recommender systems: 1) being an interaction-driven application and a critical part of online e-commerce services, 2) providing information filtering technology and suggesting recommended items, and 3) providing decision support technology for the users.

The main contribution of this paper is the development of a well-balanced evaluation framework for measuring the perceived qualities of a recommender and predicting users' behavioral intentions as a result of these qualities. Thus, it is a forecasting model that helps us understand users' motivation in adopting a certain recommender. Secondly, the framework aims to help designers and researchers easily perform a usability and user acceptance test during any stage of the design and deployment phase of a recommender. These usability tests can be performed either on a stand-alone basis or as a post-study questionnaire. The model can be further combined with measurements that address other perceived qualities of a recommender, such as security and robustness issues. For those who are interested in a quick usability evaluation, we also propose a simplified version of the model with 15 questions.

2. EVALUATION WORK FROM USERS' POINT OF VIEW

Swearingen and Sinha [38] conducted a user study on eleven recommender systems in order to understand and discover influential factors, other than algorithm accuracy, that affect users' perception. The main results are that transparent system logic, recommendation of familiar items, and sufficient supporting information to recommended items is crucial in influencing users' favorable perception towards the system. They also highlighted that trust and willingness to purchase should be noted. In addition, the users' appreciation of online recommendations is compared with that of recommendations from their friends, defining the notion of relative accuracy.

McNee et al. [20] pointed out that accuracy metrics alone and the commonly employed leave-one-out procedure was very limited in evaluating recommender systems. User satisfaction does not always correlate with high recommender accuracy. Metrics are needed to determine good and useful recommendations, such as the serendipity, salience, and diversity of the recommended items.

Tintarev and Masthoff provided a comprehensive survey of the explanation functionality used in ten academic and eight commercial recommenders [31]. They derived seven main aims of the explanation facility which can help a recommender significantly enhance users' satisfaction: transparency (explains why recommendations were generated), scrutability (the ability for the user to critique the system), trust (increase users' confidence in the system), effectiveness (help users make good decisions), persuasiveness (convince users to try or buy items recommended to them), efficiency (help users make decisions faster) and satisfaction (increase the ease of use and enjoyment). These aims are very similar to the set of criteria that we have developed in ResQue, except the fact that we focus more on the system as a whole rather than just the explanation component.

Ozok et al. [22] explored recommender systems' usability and user preferences from both the structural (how recommender systems should look) and content (what information recommender systems should contain) perspectives. A two-layer interface usability evaluation model including both micro- and macro-level interface evaluations was proposed, followed by a Survey on Usability of E-Commerce Recommender Systems (SUERS). The survey was administered on 131 college-aged online shoppers to measure and rank the importance of structural and content aspects of recommender systems from the shoppers' perspectives. The main result was a set of 14 design guidelines. The micro-level of the guidelines provided suggestions specific to the recommended product such as what attributes (name, price, image, description, rating, etc.) to include in the interface. The macro-level of the guidelines provided suggestions concerning when, where and how the recommended products should be displayed. The development process of the model was limited, as authors did not go through an iterative process of the evaluation and refinement of the model. Instead, it was purely based on a literature survey of quite limited past work of subjective evaluations of recommender system. Most importantly, it failed to explain how usability issues influence users' behavioral intentions such as their intention to buy the items recommended to them, whether they will continue using the system and recommend the system to their friends.

Jones and Pu [13] presented the first significant user study that aimed to understand users' *initial* adoption of the recommender technology and their subjective perceptions of the system. Study results show that a simple interface design, a small amount of initial effort required by the system to get to know the users, the perceived qualities such as the subjective accuracy, novelty and enjoyability of the recommended items are the key design factors that significantly enhance the website's ability to attract users.

3. MODEL DEVELOPMENT

A measurement model consists of a set of constructs, the participating questions for each construct, the scale's dimensions, and a procedure for conducting the questionnaire. Psychometric questionnaires such as the one proposed in this paper require the validation of the questions used, data gathering, and statistical analysis before they can be used with confidence. The current model and its constructs were based on our past work in investigating various interface and interaction issues between users and recommenders. In over 10 user studies, we have carefully and progressively developed and employed user satisfaction questionnaires to evaluate recommenders' perceived qualities such as ease of use, perceived usefulness and users' satisfaction and behavioral intentions [4,5,6,12,13,14,23,24]. This past research has given us a unique opportunity to synthesize and organize the accumulation of existing questionnaires and develop a well-balanced framework.

