User Model User-Adap Inter (2012) 22:317–355
DOI 10.1007/s11257-011-9115-7

ORIGINAL PAPER

Evaluating recommender systems from the user's perspective: survey of the state of the art

Pearl Pu · Li Chen · Rong Hu

Received: 3 February 2010 / Accepted in revised form: 21 May 2011 /
Published online: 10 March 2012

Abstract A recommender system is a Web technology that proactively suggests items of interest to users based on their objective behavior or explicitly stated preferences. Evaluations of recommender systems (RS) have traditionally focused on the performance of algorithms. However, many researchers have recently started investigating system effectiveness and evaluation criteria from users' perspectives. In this paper, we survey the state of the art of user experience research in RS by examining how researchers have evaluated design methods that augment RS's ability to help users find the information or product that they truly prefer, interact with ease with the system, and form trust with RS through system transparency, control and privacy preserving mechanisms finally, we examine how these system design features influence users' adoption of the technology. We summarize existing work concerning three crucial interaction activities between the user and the system: the initial preference elicitation process, the preference refinement process, and the presentation of the system's recommendation results. Additionally, we will also cover recent evaluation frameworks that measure a recommender system's overall perceptive qualities and how these qualities influence users' behavioral intentions. The key results are summarized in a set of design guidelines that can provide useful suggestions to scholars and practitioners concerning the design and development of effective recommender

P. Pu (✉) · R. Hu
Human Computer Interaction Group, School of Computer and Communication Sciences,
Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland
e-mail: pearl.pu@epfl.ch

R. Hu
e-mail: rong.hu@epfl.ch

L. Chen
Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong
e-mail: lichen@comp.hkbu.edu.hk

systems. The survey also lays groundwork for researchers to pursue future topics that have not been covered by existing methods.

Keywords Research survey · Recommender systems · User experience research · Explanation interface · Design guidelines

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. No longer a fanciful add-on, a recommender is a necessary component to any competitive website offering an array of choices: the technology is used to improve users' choice satisfaction, while reducing their effort, in finding preferred items. For commercial systems, recommender systems also facilitate the increase of business revenue by successfully persuading users to purchase items suggested to them (cross-selling). According to the 2007 ChoiceStream survey, 45% of users are more likely to shop at a website that employs recommender technology. Furthermore, the more a user spends, the more likely he or she is to prefer the support of recommendation technology.[1]

Previously, research on recommender system evaluation mainly focused on algorithm performance (Herlocker et al. 2004; Adomavicius and Tuzhilin 2005), especially objective prediction accuracy (Sarwar et al. 2000, 2001). In recent years, it has been recognized that recommendation accuracy alone is not sufficient to fulfill user satisfaction, build user loyalty or persuade them to purchase (Herlocker et al. 2000; McNee et al. 2006a,b; Swearingen and Sinha 2002). In addition to the overall quality of the recommended items, the system's effectiveness in presenting recommendations, explaining the reasons for its suggestions, and inspiring users' confidence to make decisions also weighs heavily on users' overall perception of a recommender. In contrast to accuracy measures that can be used to evaluate algorithms in online environments, *user-related issues* can only be effectively investigated via empirical studies involving real users. A number of significant works have emerged in recent years since the first publication of such user studies (Swearingen and Sinha 2002). As research outcomes accumulate, it has become necessary to analyze and summarize the divergent results using a coherent framework. Our goal is, therefore, to (1) survey the state of the art of user evaluations of recommender systems, (2) delineate crucial criteria for user satisfaction and acceptance, and (3) summarize the key results into a conclusive set of design guidelines. Hopefully, this survey can provide a basis for future investigation of user issues in recommender systems, as well as suggesting methods to practitioners concerning the design and development of effective recommender systems.

We assume that a recommender system (RS) is an interactive and adaptive application, often serving as a critical part of online services. Its main function is to provide information personalization technology (e.g., news recommendation), suggest items of interest (e.g., music or book recommendation), and provide decision support for users (e.g., suggest items to purchase). We further assume a generic model of system-user

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

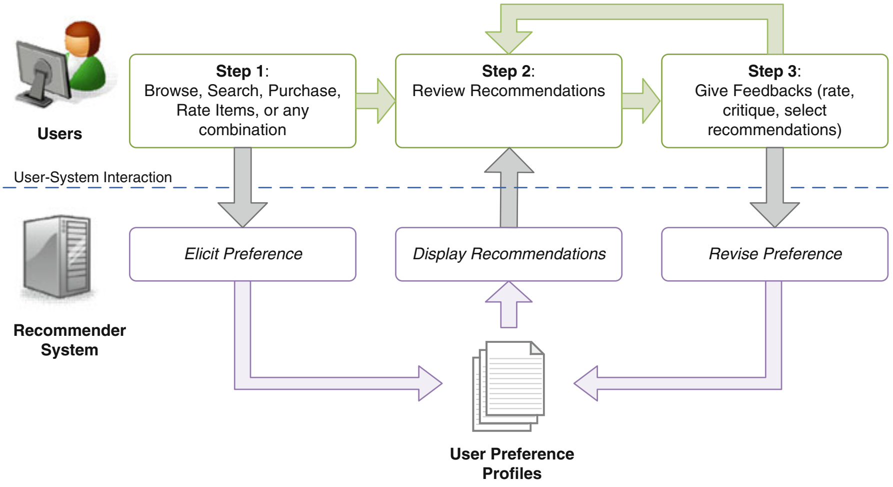

Users
Step 1: Browse, Search, Purchase, Rate Items, or any combination
Step 2: Review Recommendations
Step 3: Give Feedbacks (rate, critique, select recommendations)
User-System Interaction
Recommender System
Elicit Preference
Display Recommendations
Revise Preference
User Preference Profiles

Fig. 1 Generic interaction model between users and recommender systems

interaction as shown in Fig. 1. The three steps located at the top represent user actions arranged in a suggested sequence, typical of most RS's process models. The three processes found at the base of the figure (elicit preference, display recommendations, revise preference) suggest a generic architecture of a RS managing user actions. The details of this generic model are explained as follows.

The initial user profile can be established by users' stated preferences (explicit elicitation) or their objective behaviors (implicit elicitation). In the explicit mode, users are asked to rate items or specify preferences on the features of products. In the implicit mode, the system makes predictions of users' interests by observing users' browsing, searching, selecting, purchasing, and rating behaviors. This user profile initialization (Step 1) is called *preference elicitation* and its duration is called *time-to-recommendation*. After a user's profile has been established, a typical recommender system generates and presents a set of results, called the *recommended items* (Step 2). Users can interact with this *result set* by accepting one or several items (e.g., download or purchase the item), *critiquing* some items ("I like this camera, but I want something cheaper"), or ignoring them. This step is called user *feedback* (Step 3).

Users' subsequent interactions with the RS can improve its prediction of what may interest the users and hence the quality of user profiles. This interaction proceeds in two modes (Step 3). In the behavioral mode, a user's viewing or purchasing of an item signals her interest to the system. The RS hence uses this information to decide what to suggest to her in subsequent steps. On the other hand, if she selects an item but decides to ignore it later (e.g., she skips the song), the system may infer that she is not interested in the item. In the explicit mode of providing feedback, users directly rate or critique the recommended items. For example, at Amazon (www.amazon.com), users can click on the "Fix this recommendation" link next to a suggested item to indicate the degree they like it ("rate it") or they dislike it ("do not use for recommendation").

While Steps 1 and 3 characterize system user interactions with a RS, Step 2 is concerned with a RS's methods and strategies for effectively presenting results to users. The key issues concern the choice of the labels and layouts for the recommended items, the internal structure of the result set, and the set composition.

The *recommendation label* identifies the area on the screen where the recommended items are displayed. The most well-known commercial recommender systems use labels such as "Customer who viewed this also viewed", "Customers who bought this also bought", "Recommendation for you" (all by Amazon), "Movies you'll like" (by Netflix), or simply "Suggestion" (by YouTube). In addition to labels, many recent recommenders employ *explanation* techniques to help users understand the recommender's logic and thus augment the transparency of the system user interaction. For example, Amazon uses "because you purchased" and "because your shopping cart includes" to explain why an item was recommended to a user. In critiquing based recommenders (see detailed review in Sect. 2.2.4), the recommendation labels also explain the value tradeoffs between the user's desired item and the suggested items. Thus the *labels* chosen for the recommendation set play an important role in providing system transparency (as in simple explanation) and persuading users to select items recommended to them (as in tradeoff explanation).

The layout for the recommendation outcome is concerned with the relative location on the computer screen for displaying the content of the result set. The *right-hand longitudinal display* places the items on the right hand side of the screen (as in YouTube), and the *lower latitudinal display* places the recommendation list at the bottom of the screen (as in Amazon). Both layout methods often display the set together with the item that the user is viewing in detail.

The internal structure of the set refers to whether the set will be presented as a linear set of items (list) or whether it is further structured and organized into categories of items (as investigated in Pu and Chen (2006)). The list view is the standard interface used in most commercial websites, while the latter (sometimes called the grid view) represents an emerging and novel interface in the field. Some commercial websites (e.g., www.asos.com) have started adopting this design. The set composition refers to the choice of the set size (how many items to recommend) and how the different kinds of items are mixed and balanced to make up the final set. Studies indicate that while showing one item is too few, showing more than five items increases users' choice difficulty, but increases their perception of diversity. Composing a set with a mixture of top-ranked and mediocre items reduces users' choice difficulty (Knijnenburg et al. 2012).

To provide a framework for the derivation of the guidelines, we define a taxonomy of user criteria for the evaluation of RS as follows: (1) the overall qualities of recommended items including its perceived accuracy, novelty, attractiveness, and diversity; (2) the ease of preference elicitation and revision processes; (3) the adequacy of layout and labels of the recommendation area; (4) its ability to assist and improve the quality of decisions users make; and (5) its ability to explain the recommended results (transparency) and inspire user trust to accept the items recommended to them.

The rest of this paper is hence organized as follows. We start by presenting a basic classification of the types of recommender systems in Sect. 2. We then survey the state of the art of user experience research related to three crucial interaction aspects of a recommender: preference elicitation (Sect. 3), preference refinement (Sect. 4), and results presentation strategies (Sect. 5). For each aspect, we present key results from existing user evaluation work to derive a set of design guidelines. In preference elicitation, we are concerned with the initial interaction between the user and the recommender system. We pay particular attention to the tradeoff between accuracy and effort and draw design guidelines that ensure an optimal balance. In preference refinement and results presentation, we focus on the treatment of accuracy and user confidence. We examine various interaction techniques and develop guidelines that can help the system better motivate users to achieve high accuracy and inspire decision confidence. These sections are followed by our discussion of related works that aim at surveying and establishing an overall user evaluation framework to assess a recommender system's perceived qualities (Sect. 6). Section 7 concludes the survey.