In the model development process, we also compare our constructs with those used in TAM and SUMI, two well-known and widely adopted measurement frameworks.

TAM (Technology Acceptance Model) seeks to understand a set of perceived qualities of a system and users' intention to adopt the system as a result of these qualities, thus explaining not only the desirable outcome of a system but also users' motivation. The original TAM listed three constructs: perceived ease of use of a system, its perceived usefulness and users' intention to use the system. However, TAM was also criticized for its over-simplicity and generality. Venkatesh et al. [32] formulated an updated version of TAM, called the Unified Theory of Acceptance and Use of Technology. In this more recent theory, four key constructs (performance expectancy, effort expectancy, social influence, and facilitating conditions) were presented as direct determinants of usage intentions and behaviors.

SUMI (Software Usability Measurement Inventory) is a psychometric evaluation model developed by Kirakowski and Corbett [15] to measure the quality of software from the end-user's point of view. The model consists of 5 constructs (efficiency, affect, helpfulness, control, learnability) and 50 questions. It is widely used to help designers and developers assess the quality of use of a software product or prototype and can assist with the detection of usability flaws and the comparison between software products.

By adapting our past work to the TAM and SUMI models, we have identified 4 essential constructs of ResQue for a successful recommender system to fulfill from the users' point of view: 1) user perceived qualities of the system, 2) user beliefs as a result of these qualities in terms of ease of use, usefulness and control, 3) their subjective attitudes, and 4) their behavioral intentions. Figure 1 depicts the detailed schema of the constructs of ResQue and some of the scales for each construct.

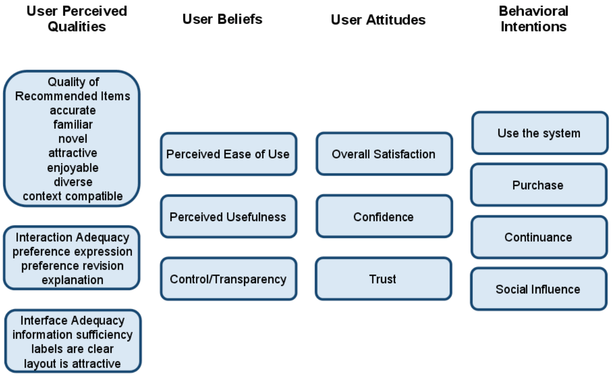

Figure 1: Constructs of an Evaluation Framework on the Perceived Qualities of Recommenders (ResQue).

When administering the questionnaires, we assume that a recommender system being evaluated is part of an online system. To make the evaluation more focused on the recommender component, we often give subjects a specific task: "*find an ideal product to buy/experience from an online site*" where the recommender in question is a constituent component.

In the following sections, the meaning of each scale as well as its subscales is defined and explained, and the sample questions that can be used in a questionnaire are suggested in the appendix at the end of the paper. It is a common practice in questionnaire development to vary the tone of items to control potential response biases. Typically some of the items elicit agreement and others elicit disagreement. For some of the items, therefore, we also suggest reverse scale questions. A 5-point Likert scale from "strongly disagree" (1) to "strongly agree" (5) is recommended to characterize users' responses.

3.1 Perceived System Qualities

This construct refers to the functional and informational aspect of a recommender and how the perceived qualities of these aspects influence users' beliefs on the ease of use, usefulness and control/transparency of a system. A recommender system is not simply part of a website, but more importantly a decision support tool. We focus on three essential dimensions: the quality of the recommended items, the interaction adequacy and the interface adequacy as the recommender helps users reach a purchase decision.

3.1.1 Quality of Recommended Items

The items proposed by a recommender can be considered one of the main features of the system. Qualities refer to the information quality and genuine usefulness of the suggested items. Presented as a collection of articles, the recommended items are often labeled and presented in a certain area of the recommender page. Some systems also propose grouping them into meaningful subareas to increase users' comprehension of the list and enable them to more effectively reach decisions [4]. In our earlier work, we have found strong correlations of the following qualities of the recommended items to users' intention to use the system.

Perceived accuracy is the degree to which users feel the recommendations match their interests and preferences. It is an overall assessment of how well the recommender has understood the users' preferences and tastes. This subjective measure is significantly easier to obtain than the measure of objective accuracy that we used in our earlier work [23]. Our studies show that they are strongly correlated [6]. In other words, if users respond well to this question, it is likely that the underlying algorithm is accurate in predicting users' interest. In addition, it is useful to use **relative accuracy** to compare the difference between recommendations a user may get from a system versus those from friends [28]. It can serve as a useful complement to perceived accuracy because it implicitly sets up friends' recommendation quality as a baseline.

Familiarity describes whether or not users have previous knowledge of, or experience with, the items recommended to them. Swearingen and Sinha [30] indicated that users like and prefer to get recommendations of previously experienced items because their presence reinforces trust in the recommender system. However, users can be frustrated by too much familiarity. Therefore, it is important to know whether or not a recommender website has achieved the proper balance of familiarity and novelty from the users' perspective.

Novelty (or discovery) is the extent to which users receive new and interesting recommendations. The core concept of novelty is related to the recommender's ability to educate users and help them discover new items [24]. In [20], a similar concept, called "serendipity", was suggested. Herlocker [11] argued that novelty is different from serendipity, because novelty only covers the concept of "new" while serendipity means not only "new" but also "surprising". However, in conducting the actual user evaluation procedure, the meticulous distinction between these two words will cause confusion for users. Therefore, we suggest novelty and discovery as two similar questions. More user trials will be needed to further delineate the serendipity question.

The **Attractiveness** of the recommended items refers to whether or not recommended items are capable of stimulating users' imagination and evoking a positive emotion of interest or desire.

Attractiveness is different from accuracy and novelty. An item can be accurate and novel, but not necessarily attractive; a novel item is different from anything a user has ever experienced, whereas an attractive item stimulates the user in a positive manner. This concept is similar to the salience factor in [20].

While judging novelty requires a user to think more about the distinguishing factors of an item, the aspect of attractiveness brings to mind the outstanding quality of an item and has a more emotional tone to it.

The **enjoyability** of recommended items refers to whether users have enjoyed experiencing the items suggested to them. It was found to have a significant correlation to users' intention to use and return to the system [13]. This is the only scale that assesses a user's actual experience of a recommender. In many online study scenarios, it is not possible to immediately measure enjoyability unless users are told to answer a questionnaire after a few weeks when they have actually received and experienced the item. In testing music or film recommenders, it is possible to allow users to answer this question if they are given the opportunity to listen to a song excerpt or watch a movie trailer.

Diversity measures the diversity level of items in the recommendation list. As the recommendation list is the first piece of information users will encounter before they examine the details of an individual recommendation, users' impression of this list is important for their perception of the whole system. At this stage, it has been found that a low diversity level might disappoint users and could cause them to leave this recommender [13]. McGinty and Smyth [17] proposed integrating diversity with similarity in order to adaptively select the appropriate strategy (either similar or diverse ones) given each individual user's past behavior and current needs. Literature also suggests that a recommendation list as a complete entity should be judged for its diversity rather than treating each recommendation as an isolated item [33].

Context compatibility evaluates whether or not the recommendations consider general or personal context requirements. For example, for a movie recommender, the necessary context information may include a user's current mood, different occasions for watching the movie, whether or not other people will be present, and whether the recommendation is timely. A good recommender system should be able to formulate recommendations considering different kinds of contextual factors that will likely take effect.