2 Recommender types

Two types of recommender systems have been broadly recognized in the field related to the way systems gather and build user preference profiles: those based on users' explicitly stated preferences, called preference-based recommenders, and those based on users' navigation or purchase behaviors, called behavior-based recommenders (Adomavicius and Tuzhilin 2005; Herlocker et al. 2004; Resnick and Varian 1997). The first category of systems, where users actively state their preferences and provide feedback, will receive more attention in our survey, especially for the sections concerning preference elicitation and refinement. In the following, we describe three types of preference-based recommenders: rating-based, feature-based (which includes case-based, utility-based, knowledge-based and critiquing-based), and personality-based systems. In the second category of systems (i.e., the behavior-based recommenders), users' interactions with the system only become noticeable after the recommendation list has been generated and displayed. Therefore, we will consider these systems only from the point of view of how the users interact with the recommendation list. Zukerman and Albrecht (2001) provide a comprehensive overview of such recommender systems; other ways to classify recommender systems are discussed by Burke (2002), and Adomavicius and Tuzhilin (2005).

2.1 Rating-based systems

In these systems, users explicitly express their preferences by giving either binary or multi-scale scores to items that they have already experienced. These initial ratings constitute users' profiles, which are used to personalize recommended items. Due to the simplicity of this preference acquisition method, most current recommender systems fall into this category. Adomavicius and Tuzhilin (2005) divided these systems into two primary categories: content-based and collaborative filtering methods. Content-based technologies filter items similar to those the active user has preferred in the past, while collaborative filtering technologies take into account the opinions of like-mined "neighbors". Collaborative filtering methods are especially suitable for the domains where item contents cannot easily be parsed automatically.

Recently, instead of overall preference ratings, ratings on items' detailed attributes have been considered in the implementation of recommender systems with the purpose of obtaining more refined user preference profiles and suggesting more precise recommendations. For example, tripadvisor.com, a travel guide and recommender system, asks users to rate a hotel on multiple criteria: value, service, rooms, sleep quality, location and cleanliness.

2.2 Feature-based systems

Rating-based recommender systems apply historical data (i.e., ratings) to estimate user preferences. However, historical data might become obsolete over time. Another major category of recommenders models user profiles and produces personalized recommendations based on their specified needs on item features. This type of recommender system can also be referred to as content-based systems in conventional taxonomy (Adomavicius and Tuzhilin 2005). However, one distinguishing characteristic of feature-based systems identified in this article is that they allow users to explicitly express their preferences on specific item features.

Feature-based recommenders do not attempt to build long-term generalizations about their users, but rather base their advice on an evaluation of the match between a user's need and the set of options available (Burke 2002). They are suitable especially for recommending products infrequently purchased in a user's life or those demanding a significant financial commitment, also called as *high involvement* products in (Spiekermann and Parachiv 2002), such as notebooks, cars, or digital cameras. In this category, four kinds of systems can be identified in the literature: case-based, utility-based, knowledge-based and critiquing-based. Generally speaking, the latter three systems are variants of case-based systems. Case-based systems can rely on utility-based or knowledge-based technologies to assess the similarity between a case and a user's query. Critiquing-based systems are a form of case-based systems which operate in a reactive fashion. We briefly introduce each system in the following sections.

2.2.1 Case-based systems

Case-based recommender systems have their origins in case-based reasoning (CBR) techniques (Smyth 2007). CBR systems solve new problems by utilizing a case base (i.e., database) of past problem solving experiences, instead of codified rules and strong domain models, and retrieving a similar case and adapting its solution to fit the target situation. In case-based recommenders, items or products are represented as cases, and recommendations are generated by retrieving those cases that are most similar to a user's query or profile. Even though case-based recommenders, like other content-based systems, generate recommendations based on item features, they have two important distinguishing characteristics. First, they rely on more structured representations of item content, while content-based systems normally employ an unstructured or semi-structured manner; therefore, case-based recommenders are particularly suitable for product domains where detailed feature descriptions are readily available. Second, they can take advantage of various sophisticated similarity assessment approaches for retrieving similar cases, due to the structured representation of items.

2.2.2 Utility-based systems

Utility-based recommenders make suggestions based on a computation of the utility of each object for a user. Users have to explicitly or implicitly specify their preferences for a set of attributes which characterize the multi-attribute product type. For example, digital cameras can be represented by the set of attributes: manufacturer, price, resolution, optical zoom, memory, screen size, thickness, weight, etc. The central problem for this type of system is creating a utility function for each user. There are different techniques for arriving at a user-specific utility function and applying it to the objects under consideration (Guttman 1998). The benefit of utility-based recommendation is that it can apply non-product attributes, such as product availability, into the utility computation; for example, a user who has an immediate need could trade off price against a delivery schedule.

2.2.3 Knowledge-based systems

Unlike other recommendation techniques, knowledge-based approaches have functional knowledge on how a particular item meets a particular user need, and can therefore reason about the relationship between a need and a possible recommendation. The user profile can be any knowledge structure that supports this inference, from a simple query to a more detailed representation of a user's needs (e.g., CBR in Burke (2002)). The knowledge used by a knowledge-based recommender can also take many forms. For example, in Burke (2002), the implemented restaurant recommender, Entree, uses knowledge of cuisines to infer similarity between restaurants.

2.2.4 Critiquing-based systems

A critiquing-based product recommender simulates an artificial salesperson that recommends options based on users' current preferences and then elicits their feedback in the form of critiques such as "I would like something cheaper" or "with faster processor speed." These critiques help the agent improve its accuracy in predicting users' needs in the next recommendation cycle. For a user to finally identify an ideal product, a number of such cycles are often required. Since users are unlikely to state all of their preferences initially, especially for products that are unfamiliar to them, the preference critiquing agent is an effective way to help them incrementally construct their preference model and refine it as they are exposed to more options. Chen and Pu (2009) provide a set of usability guidelines related to critiquing based recommenders.

2.3 Personality-based systems

As an emerging preference elicitation technique, personality acquisition methods have been employed in recommender systems to help build user profiles (Hu and Pu 2009a,b, 2010). Research has shown that personality is an enduring and primary factor that determines human behavior and that there are significant connections between personality and people's tastes and interests. In light of these findings in psychology, it is reasonable to conclude that personality-based recommenders can provide more personalized information/services, since they understand the customers better from a psychological perspective. As with rating-based systems, personality-based recommender systems also have implicit and explicit ways to measure a user's personality characteristics. Implicit methods mainly infer users' personalities by observing users' behavior while accomplishing a certain task, such as playing games (Dunn et al. 2009), while explicit methods rely on personality questionnaires. Since this is still a new research topic, more emphasis should be placed on it.

3 Preference elicitation

3.1 User effort

Effort expenditure has been a central subject of study in decision theories and lately in recommender systems. This attribute concerns the time and effort users expend interacting with the RS before they receive decision assistance. While users cannot always perceive the system's usefulness in the beginning, they are confronted with the effort requirement at the onset of RS use. The central issue involves determining how much effort users are willing to invest in RS use, and how RS designers can best manage this cost–benefit tradeoff.

In earlier rating-based recommender systems, users were asked to rate up to 50 items before receiving any recommendations. While such an elaborate preference elicitation is likely to help the system more accurately predict what may interest users, such a painstaking process may unintentionally drive them away. Drenner et al. (2008) analyzed historical data over 5,000 users who have had at least 65 ratings with MovieLens. The study compared the recommender system's prediction accuracy for these users when they had only rated 1 movie up to 15 movies. The accuracy was measured by offline experiments that computed the Mean Absolute Error (MAE) of system prediction. As expected, as users rated an increasingly larger number of movies, the prediction error decreased, resembling an inverse curve, asymptotically approaching the limit of 0.65. In other words, while users were able to increasingly obtain more precise recommendations by rating more movies, once the asymptotic line is reached, their subsequent effort does not result in any significant reward. In fact, asking users to provide 5 ratings seems to be the optimum for this accuracy versus effort tradeoff. Going from 1 to 5 movie ratings, users were able to reach a significant increase in accuracy (from 0.94 to 0.76 in MAE, a decrease of 20% in prediction error). On the other hand, users who gave 15 ratings were only able to reduce prediction errors to 0.687 in MAE.

Jones and Pu (2007) compared two music recommender systems in a controlled study involving 64 real users, in a with-in subject experiment design. This setting allows the users to experience two systems one after the other, immediately or within few days' time. It provides more statistical power as the sample size doubles (more data points collected), and, more importantly, it is possible to assess their preference between the two systems. The disadvantage is that user fatigue and participation motivation need to be carefully managed. In this study, users were able to listen to a series of recommended songs shortly after interacting with one system, by specifying the title of one song, the name of an artist, or the name of an album. In the other system, users had to install a plug-in to initialize their music profiles and wait for few days (∼5 days) before receiving personalized recommendations. The results show more than 70% of the tested users preferred the first system whose time-to-recommendation is almost instantaneous over a system which requires users to reveal their preferences over several days. General satisfaction for the first system is 4.1 versus only 3.4 for the second system, a decrease of more than 20% on the Likert scale. Basartan (2001) conducted an empirical study in which a simulated shopbot was used to evaluate users' subjective opinions. A key variable in the study was the shopbot's response time. The study found that users' preference for the shopping agent decreased when the time-to-recommendation increased.

As the field matured, it became apparent that placing more emphasis on the algorithm's accuracy than on users' effort would cause serious user experience problems. Today it is no longer a common practice to elicit 50 ratings from users or extensively request their preferences (Viappiani et al. 2006). For example, MovieLens requires 15 ratings in the registration stage; Jinni (http://www.jinni.com) and Filmaffinity (http://www.filmaffinity.com) ask users to rate around 10 movies to obtain recommendations. If other methods are used, e.g., personality quizzes (Hu and Pu 2010), the initial effort should also be limited to the range of 10–15 questions.

According to behavioral decision theories, human decision makers focus more on reducing cognitive effort than on improving decision accuracy, because feedback on effort expenditure tends to be immediate while feedback on accuracy is subject to delay and ambiguity (Einhorn and Hogarth 1978; Kleinmuntz and Schkade 1993; Haubl and Trifts 2000). This infers that the difference in accuracy among recommender engines might be less important than the difference in their demands for user effort.

Thus, we suggest:

Guideline 1 Minimize preference elicitation in profile initialization. In the tradeoff between accuracy versus effort, users are likely to settle on the immediate benefit of saving effort over delayed gratification of higher accuracy. This guideline can be applied to all types of recommender systems that require an explicit preference elicitation process.

While it is a general trend that users avoid effort, not all of them have the same predisposed effort levels. It is possible to design an adaptive system that tailors the user interaction effort to his or her individual characteristics. It is also imaginable that methods differ in terms of how they motivate users to expend this effort.

Mahmood et al. (2010) proposes treating the system-user interaction design with *adaptive recommendation strategies*, dynamically selecting the most optimal method as the system learns about the characteristics of an active user. In particular, after a user makes a request (e.g., a search for hotels), the system determines the kinds of recommending computations it will perform and the specific user interface it will show. To select an appropriate action to execute given the user's current situation, they applied reinforcement learning techniques to decide on the optimal policies. With the involvement of 544 users, they tested the adaptive learning approach within an online travel recommender and validated that this method enables more users to add products to their carts with less effort (i.e., reduced session duration and reduced number of queries).

Berger et al. (2007) examined an innovative technique to elicit user preferences for vacation destinations based on photos selected by users. An online survey of over 476 subjects was conducted, where each user selected photos from a catalog of 60 photos classified into 17 types such as Anthropologist, Escapist, Archaeologist, etc. Correspondence analysis was then used to produce a mapping of relationships between tourist types and the photographs. This work showed that preference elicitation is not restricted to methods such as rating items. It can be obtained from asking users to choose visual impressions. Further, it seems that users can answer the visual query over 60 photo items without much effort, thus suggesting that a more entertaining process can increase users' predisposed effort levels.

According to these two investigations, we suggest:

Guideline 2 Consider crafting the initial elicitation process adaptive to individuals' characteristics, such as their expertise and knowledge of the domain, and consider using more entertaining means to motivate users to provide input. This guideline has been tested in the travel product domain and has the potential to be applicable to all recommenders.

3.2 Composition of items to rate

Rashid et al. (2002) studied six elicitation strategies: pure random, MovieLens Classique, popularity, pure entropy, combination of popularity and entropy, and item-to-item personalization. *Random* selects items to present arbitrarily with uniform probability over the universe of items. *MoviesLens Classique* selects one movie randomly from an ad hoc list of popular movies, with the rest on that page randomly taken from all movies. *Popularity* lists the movies in descending order of number ratings, with the highest rated movies displayed first. The *pure entropy* strategy proposes items with higher information gains (i.e., movies that some users hate and others like). The *combination of popularity and entropy* (Pop * Ent) strategy is a hybrid approach that combines the popularity and pure entropy methods. The reasoning was to make the two strategies compensate for each other, in order to address the popularity strategy's lack of information strength and the entropy strategy's likelihood of making users rate a movie that they have not seen. The *item-to-item personalized method* learns which items users have rated thus far and then suggests some related ones for them to rate next. In an experiment where historical data was used, researchers mimicked an online sign-up process. They used each strategy to present a total of 30, 45, 60, or 90 movies to each "user". The experiment showed that the item-to-item strategy performed the best in picking movies "users" could rate, while Pure Entropy did the worst. Specifically, when presented with 30 movies, simulated "users" (for whom the database has the ratings) were able to rate up to 20 movies with the item-to-item strategy and up to 15 movies with the popularity and Pop * Ent methods. However, using the Pure Entropy list, less than 2 could be rated.

Because the Pure Entropy strategy turned out to be an inferior method, it was eliminated in the follow-up online study. In the subsequent online user study, 351 new users took part in the sign-up process during a 10-day period (49 users left during the process). The system continuously presented a set of movies until users voluntarily rated at least ten of them. The results indicate that users in the popularity group rated 10 items within an average of 1.9 pages, while the item-to-item group went through 2.3 pages on average to rate 10 items. Log Pop * Ent, a variant of Pop * Ent, required much more effort from users, as the average number of pages was 4.0 for rating 10 movies. The MovieLens Classique method involved even more effort, requiring 7.0 pages before users rated 10 movies. This user study hence suggests that the popularity and item-to-item strategies are by far the most effective methods with over 90% of users being able to sign up in five pages or less.

Based on these results, we are able to draw a guideline for rating-based recommender systems:

Guideline 3 Consider showing a mixture of popular and personalized items for users to rate when eliciting their preferences. The relative proportion of the two strategies depends on the product domains; for some products (e.g., restaurants, accommodations) users are less likely to have experienced popular items, so the proportion of personalized items should be higher.

3.3 System or user in control

McNee et al. (2003) compared three interface strategies for eliciting movie ratings from new users. In the first condition, the system asked users to rate movies chosen based on a combination of the popularity and entropy strategies (as explained above); in this case, the system decided which movies users should initially rate (system-controlled list). In the second user-controlled condition, users were prompted to recall a set of movies to rate. In the third, the mixed-initiative condition, users had both possibilities. In all cases, users were required to rate a minimum of 12 movies on a 5-star scale before receiving personalized recommendations. A total of 192 users completed the experiment. It was found that users who received the user-controlled interface took more time to complete the process and did not provide as many ratings as users in the mixed-initiative or system-controlled groups. However, the user-controlled interface had the lowest Mean Absolute Error (MAE) (i.e., better recommendation accuracy) compared to the other two. Furthermore, users of this group felt that the items that they entered more accurately represented their tastes than those entered by the other two groups, which shows that a higher level of user control over the interaction gives rise to a more accurate preference model. The experiment also revealed that although more time was spent by the user-controlled group, users did not perceive any additional effort. 70% of them perceived the sign-up process time as short, as did 69% of users from the system-controlled group. In terms of motivating users to continue providing ratings in the future, the user-controlled interface proved to be more active. Within a 25-day period of time after the experiment, the user-controlled group members rated 225 more items on average, compared to 125 more items provided by the system-controlled user group.

According to this experiment, we suggest:

Guideline 4 Consider letting users decide which items to rate. This approach will likely enhance their sense of control over the preference elicitation process and motivate them to rate more items in the future.

3.4 Elicitation of feature preferences

Thus far, we have mainly considered methods that build users' interest profiles based on ratings, which have been commonly used via rating-based recommenders. In other types of systems such case-based and utility-based recommenders, the focus lies more on the elicitation of user preferences on product features. For instance, Burke (2002) and Adomavicius and Tuzhilin (2005) discussed how the recommendation technology relies on the similarity between items' features and how to construct a multi-criteria preference model. Consider the example of a flight recommender; rather than asking users to rate flights that they have taken in the past, the system can ask users' criteria on airline, departure time, intermediate airport, and so on. Another way of eliciting users' preferences was to ask them to specify a set of example items that they preferred in the past, the method used by case-based recommenders. For example, Pandora (www.pandora.com), an internet radio and music recommender website, lets users start a radio station by entering the name of a song, an artist, or album. The system uses this initial feed's features to deduce what the user prefers and then recommend similar items.

As a matter of fact, feature-based preference elicitation has been more popularly applied to high-risk and high-involvement products, such as cars, computers, houses, cameras, with which users rarely have experience (so it is difficult to obtain their ratings), while it is feasible to ask them to identify criteria on specific features. Furthermore, besides eliciting users' preferences on features' values (e.g., brand, price, size, weight), most related works have aimed at identifying the relative importance (i.e., weight) between different features. For example, in the ATA system, which is a flight selection/recommendation agent (Linden et al. 1997), a user's preference structure is assumed to be weighted additive independent. Each constraint (e.g., on the arrival time) is associated with a weight to indicate its importance. The Apt decision agent (used in the domain of rental apartments) also builds user model as a weighted feature vector (Shearin and Lieberman 2001). Each feature of an apartment has a base weight determined as a part of domain analysis. The system allows users to freely change the weight on individual features by dragging the feature onto a slot (from -1 to 6). In a series of preference-based product search systems developed by Pu and Chen (2005, 2006, 2007), users are also supported to provide the weight of each participating attribute (e.g., from 1 "least important" to 5 "most important"). If a user expresses a preference, but does not know at this point how important this preference is, the default value is set to "somewhat important". A preference structure is hence a set of (attribute value, weight) pairs of all participating attributes. When obtaining users' preferences, the system applies the weighted additive sum rule (Chen and Pu 2007) to compute recommendations. They performed a user study to test the interfaces in the real estate domain and found that most users were active in specifying and adjusting the weights during the whole interaction process. This type of weighted preference structure has also been proven to enable users to conduct tradeoff navigation and make effective recommendation critiques. Similar phenomena were later verified in other product domains, such as digital cameras and computers. Furthermore, such feature-based preference elicitation methods relying on explicit user feedback are especially suited for the context of providing recommendations to anonymous or first-time online users who have no item ratings available for learning (Zanker and Jessenitschnig 2009).

We thus suggest:

Guideline 5 Consider allowing users to indicate the relative importance of the features of a product/item that they prefer. This approach will likely help establish a more accurate preference model for decision support recommenders in high-risk product domains.

3.5 Preference elicitation via personality questionnaire

Personality, one of the most important user characteristics and resources for preference modeling, is increasingly attracting researchers' attention in the realm of recommender and other adaptive systems. Personalities reflect more innate and permanent characteristics of a person, which is less likely to be affected by specific item domains. Hu and Pu (2009a,b) conducted a real user study to compare a personality quiz-based movie recommender system with a rating-based movie recommender system used as the baseline. The objective was to find out if users would find the two types of systems comparable, concerning the perceived quality of the recommended items, initial elicitation effort, and system effectiveness. Users had to answer a personality questionnaire (20 questions) versus rating 15 movies before receiving any movie recommendations. The results from that study show that while the personality-based recommender is perceived to be only slightly more accurate than the rating-based one, users found it much easier to use. This leads to a higher user acceptance of the personality quiz-based recommender system. Moreover, the actual completion time has no significant correlation to users' cognitive effort in the personality-based recommender (a sign that users do not notice the lapse of time), while such a correlation could be found in the rating-based recommender system. Further analysis of users' comments after the experiment show that the questions used in the personality quizzes were perceived as more entertaining, while the rating elicitation was not. Therefore, personality-based recommenders are viable methods that can be appreciated by users, and the elicitation method could be used to alleviate the cold-start problem. Commercial systems that employ personality modeling include gifts.com and Hunch.

Therefore, we propose:

Guideline 6 Personality traits can be considered as useful resources for user preference modeling and personality-based recommenders. This method can alleviate the cold-start problem, especially if the personality quizzes are designed in an entertaining way. In general, a more enjoyable preference elicitation process reduces users' perceived cognitive effort and increases their satisfaction.

3.6 Comparison of elicitation methods

As shown in the previous section, the realm of elicitation methods has expanded, and preference elicitation is no longer restricted to only rating based approaches. With more methods used in practice, it is appropriate to evaluate whether users prefer one way to another. Hu and Pu (2009a) recently reported a comparative study where 30 participants were told to evaluate two movie recommenders side by side in a within-subject experiment setting. This experiment aims at comparing users' preferences over how they would like to reveal their preferences. One system was MovieLens where users had to rate at least 15 movies before receiving any recommendations. The second system was What-to-Rent, a personality based recommender where users had to answer 20 personality questions in the initial preference elicitation phrase. The study measured and compared the perceived accuracy of the recommended movies, the perceived cognitive effort and the actual effort (completion time) focusing on preference elicitation. Users were also asked whether they would like to reuse the tested systems and introduce them to friends, and they were asked to indicate their final preference between these two systems.