3.1.2 Interaction Adequacy

Besides issues related to the quality of recommended items, the system's ability to present recommendations, to allow for user feedback and to explain the reasons why recommendations facilitate purchasing decisions also weighs highly on users' overall perception of a recommender. Thus, three main interaction mechanisms are usually suggested in various recommenders: initial preference elicitation, preference revision, and the system's ability to explain its results. Behavioral based recommenders do not require users to explicitly indicate their preferences, but collect such information via users' browsing and purchasing history. For rating and preference based recommenders, this process requires a user to rate a set of items or state their preferences on desired items in a graphical user interface [23]. Some conversational recommenders provide explicit mechanisms for users to provide feedback in the form of critiques [6]. The simplest critiques indicate whether the recommended item is good or bad, while the more sophisticated ones show users a set of alternative items that take into account users' desire for these items and the potential superior values they offer, such as better price, more popularity, etc [6].

The final interaction quality being measured is the system's ability to explain the recommended results. Herlocker et al. [10], Sinha and Swearingen [30] and Tintarev and Masthoff [31] demonstrated that a good explanation interface could help inspire users' trust and satisfaction by giving them information to personally justify recommendations, increasing user involvement and educating users on the internal logic of the system [10, 31]. In addition, Tintarev and Masthoff [31] defined in detail possible aims of explanation facilities: transparency, scrutability, trust, effectiveness, persuasiveness, efficiency, and satisfaction. Pu and Chen extensively investigated design guidelines for developing explanation-based recommender interfaces [4]. They found that organization interfaces are particularly effective in promoting users' satisfaction of the system, convincing them to buy items recommended to them, and bringing them back to the store in the future.

3.1.3 Interface Adequacy

Interface design issues related to recommenders have also been extensively investigated in [10, 20, 31,22]. Most of the existing work is concerned with how to optimize the recommender page layout to achieve the maximum visibility of the recommendation, i.e. whether to use image, text, or a combination of the two. A detailed set of design guidelines were investigated and proposed [22]. In our current model, we mainly emphasize users' subjective evaluations of a recommender interface in terms of its information sufficiency, the interface label and layout adequacy and clarity.

3.2 Beliefs

3.2.1 Perceived Ease of Use

Perceived ease of use, also known as efficiency in SUMI and perceived cognitive effort in our existing work [6,14], measures users' ability to quickly and correctly accomplish tasks with ease and without frustration. We also use it to refer to decision efficiency, i.e. the extent to which a recommender system facilitates users to find their preferential items quickly. Although task completion and learning time can be measured objectively, it can be difficult to distinguish the actual task completion time from the measured task time for various reasons. Users can be exploring the website and discovering information unrelated to the assigned task. This is especially true if a system is entertaining and educational, and its interface and content is very appealing. It is also possible that the user perceives that he/she has consumed less time while the measured task completion time is in fact high. Therefore, evaluating perceived ease of use may be more appropriate than using the objective task completion time to measure a system's ease of use.

Besides the overall perceived ease of use, **perceived initial effort** should also be taken into account, given the new user problem. Perceived initial effort is the perceived effort users contribute to the system before they get the first set of recommendations. The initial effort could be spent on rating items [19], specifying preferences, or answering personality quizzes [12]. Theoretically speaking, recommender systems should try to minimize the effort users expend for a good recommendation [30].

Proceedings of the ACM RecSys 2010 Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces (UCERSTI),
Barcelona, Spain, Sep 30, 2010
Published by CEUR-WS.org, ISSN 1613-0073, online ceur-ws.org/Vol-612/paper3.pdf

Easy to learn, known as "learnability" in SUMI, initially appears to be an inadequate dimension since most recommenders require a minimal amount of learning by design. However, since some users may not initially notice the recommended items or know exactly what they were intended for, especially without clear labels or explicit explanations on the interface, the learning aspect should be included to measure the level of ease for users to discover the recommended items. In addition, some recommenders, such as critiquing-based recommenders, do allow users to provide feedback to increase the personalization of the recommender. In this case, the learning construct measures how easy it is for users to alter their personal profile information in order to receive different recommendations.

3.2.2 Perceived Usefulness

Perceived usefulness of a recommender (called perceived competence in our previous work) is the extent to which a user finds that using a recommender system would improve his/her performance, compared with their previous experiences without the help of a recommender [4]. This element requests users' opinion as to whether or not this system is useful to them. Since recommenders used in e-commerce environments mainly assist users in finding relevant information to support their purchase decision, we further qualify the usefulness in two aspects: decision support and decision quality.