The results showed that the personality quiz method scored slightly higher in perceived accuracy than the rating-based system, but the difference is not significant. However, when participants were asked about their perceived effort in preference elicitation, the personality-based system was rated as requiring significantly lower effort than the rating-based system. Additionally, the average time-to-recommendation was 6.8 min for the personality-based system versus 18.7 min for the rating-based system. 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 (the remaining users had a neutral opinion). This preference represents a significant distribution ($p < 0.001$ by the Chi-squared test). Participants also stated a significantly stronger intention to reuse the personality-based system in the future. Likewise, more participants intended to introduce the personality system to their friends.

This user study suggests that if two systems are comparable in terms of users' perceived accuracy, but one system requires significantly more preference elicitation effort, users will more likely adopt the system that demands less effort. They are also more likely to be loyal to it and recommend it to their friends. This study also pointed out that the comparison of systems in a side-by-side experiment is effective in revealing factors that influence users' perceptions and attitudes towards a particular system design.

We therefore suggest:

Guideline 7 Conducting a comparative user study of two designs can help reveal the elicitation method that most optimally balances accuracy and effort.

3.7 User contribution

As Beenen et al. (2004) discovered, most users are social loafers rather than contributors. Motivated by such problems, researchers referred to social psychology to gain some insights; they particularly focused on the effect of community on users' behavior since people are increasingly aware of other users and contributors. One of the questions they asked was: *if a user is not motivated to contribute ratings for her/his own sake, would s/he do it for the sake of the community?* Karau and Williams (2001) developed a collective effort model (CEM) and identified some conditions under which people will contribute more to a group. For example, when users believe that their effort is important to the group's performance and their contributions to the group can be identifiable, they will work harder. The second question the researchers asked was: *if people are motivated to contribute, should they be given specific goals or just be told to do their best?*

Two sets of large-scale user studies were conducted in the MovieLens setting (Beenen et al. 2004). The findings offered a rare opportunity to observe how social psychology theories can be adopted to understand design issues for recommender systems. Specifically, in their experiment, a rating campaign was launched among 830 active MovieLens users (those who have rated at least three movies). After receiving an email invitation, 397 members logged in and rated at least one movie. These users had actually rated an average of 5.4 movies in the 6 months before the invitation. One important finding from this study is that members who were told that they were going to make a unique contribution provided significantly more ratings on average (i.e., 31.53 ratings versus 25.16 for those who were not told that their contribution was unique). Another result is that users contributed lower amounts of ratings when they were told that their contribution would only benefit themselves (13.19 ratings), compared to the condition in which their contribution was stated to benefit both themselves and others (19.21 ratings). Based on this study, our design suggestion is:

Guideline 8 Consider making users aware of their unique contributions to communities. This approach will likely motivate users to contribute.

At Amazon, there is a feature called "be the first reviewer for this product", which can be regarded as an application of this guideline. It might make users aware of their contribution to the community.

A second study reported in the same article (Beenen et al. 2004) investigated whether goal-setting theories (Locke and Latham 2002) could boost members' contributions to a community or not. According to this theory, assigning people with challenging, specific goals can lead them to achieve more. However, pushing users past a certain point will result in their performance leveling off. The experiment tested the theory's benefits and limitations in a similar email campaign, in which 834 new users participated. An interesting result of this study is that participants who were given specific goals contributed significantly more than those who were given the "do your best" goal. Moreover, they provided more ratings if they were told that they belonged to a group of "explorers", where the whole group was required to rate at least 32 movies.

From this study's results, we thus propose:

Guideline 9 Consider setting specific goals for users to achieve. This approach will also motivate users to contribute. In addition, if users can identify themselves within specific communities (e.g., "explorer", "comedy fan club"), they will likely contribute more.

3.8 Privacy issues

Privacy is a critical issue for recommender systems, regardless of whether the adopted user modeling method is explicit or implicit (Resnick and Varian 1997; Riedl 2001). To understand users well enough and make effective recommendations, a recommender must acquire sufficient information (e.g., demographic information, preference information, personality information etc.) about the users. The privacy concern becomes more important when the required information is more personal and users want to keep it confidential. Finding the optimal balance between privacy protection and personalization remains a challenging task.

There is a general assumption that people are sensitive to privacy issues. However, a mismatch exists between people's privacy preferences and their actual behavior. Spiekermann et al. (2001) compared self-reported privacy preferences of 171 participants with their actual disclosing behavior during an online shopping episode. In their study, most individuals stated that privacy was important to them, with concern centering on the disclosure of different aspects of personal information. However, regardless of their specific privacy concerns, most participants did not adhere to their self-reported privacy preferences. The results suggest that people overlook privacy concerns once they are highly involved in the system. Knijnenburg et al. (2010) validated the results of Spiekermann et al. They carried out a user study with 43 participants in the Microsoft ClipClub system (a multimedia recommender) with the aim of determining the factors that influence users' intention to provide feedback. Their results show that users' intention to provide feedback is influenced by both user experience (choice satisfaction, the perceived effectiveness of the system) and users' privacy concerns. That is, initial privacy concerns can be overcome when users perceive an improvement of their experience when they provide feedback. Moreover, studies show that explicitly indicating the resulting benefits could stimulate users to give more information (Brodie et al. 2004).

Another factor which affects users' intentions to divulge personal information is their trust in the system (Lam et al. 2006). Users must trust that the system will protect their information appropriately; reputed websites are able retain their users with considerable loyalty based on a high level of trust. For example, at Amazon, users are more apt to give personal information since they believe their information will be kept confidential. In addition, when users decide whether to provide personal information, they would like to know who is able to access the information and its specific purpose (Kobsa and Schreck 2003). A serial of user studies conducted by Pu and Chen (2006, 2007) have indicated that explanation interfaces could effectively help build users' trust in the system. Other researchers also contend that explanation interfaces can cultivate user trust (Sinha and Swearingen 2002).

Therefore, even though users are concerned with revealing their personal information, their initial privacy concerns can be overcome by the benefits.

Thus, we propose:

Guideline 10 Consider the tradeoff between users' privacy concerns and their willingness to reveal reference information. Designing explanation interfaces can be an effective method for achieving an optimal balance. The interfaces should explicitly indicate the benefits of providing information, as well as fully disclosing its use, in order to build users' trust in the system.

4 Preference refinement

As indicated previously, users' preferences are context-dependent and will be constructed gradually as users are exposed to more domain information (Payne et al. 1999). This phenomenon has motivated some researchers to develop intelligent tools to help users adaptively build a preference model as they view more recommendations. An effective method of facilitating users to refine their preferences is the *example critiquing* technique (discussed in Sect. 5.2). The system starts the process by first eliciting users' initial preferences and then presenting a few representative items (called *examples*) that may correspond to users' interests. Via a sequence of interaction cycles, called *critiquing*, users gradually refine their profile model until they are satisfied with the outcome proposed by the system.

4.1 Importance of preference refinement

Viappiani et al. (2007) conducted a user study with 22 subjects that compared the number of preferences elicited before and after the critiquing process. The average number of preferences stated was 2.1 on a total of 10 attributes, but this increased to 4.19 attributes when users came to the stage of selecting the final choice. The results suggest that an effective recommender mechanism should allow users to *refine* their preferences. Furthermore, as Swearingen and Sinha (2002) indicated, the initial recommended results are not likely to be 100% accurate, so it is important to design a system that does not look like a dead-end.

For rating-based systems, profile refinement has mainly been handled by requiring users to provide more ratings (Herlocker et al. 2004). For example, in Pandora (introduced previously), users are given the possibility to give a thumb up or down next to each of recommended songs (Jones and Pu 2007). For case-based systems and utility-based systems, this issue is addressed by initiating a preference critiquing support that enhances the refinement of features through a successive series of interactions (Pu et al. 2011).

In commercial sites, Amazon employs a simple preference refinement method. It asks users to rate some specific items under the box that says "Improve Your Recommendations". This facility may convince users that their work will lead to more accurate recommendations and encourage them to put forth more effort.

Therefore, we propose:

Guideline 11 Consider providing preference refinement facilities, such as critiquing, which helps the system increase recommendation accuracy.

4.2 Critiquing support

In this section, we survey critiquing support techniques based on work proposed in case-based and utility-based systems in the past several years. Critiquing support essentially acts like an artificial salesperson. It engages users in a conversational dialog where users can provide feedback in the form of critiques (e.g., “I would like something cheaper”) to the sample items. The feedback, in turn, enables the system to refine its understanding of the user’s preferences and its prediction of what the user truly wants. In the next interaction cycle, the system is then able to recommend products that may better stimulate the user’s interest.

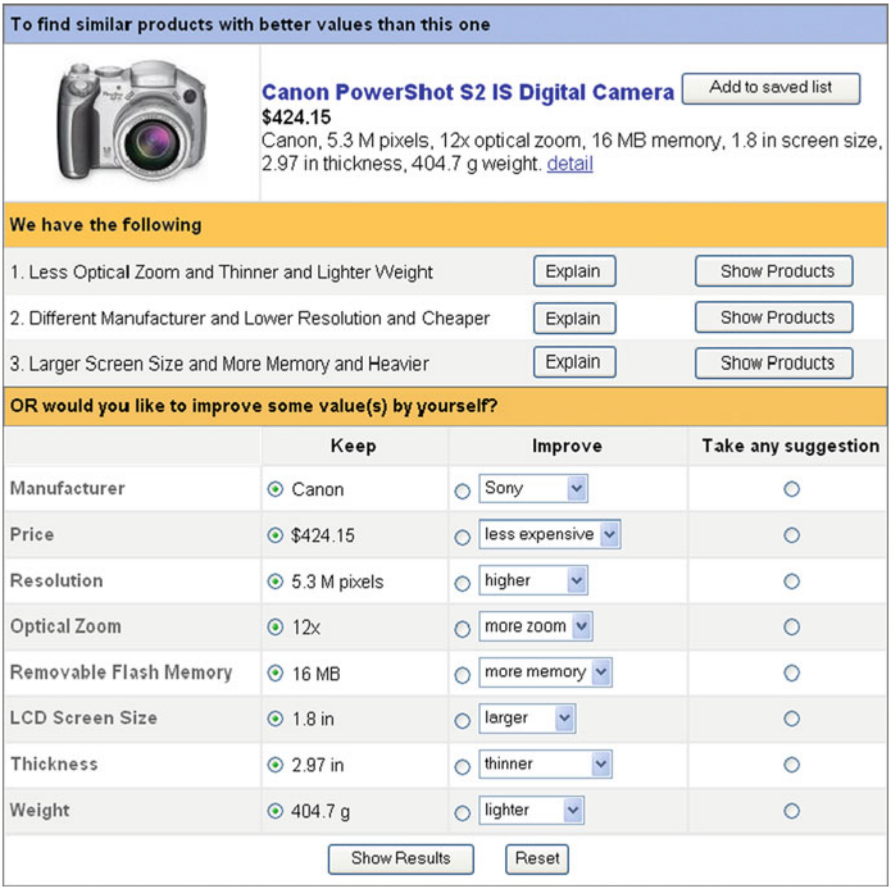
To our knowledge, this critiquing concept was first mentioned in the RABBIT system (Williams and Tou 1982) as a new interface paradigm for formulating queries to a database. In recent years, it has evolved into two principal branches. One aims to proactively generate a set of knowledge-based critiques that users may be prepared to accept as ways to improve the current product, termed *system-suggested critiques* in Chen and Pu (2009). This mechanism was initially used in FindMe systems (Burke et al. 1997), and more recently in Dynamic Critiquing (DC) agents (Reilly et al. 2004; McCarthy et al. 2005). The main advantage, as detailed in related literature (Reilly et al. 2004; McCarthy et al. 2004; McSherry 2005), is that *system-suggested critiques* not only exposes the knowledge of remaining recommendation opportunities, but also accelerates the user’s critiquing process if they correspond well to the user’s intended feedback criteria.

An alternative critiquing mechanism does not propose pre-computed critiques, but provides a facility to stimulate users to freely create and combine critiques themselves (*user-initiated critiquing support* in Chen and Pu (2009)). As a typical application, the Example Critiquing (EC) agent has been developed for this goal, and its focus is on showing examples and facilitating users’ composition of self-initiated critiques (Pu and Faltings 2004; Pu and Kumar 2004; Pu et al. 2006). In essence, the EC agent is capable of allowing users to choose any combinations of feature(s) to critique and the amount of variation they want to apply to the critiques. Previous work has proven that this enables users to obtain significantly higher decision accuracy, preference certainty, and sense of control, compared to non critiquing-based product search systems (Pu and Kumar 2004; Pu and Chen 2005).

Chen and Pu (2009) have further studied how to design effective critiquing components. Through a series of three user-trials, they identified the effects of two crucial elements: *critiquing coverage* (the number of items to be critiqued) and *critiquing aid* (the specific critiquing support), on users’ decision accuracy/effort, and behavioral intentions.

Specifically, the first trial (with 36 subjects) compared two typical existing applications: DC, which suggests one item during each recommendation cycle and a list of system-suggested compound critiques for users to select as improvements to the recommended item, and EC, which returns multiple items at a time and provides a user-initiated critiquing facility to assist users in freely creating their own tradeoff criteria. The results showed that EC performed significantly better than DC in improving users’ objective/subjective accuracy, reducing their interaction effort and cognitive effort, and increasing their purchase and return intentions.

In the second trial (with 36 new subjects), both EC and DC were modified so that their only differing factor was the critiquing aid. Although there is no significant difference between the two modified versions, participants' written comments revealed their respective strengths: *system-suggested compound critiques* made users feel that they obtained more knowledge about the remaining recommendation opportunities and it was easier to make critiques; a *user-initiated critiquing aid* allowed for detailed preference refinement and more user-control over composing users' own search criteria. Moreover, the significant effects of k recommendations in the first round (k-NIR) and k recommended items after each critiquing action (k-NCR) were identified ($k = 7$ in the experiments). That is, k-NIR significantly contributed to improving users' decision accuracy and trusting intentions, and k-NCR performed effectively in saving users' objective effort, including task time and critiquing cycles. In addition, it implies that this multi-item display strategy represents the key factor that led to EC's success in the first trial.



To find similar products with better values than this one

Canon PowerShot S2 IS Digital Camera

$424.15

Canon, 5.3 M pixels, 12x optical zoom, 16 MB memory, 1.8 in screen size, 2.97 in thickness, 404.7 g weight. detail

We have the following

1. Less Optical Zoom and Thinner and Lighter Weight	Explain	Show Products
2. Different Manufacturer and Lower Resolution and Cheaper	Explain	Show Products
3. Larger Screen Size and More Memory and Heavier	Explain	Show Products

OR would you like to improve some value(s) by yourself?

	Keep	Improve	Take any suggestion
Manufacturer	◉ Canon	○ Sony	○
Price	◉ $424.15	○ less expensive	○
Resolution	◉ 5.3 M pixels	○ higher	○
Optical Zoom	◉ 12x	○ more zoom	○
Removable Flash Memory	◉ 16 MB	○ more memory	○
LCD Screen Size	◉ 1.8 in	○ larger	○
Thickness	◉ 2.97 in	○ thinner	○
Weight	◉ 404.7 g	○ lighter	○

Fig. 2 The hybrid-critiquing interface with both system-suggested critiques and user-initiated critiquing facility (Chen and Pu 2007, 2009)

The third trial (with 18 subjects) evaluated user performance in a hybrid-critiquing interface that combines both system-suggested critiques and the user-initiated critiquing facility on the same screen (see Fig. 2). The experiment found that users were active in making critiques using this interface. Furthermore, in comparison with user data in the system without system-suggested compound critiques or the one without user-initiated critiquing support, the hybrid system enabled users to reach significantly higher levels of decision confidence and return intentions.

Based on these findings, we suggest two guidelines as follows:

Guideline 12 Consider presenting several items for users to critique. This will increase their sense of control and therefore their satisfaction with the critiquing support.

Guideline 13 Consider offering a hybrid-critiquing method with two types of critiquing aids: *system-suggested critiques* and *user-initiated critiquing*. The former provides more critiquing assistance, especially when the system-generated critiques closely match users' interest, while the latter provides users with a higher degree of control.

5 Recommendation presentation

Recommendation results are the set of items a recommender system produces and presents to an active user. In most cases, such results are labeled and displayed in a particular area of the screen (the "North", "South", and "East" sections of the screen are the most common areas). Users are likely to review recommended items as a set rather than reviewing them as individual items. In another words, reviewing one item from the set influences how the user reviews the other items (McNee et al. 2006a; Ziegler et al. 2005).

Because presentation is a crucial factor in persuading users to accept the items recommended to them, each recommender system must carefully employ special strategies that are sensible to users' information needs as well as the business goals of the RS. For example, would the result set include only highly-ranked items, but similar ones? Or would it include a well-balanced top-ranked one and popular items? Should it intentionally include familiar items that the system predicted for users? Should a recommender explain the recommendation's inner logic? According to prior work, users' perception of recommendations is directly related to their trust in the system (Pu and Chen 2006). Trust can further be seen as a long-term relationship between the user and the organization that the recommender system represents. Therefore, recommended results could affect users' confidence in whether the system is providing increasingly useful information. If the recommendation succeeds in inspiring user confidence and motivation to expend more interaction effort, users will increasingly like the system. In the following sections, we will discuss these presentation issues in greater detail.

5.1 Accuracy

In the literature, accuracy has been used as a dominating criterion for evaluating recommender systems (McNee et al. 2006a,b; Herlocker et al. 2004). Over the past decade, a wide variety of algorithms have been developed with the goal of increasing objective accuracy (Herlocker et al. 2004). For example, Mean Absolute Error (MAE), defined as the difference between the predicted ratings of an algorithm and actual user ratings, is the commonly adopted measure to evaluate the objective accuracy of rating-based systems. For content- or feature-based recommenders, Pu et al. (2010) established a procedure, called the switching task, to measure the objective accuracy. Specifically, it measures whether a user changes his/her initially preferred item, which was identified by using a recommender agent, to another one when instructed to view all available items in a later round. If some users change their mind (switching), the accuracy of the recommender will be discounted.

More recently, researchers have been paying attention to users' perceived accuracy. This measure is defined as the degree to which users feel proposed recommendations match their interests and preferences. Compared to objective algorithm accuracy, perceived accuracy has a more direct influence on users' trust formation with the system and behavioral intentions, such as their intention to purchase items recommended to them and intention to return to the system (Chen and Pu 2009).

Some studies indicate that user perception is directly correlated with objective differences in recommendation quality (Ziegler et al. 2005; Chen and Pu 2007). Knijnenburg et al. (2010) investigated the effects of two recommendation algorithms on users' perceived quality. Their results show that users given personalized recommendations scored higher in terms of perceived accuracy, in contrast to the other group of users given randomly selected items. The high correlation suggests that users are aware of the quality distinctions between random and personalized recommendations. These results have also been validated in a series of their follow-up studies (Knijnenburg et al. 2012).

However, some researchers have shown that objective accuracy may not have a direct correlation with perceived accuracy. McNee et al. (2002) compared six recommendation algorithms through both offline and online experiments. In the offline experiment, they found that the "item-to-item" algorithm that achieved the best prediction performance gave the fewest recommendations that users believed to be relevant and helpful.

On the other hand, perceived accuracy can be enhanced by other aspects, such as the level of domain knowledge of individuals (Hu and Pu 2010; Knijnenburg et al. 2012), contexts and interface design (Xiao and Benbasat 2007; Cosley et al. 2003; Chen and Pu 2010), and interface labels (Ochi et al. 2010). The study of Cosley et al. (2003) shows that the display of predicted ratings could manipulate users' opinions on the recommended items, with a bias towards the presented rating values. In the experiment conducted by Ochi et al. (2010), users were manipulated to think that they were evaluating two recommender system approaches by interface labels, when in reality they received the same recommendations. Their empirical results show that for search products (defined as those whose qualities can be determined before their purchase, e.g., rugs), participants tended to like the recommendations from collaborative-based systems, while for experience products (defined as those whose quality cannot be determined until after usage and purchase e.g., perfume), participants tended to feel better about the recommendations from content-based system.

Considering the existence of the gap between objective accuracy and perceived accuracy, we propose:

Guideline 14 Consider enhancing users' perceived accuracy with more attractive layout design and effective labels, and explaining how the systems compute the recommendations. Doing so can increase users' perception of the system's effectiveness, their overall satisfaction of the system, their readiness to accept the recommended items, and their trust in the system.

5.2 Familiarity

Swearingen and Sinha (2002) first noticed the importance of "familiarity" in their user studies. After they found preliminary results that the presence of well-known items reinforces trust in the recommender system in their first user study, they examined this issue in depth in another user study and found that users' mean liking for familiar recommendations is higher than that for unfamiliar recommendations. In addition, most of the users in the study ($>50\%$) agreed that the inclusion of previously liked items in the recommendation set could increase their trust in the system. The study also showed that the users expressed greater willingness to buy familiar recommended items than unfamiliar ones. However, although users prefer familiar items, they did not like that the recommendations were too directly related to their input ratings. For example, if the system always shows albums from the same artist that the user had entered, they felt that the system was not capable of helping them discover new directions and broadening their tastes. This raises the next question about "novel recommendation" (see the next section). Thus, regarding familiarity, as suggested by Swearingen and Sinha, the system should better understand users' needs with relation to familiarity. For example, it could ask users to explicitly indicate how familiar they would like the recommendation set to be (e.g., with a slider bar), which would enable the system to better tailor to their needs.