Recommender technology provides decision support to users in the process of selecting preferential items, for example making a purchase in an e-commerce environment. The objective of decision technologies in general is to overcome the limit of users' bounded rationality and to help them make more satisfying decisions with a minimal amount of effort [29]. Recommender systems specifically help users manage an overwhelming flood of information and make high-quality decisions under limited time and knowledge constraints. **Decision support** thus measures the extent to which users feel assisted by the recommended system.

In addition to the efficiency of decision making, the quality of the decision (**decision quality)** also matters. The quality of a system-facilitated decision can be assessed by confidence criterion, which is the level of a user's certainty in believing that he/she has made a correct choice with the assistance of a recommender.

3.2.3 Control and Transparency

User control measures whether users felt in control in their interaction with the recommender. The concept of user control includes the system's ability to allow users to revise their preferences, to customize received recommendations, and to request a new set of recommendations. This aspect weighs heavily in the overall user experience of the system. If the system does not provide a mechanism for a user to reject recommendations that he/she dislikes, a user will be unable to stop the system from continuously recommending items which might cause him/her to be disappointed with the system.

Transparency determines whether or not a system allows users to understand its inner logic, i.e. why a particular item is recommended to them. A recommender system can convey its inner logic to the user via an explanation interface [4,10,30,31]. To date, many researchers have emphasized that transparency has a certain impact on other critical aspects of users' perception. Swearingen and Sinha [30] showed that the more transparent a recommended product is, the more likely users would be to purchase it. In addition, Simonson [27] suggested that perceived accuracy of a recommendation is dependent on whether or not the user sees a correspondence between the preferences expressed in the measurement process and the recommendation presented by the system.

3.3 Attitudes

Attitude is a user's overall feeling towards a recommender, which is most likely derived from his/her experience as she interacts with a recommender. An attitude is generally believed to be more long-lasting than a belief. Users' attitudes towards a recommender are highly influential on their subsequent behavioral intentions. Many researchers attribute positive attitudes, including users' satisfaction and trust of a recommender, as important factors.

Evaluating **overall satisfaction** determines what users think and feel while using a recommender system. It gives users an opportunity to express their preferences and opinions about a system in a direct way. **Confidence inspiring** refers to the recommender's ability to inspire confidence in users, or its ability to convince users of the information or products recommended to them. **Trust** indicates whether or not users find the whole system trustworthy. Studies show that consumer trust is positively associated with their intentions to transact, purchase a product, and return to the website [8]. The trust level is determined by the reputation of online systems [8], as well as the recommender system's ability to formulate good recommendations and provide useful explanation interfaces [4,10,19]. However, as trust is a long-term relationship between a user and an online system, it is sometimes difficult to measure trust purely after a short-period interaction with a system. Thus, we recommend observing the trust formation over time, as users are incrementally exposed to the same recommender.

3.4 Behavioral Intentions

Behavioral intentions towards a system is related to whether or not the system is able to influence users' decision to use the system and purchase some of the recommended results.

One of the fundamental goals for an e-commerce website is to maximize user loyalty and the lifetime value to stimulate users' future visits and purchases. User loyalty evaluates the system's ability to convince users to reuse the system, or persuade them to introduce the system to their friends in order to increase the number of clients. Accordingly, this dimension consists of the following criteria: user agreement to use the system, user acceptance of the recommended items (resulting in a purchase), user retention and **intention to introduce this system to her/his friends**. By using a questionnaire, the user's **intention to return** can be measured as a satisfactory approximation of actual user retention, because the Theory of Planned Behavior [32] states that behavioral intention can be a strong predictor of actual behavior. Although the website's integrity, reputation and price quality will also likely impact user loyalty, the most important factor for a recommender system is to help users effectively find a satisfying product, i.e. the quality of its recommendations [7].

4. SIMPLIFIED MODEL

In the previous sections, we described the ***development process*** of a subjective evaluation framework to measure users' perceived qualities of a recommender as well as users' behavioral intentions such as their intention to buy or use the items suggested to them, continue to use the system, and tell their friends about the recommender. We described both the constructs and corresponding sample questions (see Appendix A for a summary).

Our overall motivation for this research was to understand the crucial factors that influence the user adoption of recommenders. Another motivation is to come up with a subjective evaluation questionnaire that other researchers and practitioners can employ. However, it is unlikely that a 60-item questionnaire can be administered for a quick and easy evaluation. This has motivated us in proposing a simplified model based on our past research. Between 2005 and 2010, we have administered 11 subjective questionnaires on a total of 807 subjects [4,5,6,12,13,14,23,24]. Initial questionnaires covered some of the four categories identified in the ResQue. As we conducted more experiments, we became more convinced of the four categories and used all of them in recent studies. On average, between 12 and 15 questions were used. Based this previous work, we have synthesized and organized a total of 15 questions as a simplified model for the purpose of performing a quick and easy usability and adoption evaluation of a recommender (see questions with * sign).

5. CONCLUSION AND FUTURE WORK

User evaluation of recommender systems is a crucial subject of study that requires a deep understanding, development and testing of the right dimensions (or constructs) and the standardization of the questions used. The framework described in this paper presents the first attempt to develop a complete and balanced evaluation framework that measures users' subjective attitudes based on their experience towards a recommender.

ResQue consists of a set of 13 constructs and 60 questions for a high-quality recommender system from the user point of view and can be used as a standard guideline for a user evaluation. It can also be adapted to a custom-made user evaluation by tailoring it in an individual research context. Researchers and practitioners can use these questionnaires with ease to measure users' general satisfaction with recommenders, their readiness to adopt the technology, and their intention to purchase recommended items and return to the site in the future.

After ResQue was finalized, we asked several expert researchers in the community of recommender systems to review the model. Their feedback and comments were then incorporated into the final version of the model. This method, known as the Delphi method, is one of the first validation attempts on the model. Since the work was submitted, we have started conducting a survey to further validate the model's reliability, validity and sensitivity using factor analysis, structural equation modeling (SEM), and other techniques described in [21]. Initial results based on 150 participants indicate how the model can be interpreted and show factors that correspond to the original model. At the same time, analysis also gives some indications on how to refine the model. More users are expected to participate in the survey and the final outcome will be soon reported.

APPENDIX

A. Constructs and Questions of ResQue

The following contains the questionnaire statements that can be used in a survey. They are developed based on the ResQue model described in this paper. Users should be asked to indicate their answers to each of the questions using the 1-5 Likert scales, where 1 indicates "strongly disagree" and 5 is "strongly agree."

A1. Quality of Recommended Items

A.1.1 Accuracy

- The items recommended to me matched my interests.*
- The recommender gave me good suggestions.
- I am not interested in the items recommended to me (reverse scale).

A.1.2 Relative Accuracy

- The recommendation I received better fits my interests than what I may receive from a friend.
- A recommendation from my friends better suits my interests than the recommendation from this system (reverse scale).

A.1.3 Familiarity

- Some of the recommended items are familiar to me.
- I am not familiar with the items that were recommended to me (reverse scale).

A.1.4 Attractiveness

- The items recommended to me are attractive.

A.1.5 Enjoyability

- I enjoyed the items recommended to me.

A.1.6 Novelty

- The items recommended to me are novel and interesting.*
- The recommender system is educational.
- The recommender system helps me discover new products.
- I could not find new items through the recommender (reverse scale).

A.1.6 Diversity

- The items recommended to me are diverse.*
- The items recommended to me are similar to each other (reverse scale).*

A.1.7 Context Compatibility

- I was only provided with general recommendations.
- The items recommended to me took my personal context requirements into consideration.
- The recommendations are timely**.**

A2. Interaction Adequacy

- The recommender provides an adequate way for me to express my preferences.
- The recommender provides an adequate way for me to revise my preferences.
- The recommender explains why the products are recommended to me.*

A3. Interface Adequacy

- The recommender's interface provides sufficient information.
- The information provided for the recommended items is sufficient for me.
- The labels of the recommender interface are clear and adequate.
- The layout of the recommender interface is attractive and adequate.*

A4. Perceived Ease of Use

A.4.1 Ease of Initial Learning

- I became familiar with the recommender system very quickly.
- I easily found the recommended items.
- Looking for a recommended item required too much effort (reverse scale).