Thus we suggest:

Guideline 15 Consider including familiar items in the recommendation set, as this aspect is highly correlated to users' trust in the system.

5.3 Novelty

As indicated above, novelty is another dimension that the designers of a RS should consider. The core concept of this dimension is the unexpectedness of recommendations for users. McNee et al. (2006a) also called this criterion "serendipity", which emphasizes the experience of receiving unexpected and fortuitous item recommendations. Herlocker et al. (2004) argued that novelty is different from serendipity, because novelty only covers the concept "new", while serendipity means not only "new" but also "surprising". In this paper, we unify the two understandings and measure "novelty" as the extent to which users receive new and interesting recommendations.

The effect of novelty has experimentally been demonstrated in music recommenders by Jones and Pu (2007). They discovered that users found the songs of Pandora (a popular music recommender site that offers more novel recommendations) to be significantly more enjoyable and better than songs suggested by their friends. This user study also showed that users will more likely use recommender systems that make novel music recommendations, as such a system is perceived to be more useful.

Thus, our suggestion in this regard is:

Guideline 16 Consider achieving a proper balance between familiarity and novelty among recommendations, since novelty can be correlated to the perceived usefulness of the system, especially in entertainment product domains.

5.4 Diversity

Recent research work suggests that the set of recommended items should maintain a certain level of diversity, even at the risk of compromising its overall accuracy. For example, if users only get a recommendation list full of albums by the same artist, the diversity level in this list is low. Jones and Pu (2007) pointed out that low diversity could disappoint users and decrease their decision confidence. McNee et al. (2006a) noticed the same problem and stated that the item-to-item collaborative filtering algorithm may trap users in a "similarity hole", only giving exceptionally similar recommendations. Thus, in recent years, more researchers have focused on generating diverse recommendations and aiming to reach an optimal balance between diversity and similarity (McGinty and Smyth 2003; Ziegler et al. 2005; Smyth and McClave 2001; Pu and Chen 2006). McGinty and Smyth (2003) highlighted the pitfalls of naively incorporating diversity-enhancing techniques into existing recommender systems. They pointed out that diversity should be provided adaptively rather than being enhanced in each and every recommender cycle. If a recommender system appears to be close to the target item, then diversity should be limited to avoid missing this item. But if the recommender system is not correctly focused, diversity can be used to help refocus more effectively. Ziegler et al. (2005) proposed a topic diversification approach based on taxonomy-based similarity. They compared not only the accuracy measures in different levels of diversification for both user-based and item-based CF, but also subjective satisfaction results from a large scale user survey. Their results showed that users' overall satisfaction with recommendation lists goes beyond accuracy and involves other factors, e.g., the diversity of the result set.

In the same study, they found that human perception seems to capture only a certain level of diversification inherent to a list. Beyond that point, it is difficult for users to notice the increasing diversity. In another words, they suggest that users' diversity perception may not correlate to the diversity controlled by algorithms. Bollen et al. (2010) reported an experiment aimed at understanding if objective diversity has a straightforward effect on perceived diversity. After rating 10 movies from the MovieLens dataset, users were told to select a movie to watch and then answer a satisfaction questionnaire. The 137 subjects were divided into 3×3 conditions, independently varying the underlying recommendation algorithms (Most Popular, kNN, Matrix Factorization) and the degree of diversifications (none, little, a lot). Objective diversification was implemented based on movie genre, using a greedy search method proposed by Ziegler et al. (2005). The main outcomes from this user study seem to confirm that users do not necessarily perceive the diversified recommendations as more diverse.

In Pu et al. (2009), an interface design method was described to increase users' perception of recommendation diversity by grouping and categorizing any result set into an organization interface. The method, called the Editorial Picked Critiques (EPC), creates up to seven category titles such as "more popular and cheaper", "same brand and cheaper", "just as popular but cheaper" and put items into the right group. Hu and Pu (2011) conducted an in-depth user study to compare the organization interface with the standard list interface, while keeping the recommendation contents identical. The results show that the organization interface indeed effectively increased users' perceived diversity of recommendations, especially perceived categorical variety. Correlation results further show that perceived diversity more significantly influences users' perceived ease of use and usefulness of the recommender, positive attitudes toward the system and behavioral intentions. More than 65% of users prefer the organization interface, versus 20% for the list interface. 50% of users thought the organization interface was better at recommending products to them versus only 10% for the list interface.

In addition to using organization interfaces, the result set size and mixing best items with low-ranked items could also affect users' perceived diversity. Knijnenburg et al. (2012) conducted an experiment with 174 Dutch participants, where each user first rated 10 movies from the MovieLens dataset and was told to select a movie from a recommendation set to watch. Each user was randomly assigned to one of three conditions. In Top-5, users reviewed the best five recommendations to select a movie to watch; in Top-20, users reviewed the best 20 recommendations; and in Lin-20, users reviewed the best five recommendations, appended with items which only ranked 99, 199, 299, continuing through 1499. In other words, in Lin-20, the result set was intentionally manipulated to include some low-ranked items. Results from this study show that while a larger result set (Top-20) increases the perceived diversity of the recommended items, it also increased users' choice difficulty because more information processing is required. When the result set includes both attractive and low-ranked items, even though this approach compromised on the overall quality of the items, it increased choice satisfaction because the inferior items, when contrasted to the accurate items, provided a decision context that reduced choice difficulty.

We therefore suggest:

Guideline 17 Consider adding diversity to the recommendation set. Since perceived diversity is not directly correlated with predicted diversity, additional methods, such as interface design and adding low-ranked items to the set, can be considered. Perceived diversity is strongly correlated to the perceived usefulness of the system and users' satisfaction.

5.5 Context compatibility

A good recommender system should also be able to formulate recommendations in contextual occasions. For example, for movie recommendations, the context factors may include a user's current mood, occasion for watching the movie, whether or not other people will be present, and whether the recommendation is timely. We propose a concept called *context compatibility* here, which assesses whether or not the recommendations consider general or personal context requirements. Adomavicius and

Tuzhilin (2008) mentioned a similar idea and indicated that the utility of a certain product to a user may depend not only on the two-dimensional user-item space, but also on other dimensions including time, the person(s) with whom the product will be consumed or shared and so on. In these situations, it may not be sufficient to simply generate user-item recommendations. Instead, the recommender system must take additional contextual information into account. For example, a user can have significantly different preferences for the types of movies she wants to see when she is going out with a boyfriend on a Saturday night, as opposed to watching a rental movie at home with her parents on a Wednesday evening. Applying contextual consideration may enable new users to get suitable recommendations even if they have not established robust profiles, and help repeat users to get recommendations on the basis of their instant needs, which increases both types of users' satisfaction with the recommender.

We thus suggest:

Guideline 18 Consider providing recommended items that are context compatible. This characteristic can be highly correlated to the perceived usefulness of the system and users' satisfaction.

5.6 Explanation of recommendations

Providing recommendations that satisfy the criteria listed above is not enough for effective result display. In addition, it is important to assist users in understanding why these items are presented to them. To date, many researchers have demonstrated that providing good explanations for recommendations could help inspire users' trust and satisfaction, increase users' involvement and educate users on the internal logic of the system (Herlocker et al. 2000; Sinha and Swearingen 2001, 2002; Simonson 2005; Tintarev and Masthoff 2007a,b).

For instance, Herlocker et al. (2000) discussed the concept of explanation for automated collaborative filtering systems, such as how the explanation should be implemented based on users' conceptual model of the recommending process and whether providing explanations could improve users' acceptance of a system. They highlighted that the benefits of explanation interface are *justification* (users could decide how much confidence to place), *user involvement* (allowing users to add their knowledge to complete the decision process), *education* (educating users about the computational processes used in generating recommendations), and *acceptance* (improving the acceptance of a recommender system because it helps users make a decision).

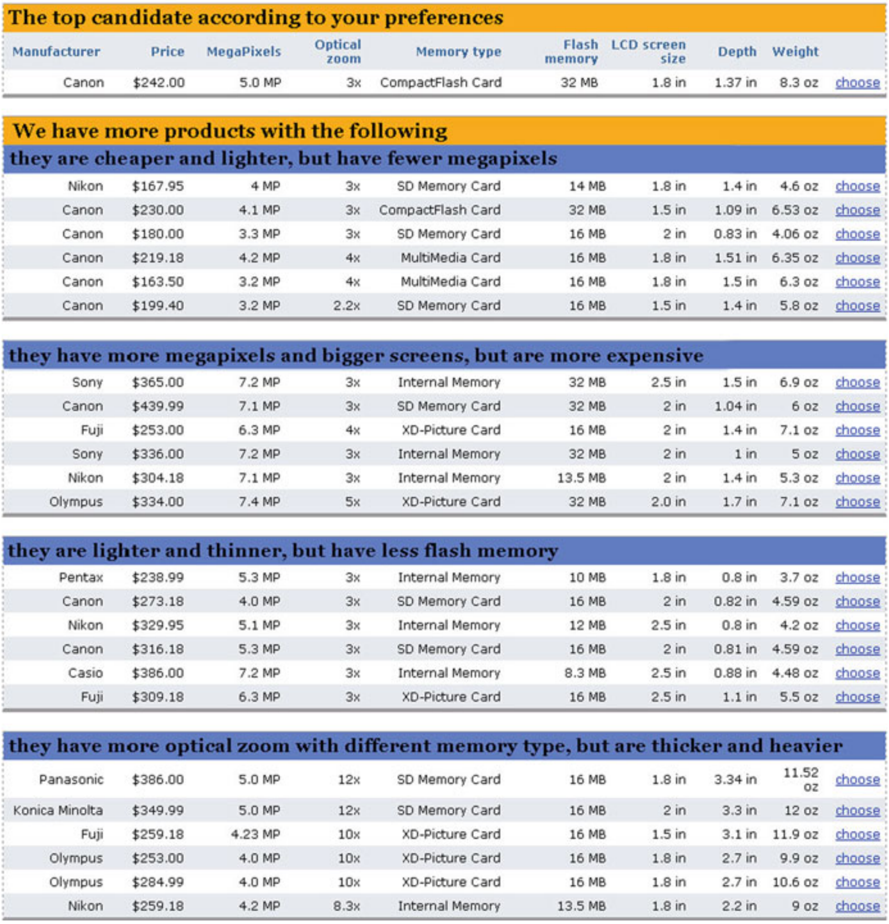
Sinha and Swearingen (2002) found that users like transparent recommendations. The more they have to "pay", the more transparency they require. They stated that the recipient of a recommendation has a number of ways to decide whether to trust the recommendation: scrutinizing the similarity between the taste of the recipient and the recommender; assessing the success of prior suggestions from the recommender; and asking the recommender for more information about why the recommendation was made. Similarly, recommender systems need to offer the user some ways to judge the appropriateness of recommendations. Understanding the relationship between the system input (e.g., ratings made by user) and output (recommendations) allows the user to initiate a predictable and efficient interaction with the system. As we indicated previously, perceived accuracy of a recommendation is also dependent on whether or not the user sees a correspondence between the preferences they express and the recommendation presented by the system (Simonson 2005).