A.4.2 Ease of Preference Elicitation

- I found it easy to tell the system about my preferences.
- It is easy to learn to tell the system what I like.
- It required too much effort to tell the system what I like (reversed scale).

A.4.3 Ease of Preference Revision

- I found it easy to make the system recommend different things to me.
- It is easy to train the system to update my preferences.
- I found it easy to alter the outcome of the recommended items due to my preference changes.
- It is easy for me to inform the system if I dislike/like the recommended item.
- It is easy for me to get a new set of recommendations.

A.4.4 Ease of Decision Making

- Using the recommender to find what I like is easy.
- I was able to take advantage of the recommender very quickly.
- I quickly became productive with the recommender.
- Finding an item to buy with the help of the recommender is easy.*
- Finding an item to buy, even with the help of the recommender, consumes too much time.

A5. Perceived Usefulness

- The recommended items effectively helped me find the ideal product.*
- The recommended items influence my selection of products.
- I feel supported to find what I like with the help of the recommender.*
- I feel supported in selecting the items to buy with the help of the recommender.

A6. Control/Transparency

- I feel in control of telling the recommender what I want.
- I don’t feel in control of telling the system what I want.
- I don’t feel in control of specifying and changing my preferences (reverse scale).
- I understood why the items were recommended to me.
- The system helps me understand why the items were recommended to me.
- The system seems to control my decision process rather than me (reverse scale).

A7. Attitudes

- Overall, I am satisfied with the recommender.*
- I am convinced of the products recommended to me.*
- I am confident I will like the items recommended to me. *
- The recommender made me more confident about my selection/decision.
- The recommended items made me confused about my choice (reverse scale).
- The recommender can be trusted.

A8. Behavioral Intentions

A.8.1 Intention to Use the System

- If a recommender such as this exists, I will use it to find products to buy.

A.8.2 Continuance and Frequency

- I will use this recommender again.*
- I will use this type of recommender frequently.
- I prefer to use this type of recommender in the future.

A.8.3 Recommendation to Friends

- I will tell my friends about this recommender.*

A.8.4 Purchase Intention

- I would buy the items recommended, given the opportunity.*

6. REFERENCES

[1] Adomavicius, G. and Tuzhilin, A. 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. Knowl. Data Eng. 17(6), 734-749.

[2] Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., et al. 2004. Using social psychology to motivate contributions to online communities. In CSCW '04: Proceedings of the ACM Conference On Computer Supported Cooperative Work. New York: ACM Press.

[3] Castagnos, S., Jones, N., and Pu, P. 2009. Recommenders' Influence on Buyers' Decision Process. In proceedings of the 3rd ACM Conference on Recommender Systems (RecSys 2009), 361-364.

[4] Chen, L. and Pu, P. 2006. Trust Building with Explanation Interfaces. In Proceedings of International Conference on Intelligent User Interface (IUI’06), 93-100.

[5] Chen, L. and Pu, P. 2008. A Cross-Cultural User Evaluation of Product Recommender Interfaces. RecSys 2008, 75-82.

[6] Chen, L. and Pu, P. 2009. Interaction Design Guidelines on Critiquing-based Recommender Systems. User Modeling and User-Adapted Interaction Journal (UMUAI), Springer Netherlands, Volume 19, Issue3, 167-206.

[7] Davis, F.D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quart.* 13 319-339.

[8] Grabner-Kräuter, S. and Kaluscha, E.A. 2003. Empirical research in on-line trust: a review and critical assessment Int. J. Hum.-Comput. Stud. (IJMMS) 58(6), 783-812.

[9] Herlocker, J.L., Konstan, J.A., Borchers, A., and Riedl, J. An algorithmic framework for performing collaborative filtering. In *Proc. of ACM SIGIR 1999*, ACM Press (1999), 230-237.