Grounded on prior works on explanation, Tintarev and Masthoff (2007a) concluded several possible goals of explanations, including transparency, scrutability, trust, effectiveness, persuasiveness, efficiency, and satisfaction. *Transparency* explains how the system works, i.e. how a recommendation was chosen. *Scrutability* allows users to correct reasoning and system assumptions where needed. *Trust* could increase users' confidence in the system. A user may be more forgiving if they understand why a bad recommendation has been made, and hence prevent it from occurring again. *Persuasiveness* is the ability to convince users to try or buy. It is important to consider that too much persuasion may backfire if certain users realize that they have bought items that they do not really want. *Effectiveness* is the ability of a system to assist the user in making accurate decision about which recommendations to utilize. *Efficiency* is the ability that explanations may make it faster for users to decide which recommended item is the best for them. Efficiency may be improved by allowing the user to understand the relation between competing options. *Satisfaction* is a quality that makes the use of the system fun.

Several papers also reported user studies to determine effective explanation mechanisms. Herlocker et al. (2000) performed experiments on the use of 23 different graphical displays to "explain" why each recommendation was given. They suggested that there are multiple ways for the system to convey its inner logic to the user: *an explanation* (e.g. "this item was recommended to you because you rated 'x' positively"); *predicted ratings* (e.g. "we think you'll give this item an 8.5 out of 10"); including *a few familiar recommendations* (i.e., ones from artists or writers who are close to input items); *community opinions* (showing both reviews and numerical ratings).

Pu and Chen (2006) presented an organization-based explanation interface where the best matching item is displayed at the top of the interface along with several categories of other recommended products (see Fig. 3). Each category is labeled with a title explaining the characteristics of the items the respective category contains. In order to understand whether the organization interface is a more effective way to explain recommendations, an empirical study was conducted to compare the organization interface with the traditional explanation view (where each item includes an explanation), in a within-subjects design. A total of 72 volunteers (19 females) were recruited as participants in the user study. The results showed that the organization interface significantly increases users' perception of competence, inspires their trust, and enhances their intention to save cognitive effort and use the interface again in the future. Moreover, the study found that the actual time spent looking for a product did not have a significant impact on users' subjective perceptions. This indicates that less time spent on the interface, while important in reducing decision effort, cannot be used alone in predicting what users may subjectively experience. Based on these empirical findings, the authors concluded a set of principles for the effective design of organization interfaces. More details (including the algorithm steps of generating the content for such interfaces) can be referred to in Chen and Pu (2010).

On the other hand, user-generated social contents, such as tags, have also been exploited to generate explanations. One recent method by Vig et al. (2009) developed



The top candidate according to your preferences

Manufacturer	Price	MegaPixels	Optical zoom	Memory type	Flash memory	LCD screen size	Depth	Weight	
Canon	$242.00	5.0 MP	3x	CompactFlash Card	32 MB	1.8 in	1.37 in	8.3 oz	choose

We have more products with the following

they are cheaper and lighter, but have fewer megapixels

Nikon	$167.95	4 MP	3x	SD Memory Card	14 MB	1.8 in	1.4 in	4.6 oz	choose
Canon	$230.00	4.1 MP	3x	CompactFlash Card	32 MB	1.5 in	1.09 in	6.53 oz	choose
Canon	$180.00	3.3 MP	3x	SD Memory Card	16 MB	2 in	0.83 in	4.06 oz	choose
Canon	$219.18	4.2 MP	4x	MultiMedia Card	16 MB	1.8 in	1.51 in	6.35 oz	choose
Canon	$163.50	3.2 MP	4x	MultiMedia Card	16 MB	1.8 in	1.5 in	6.3 oz	choose
Canon	$199.40	3.2 MP	2.2x	SD Memory Card	16 MB	1.5 in	1.4 in	5.8 oz	choose

they have more megapixels and bigger screens, but are more expensive

Sony	$365.00	7.2 MP	3x	Internal Memory	32 MB	2.5 in	1.5 in	6.9 oz	choose
Canon	$439.99	7.1 MP	3x	SD Memory Card	32 MB	2 in	1.04 in	6 oz	choose
Fuji	$253.00	6.3 MP	4x	XD-Picture Card	16 MB	2 in	1.4 in	7.1 oz	choose
Sony	$336.00	7.2 MP	3x	Internal Memory	32 MB	2 in	1 in	5 oz	choose
Nikon	$304.18	7.1 MP	3x	Internal Memory	13.5 MB	2 in	1.4 in	5.3 oz	choose
Olympus	$334.00	7.4 MP	5x	XD-Picture Card	32 MB	2.0 in	1.7 in	7.1 oz	choose

they are lighter and thinner, but have less flash memory

Pentax	$238.99	5.3 MP	3x	Internal Memory	10 MB	1.8 in	0.8 in	3.7 oz	choose
Canon	$273.18	4.0 MP	3x	SD Memory Card	16 MB	2 in	0.82 in	4.59 oz	choose
Nikon	$329.95	5.1 MP	3x	Internal Memory	12 MB	2.5 in	0.8 in	4.2 oz	choose
Canon	$316.18	5.3 MP	3x	SD Memory Card	16 MB	2 in	0.81 in	4.59 oz	choose
Casio	$386.00	7.2 MP	3x	Internal Memory	8.3 MB	2.5 in	0.88 in	4.48 oz	choose
Fuji	$309.18	6.3 MP	3x	XD-Picture Card	16 MB	2.5 in	1.1 in	5.5 oz	choose

they have more optical zoom with different memory type, but are thicker and heavier

Panasonic	$386.00	5.0 MP	12x	SD Memory Card	16 MB	1.8 in	3.34 in	11.52 oz	choose
Konica Minolta	$349.99	5.0 MP	12x	SD Memory Card	16 MB	2 in	3.3 in	12 oz	choose
Fuji	$259.18	4.23 MP	10x	XD-Picture Card	16 MB	1.5 in	3.1 in	11.9 oz	choose
Olympus	$253.00	4.0 MP	10x	XD-Picture Card	16 MB	1.8 in	2.7 in	9.9 oz	choose
Olympus	$284.99	4.0 MP	10x	XD-Picture Card	16 MB	1.8 in	2.7 in	10.6 oz	choose
Nikon	$259.18	4.2 MP	8.3x	Internal Memory	13.5 MB	1.8 in	2.2 in	9 oz	choose

Fig. 3 The organization-based explanation interface to inspire user trust in recommendations (Pu and Chen 2006)

tagsplanation to show tags along with a recommended item. The displayed tags are most relevant to the item and also most likely preferred by the user. They integrated the approach in a rating-based recommender and identified that this tag-based explanation can help users understand why the item was recommended, decide if they like the item, and determine if the item fits their current mood.

In recent years, commercial websites have also realized the importance of explanations. For example, features such as "why this was recommended" have appeared on Netflix, Amazon and Pandora.

The corresponding guideline is hence:

Guideline 19 Consider explaining why the system recommends the suggested items. These aspects can be highly correlated to users' satisfaction, sense of control, and trust-inspired behavior intentions, such as the intention to save effort and the intention to return.

5.7 Sufficiency of information

Finally, it is crucial to provide sufficient information about the recommended items in order to facilitate users' decision-making processes. Sinha and Swearingen (2002) suggested that the presence of longer descriptions of individual items correlates positively with both the perceived usefulness and ease of use of the recommender system. This suggests that users like to have more information about the recommended item, such as book description, author information, genre information and other users' reviews. Their research reinforced this finding by examining the difference between the two versions of RatingZone (one of their sample prototypes). The first version of RatingZone's Quick Picks showed only the book title and author name in the list of recommendations. The second version of RatingZone included simply the title and author, as well as providing a link to item-specific information at Amazon. The perceived usefulness of the second version was almost three times higher than the first. Sinha and Swearingen (2002) also provided several types of information users found useful: *basic item information* (i.e., song, album, artist name, genre information, the time when album was released, the album cover); *expert and community ratings* (reviews and ratings provided by other users); *item sample* (e.g., audio clips of songs, sample chapters of books). If the designer does not have access to detailed item information, offering a community forum with other users' comments can be a relatively easy way to increase the system's efficacy (Sinha and Swearingen 2002). They also suggested organizing the information in an obvious way by using an effective navigation structure; when detailed information was offered, users had trouble actually finding it, since locating item information demanded several mouse clicks and the site had poor navigation design.

Given these results, we suggest:

Guideline 20 Consider providing sufficient information related to the recommended items and controlling the information's quality and navigation structure.

6 Related work

Knijnenburg et al. (2012) provided a framework to explain how objective system aspects (such as its input, process and output) influence users' perception of these aspects, and how this perception eventually influences users' choice satisfaction and their intent to provide feedback (see Fig. 4). A set of six structurally related constructs were proposed in their framework: objective recommender system aspects, subjective evaluations, subjective experiences, objective behaviors, situational, and personal characteristics. The *objective system aspects* refer to the input, output and process aspects of a system, including the visual and interaction design, the recommender algorithm, the feedback mechanisms, the recommendation presentation, and additional system features (e.g., social networking). The *subjective system aspects* are users' evaluations of the objective system aspects. The *situational* (different tasks) *and personal characteristics* (general trust, control and social factors) are external factors. The *subjective experience* also represents an evaluation of the system by the user, but

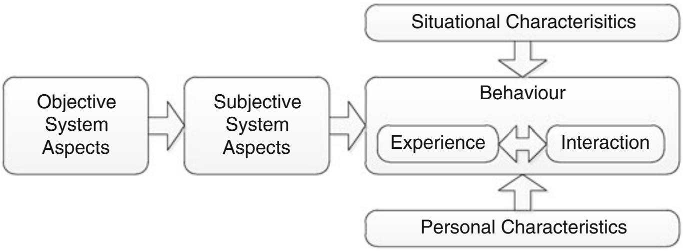


Fig. 4 User-centric evaluation framework for recommender systems (Knijnenburg et al. 2012)

it may primarily depend on personal characteristics. The *interaction* is objectively measurable, and determines users' behavior in the system. For example, a system that is evaluated positively will be used more extensively. The model postulates that objective system aspects influence users' subjective evaluation, and eventually affect users' experience and objective interactions. Furthermore, situational and personal characteristics could mediate this process. A set of six user studies, four field studies and two controlled experiments, were conducted using this framework.

The main findings from these studies show that personalized recommendations (versus randomly selected items) have higher perceived recommendation quality which causes higher perceived system effectiveness and choice satisfaction; despite privacy issues, higher choice satisfaction and system effectiveness can increase users' intention to provide feedback about their preferences; higher perceived recommendation quality leads to less perceived cognitive effort of using the system, and less cognitive effort causes higher perceived effectiveness and fun, which in turn increases users' choice satisfaction; and a higher perceived variety of recommendations can increase the perceived recommendation quality.