[10] Herlocker, J.L., Konstan, J.A., and Riedl, J. 2000. Explaining collaborative filtering recommendations. CSCW 2000, 241-250.

[11] Herlocker, J.L., Konstan, J.A., Terveen, L.G., and Riedl, J. 2004. Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. 22(1), 5-53.

[12] Hu, R. and Pu, P. Potential Acceptance Issues of Personality-based Recommender Systems. In Proceedings of ACM Conference on Recommender Systems (RecSys'09), New York City, NY, USA, October 22-25, 2009.

[13] Jones, N., and Pu, P. 2007. User Technology Adoption Issues in Recommender Systems. In Proceedings of Networking and Electronic Commerce Research Conference (NAEC2007), 379-394.

[14] Jones, N., Pu, P., and Chen, L. 2009. How Users Perceive and Appraise Personalized Recommendations. Proceedings of User Modeling, Adaptation, and Personalization conference (UMAP09), 461-466.

[15] Kirakowski, J. 1993. SUMI: the Software Usability Measurement Inventory. *British Journal of Educational Technology***,** 24 (3) 210-214.

[16] Lewis, J.R. 1993. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use.

[17] McGinty, L. and Smyth, B. On the role of diversity in conversational recommender systems. In *Proceedings of the Fifth International Conference on Case-Based Reasoning (ICCBR'03)*, 2003, 276-290.

[18] McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A. and Riedl, J. On the Recommending of Citations for Research Papers. In *Proc. of ACM CSCW 2002*, ACM Press (2002), 116-125.

[19] McNee, S.M., Lam, S.K., Konstan, J.A., Riedal, J. 2003. Interfaces for eliciting new user preferences in recommender systems. User Modeling 2003, 178-187.

[20] McNee, S.M., Riedl, J., and Konstan, J.A. 2006. Being accurate is not enough: How accuracy metrics have hurt recommender systems. CHI Extended Abstracts 2006,1097-1101.

[21] Nunnally, J. C. 1978. Psychometric Theory.

[22] Ozok, A.A, Fan, Q., Norcio, A.F. 2010. Design guidelines for effective recommender system interfaces based on a usability criteria conceptual model: results from a college student population. Behaviour & Information Technology, Volume 29, Issue 1, 57 - 83.

[23] Pu, P., Chen, L., and Kumar, P. 2008. Evaluating Product Search and Recommender Systems for E-Commerce Environments. Electronic Commerce Research Journal, 8(1-2), June,1-27.

[24] Pu, P., Zhou, M., and Castagnos, S. 2009. Critiquing Recommenders for Public Taste Products. In proceedings of the 3rd ACM Conference on Recommender Systems (RecSys 2009), 249-252.

[25] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. 2000. Analysis of recommendation algorithms for e-commerce. ACM Conference on Electronic Commerce, 158-167.

[26] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. 2001. Item-based collaborative filtering recommendation algorithms. WWW, 285-295.

[27] Simonson, I. 2005. Determinants of customers' responses to customized offers: Conceptual framework and research propositions. Journal of Marketing, 69 (January 2005), 32–45.

[28] Sinha, R. and Swearingen, K. 2001. Comparing Recommendations made by Online Systems and Friends. Proceedings of the DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries, 2001.

[29] Stohr, E.A. and Viswanathan, S. 1999. Recommendation systems: Decision support for the information economy. Emerging Information Technologies, K. E. Kendall, Ed. Thousand Oaks, CA: SAGE, 1999, 21-44.

[30] Swearingen, K. and Sinha, R. 2002. Interaction design for recommender systems. In Interactive Systems (DIS2002).

[31] Tintarev, N. and Masthoff, J. 2007. Survey of explanations in recommender systems. ICDE Workshops 2007, 801-810.

[32] Venkatesh,V., Morris, M.G., Davis, G.B. and Davis, F.D. 2003. User acceptance of information technology: Toward a unified view. MIS Quarterly, 2003, 27, 3, 425-478.

[33] Ziegler, C.N., McNee, S.M., Konstan, J.A., and Lausen, G., Improving Recommendation Lists through Topic Diversification. In *Proc. of WWW 2005*, ACM Press (2005), 22-32.