The main limitation is that the outcomes were limited to the specific recommender systems (multimedia recommenders) used in all of the six experiments and the specific algorithms used to implement these systems. While the design and procedure can serve as illustrative examples of future work in this area, if some of the system's aspects were to change, users might also alter their perceptions and behavioral intentions. In addition, an important aspect of system, the recommender system's ability to explain its result, was not considered and behaviors are limited to users' intention to provide feedback (preference revision).

Controlled experiments, such as those described in Knijnenburg et al. (2012), are often used to investigate and understand one or a few aspects of the system's interface and interaction design. This method is called layered evaluation in Paramythis et al. (2010). It proposed systematically guiding the evaluations of adaptive systems by breaking down a system into its "constituents" and evaluating each of these constituents separately. Although the layered evaluation method has not been applied to RS, it can be used as a powerful technique in identifying areas of RS that require more focused future work.

In contrast to layered evaluation and controlled experiments, summative evaluations aim at examining the overall impact of the system on users' behavior and explaining users' motivations for using the system. Some researchers have recently started developing evaluation frameworks to measure the overall quality and benefits of a recommender system to its users.

Based on an extensive survey of existing usability experience research on recommender systems, Pu and Chen (2010) proposed a research model to understand the motivations of RS use including intentions to use the system and purchase the recommended items. This model is called *ResQue* (Recommender system's Quality of user experience), and was developed based on TAM theories (Davis 1989). *ResQue*'s constructs (composed of 43 question items) are categorized into four main dimensions: perceived system qualities (such as the quality of recommended items, the interaction characteristics to handle user preferences, the interface qualities to present and explain results, and to provide information sufficiency), users' beliefs as a result of perceived qualities (the beliefs include perceived ease of use, perceived usefulness, control and transparency), users' subjective attitudes (including overall satisfaction, confidence, and trust), and their behavioral intentions (including users' agreement to use the system, purchase intention, return intention and intention to introduce this system to friends). Recently, Pu et al. (2011) reported a large-scale Web-based survey which successfully validated *ResQue* in terms of its structural consistency, validity and reliability, and the model's fitness with respect to the original hypotheses (see Fig. 5). One of the major extensions to TAM is that two types of users' behavioral intentions, i.e., purchase intention and use intention, have been identified to explain RS' role in e-commerce, and entertainment websites. That is, while overall satisfaction of the system (also in terms of ease of use and perceived usefulness) is important for use intention, trust of the system and choice confidence are fundamental for purchase intentions.

ResQue is the first validated framework providing a well-balanced set of user-centric criteria for measuring the success of a recommender system and explaining user adoption of RS by providing an account of how the perceived qualities of a RA influence user beliefs and attitudes, and inspire users' behavioral intentions. The framework merged divergent criteria in past user experience research of RS and, via validation, produced a model defining the essential qualities of an effective and satisfying recommender system. ResQue can be presented as a long or a short version of an assessment questionnaire, 43 and 14 questions respectively, that can help designers and researchers forecast the adoption of recommender systems. The model can be applied to assess different types of recommenders, including rating-based, utility-based, and knowledge-based systems, regardless of the backend engines used. In comparison with Knijnenburg et al. (2012), the *ResQue* framework and its outcome were validated independently of the underlying recommender systems or the algorithms. On the other hand, the model proposed in Knijnenburg et al. (2012) accounts for personal characteristics the users and contextual characteristics of RS use.

Xiao and Benbasat (2007) surveyed previous theoretical and empirical studies of Recommendation Agents (RAs) in both online and offline shopping environments, mainly based on existing literature in management of information systems (MIS). It focused on the RA's function to assist consumers in deciding which products to purchase, as well as how the use of RA, RA properties, and other factors influence consumers' evaluations of RAs. The summary of results were combined into a conceptual model of 28 hypotheses using theories of human information processing, interpersonal

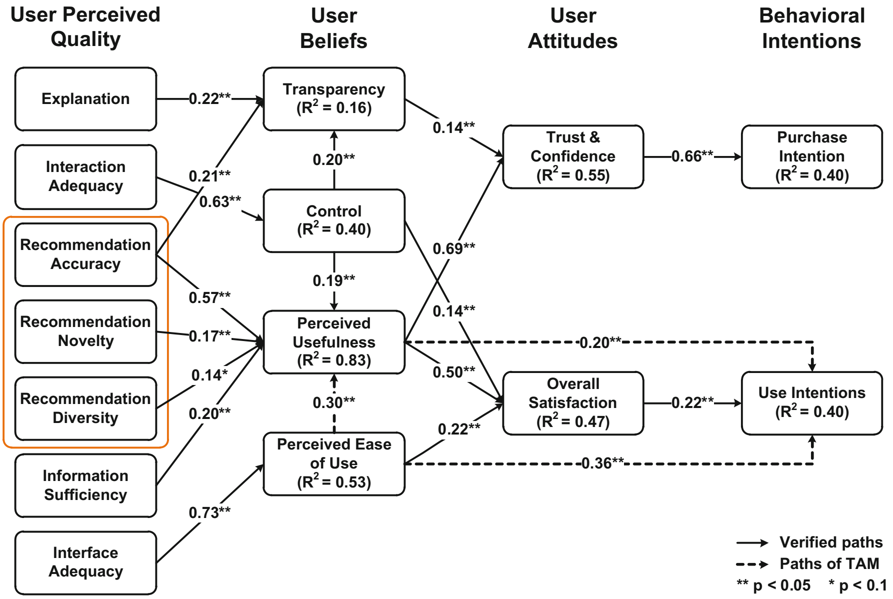


Fig. 5 Constructs of an evaluation framework on the perceived qualities of recommenders (*ResQue*) (Pu et al. 2011)

similarity, trust formation, technology acceptance model (TAM), and satisfaction. The majority of the 28 hypotheses were supported by corresponding empirical work. For example, RA use improved consumers' decision quality; the explicit preference elicitation method leads to better decision quality and higher decision effort than does the implicit preference elicitation method; recommending more alternatives increased information searched, decreased the quality of the consideration set, led to poor product choices, and reduced consumers' selectivity; ease of generating new or additional recommendations improved estimations of the ease of use of an RA; increased user control over interaction with the RA resulted in increased users' trust in and satisfaction; and explanations of an RA's inner logic strengthened users' trusting beliefs in the RA's competence and benevolence. The main limitation, as outlined by the authors, was the lack of results to support user behavioral intentions such as the intention to use RA in the future and the intention to purchase/accept the items recommended to them.

The reviewed work was also limited to existing RA methods that were employed in existing online environments. Innovative techniques experimented in research laboratories were not included. For example, recent work on trust-inspiring interfaces, critiquing based RAs which offer compensatory decision aid, and diversity-compatible recommenders were not examined, although these results are covered in this article. In addition, while the summary from Xiao and Benbasat (2007) provided recommendations relating features of RA or RA types to users' decision outcome and their evaluations of RAs, the 28 hypotheses did not provide suggestive design guidelines

Table 1 Summary of surveyed typical works and derived guidelines

Sections	Usability issues	Typical works	Tested recommender types and/or product domains	Guidelines
Preference elicitation	User effort	Haubl and Trifts (2000)	To recommenders that require an explicit preference elicitation	Guideline 1: reduce users' initial effort to the minimum
		Mahmood et al. (2010), Berger et al. (2007)	Tested in travel product domain	Guideline 2: make the elicitation process adaptive and entertaining
	What items to rate	Rashid et al. (2002)	Rating-based recommenders	Guideline 3: show a mixture of popular and item-to-item personalized items for users to rate initially
	System or user in control	McNee et al. (2003)	Rating-based recommenders	Guideline 4: let users to decide what to rate
	Elicitation of feature preferences	Shearin and Lieberman (2001), Pu and Chen (2005), Zanker and Jessenitschnig (2009)	Case-based and utility-based recommenders; applicable to inexperienced, high-risk products	Guideline 5: allowing users to indicate the relative importance of the features of a product/item they prefer
	Preference elicitation via personality questionnaire	Hu and Pu (2009a,b)	Personality-based recommenders; tested for movie and music	Guideline 6: consider using entertaining personality quizzes as a preference elicitation method
	Comparison of elicitation methods	Hu and Pu (2009a)	To all types	Guideline 7: conduct comparative user studies to identify the optimal elicitation method balancing between accuracy and effort
	User contribution	Beenen et al. (2004)	To all types	Guidelines 8 and 9: make users aware of their unique contributions, and set specific goals for users to achieve

Table 1 continued

Sections	Usability issues	Typical works	Tested recommender types and/or product domains	Guidelines
	Privacy issues	Kobsa and Schreck (2003), Lam et al. (2006), Knijnenburg et al. (2010)	To all types	Guideline 10: use explanation interfaces to increase system transparency and reduce privacy concerns
Preference refinement	Importance of preference refinement	Swearingen and Sinha (2002), Herlocker et al. (2004), Pu and Kumar (2004), Viappiani et al. (2007)	To all types	Guideline 11: provide preference refinement facilities such as critiquing
	Critiquing support	Chen and Pu (2009), Burke et al. (1997), Reilly et al. (2004), McCarthy et al. (2005), Pu and Chen (2005)	Tested in preference-based recommenders for high-risk products	Guidelines 12 and 13: return multiple recommended items for users to critique, and offer a hybrid-critiquing interface
Recommendation presentation	Accuracy	Knijnenburg et al. (2010), McNee et al. (2002), Cosley et al. (2003), Ochi et al. (2010)	To all types	Guideline 14: enhance accuracy with better interface design, such as labeling, structuring items, and explanation
	Familiarity	Swearingen and Sinha (2002)	Tested in rating-based recommenders for music	Guideline 15: include familiar items in the result set to increase trust
	Novelty	Swearingen and Sinha 2002, Jones and Pu (2007), Herlocker et al. (2004), McNee et al. (2006a)	Tested in content-based recommenders for entertainment products	Guideline 16: achieve a proper balance between familiarity and novelty

Table 1 continued

Sections	Usability issues	Typical works	Tested recommender types and/or product domains	Guidelines
	Diversity	Jones and Pu (2007), McNee et al. (2006a); Smyth and McClave (2001)	Tested in content-based, item-item CF, and case-based recommenders	Guideline 17: include diversity and variety in the result set
	Context compatibility	Adomavicius and Tuzhilin (2008)	Tested for context-related items such as movie, music, travel	Guideline 18: provide context compatible recommendations
	Explanation of recommendations	Herlocker et al. (2000), Sinha and Swearingen (2001, 2002), Simonson (2005), Tintarev and Masthoff (2007a,b), Pu and Chen (2006), Pu et al. (2009)	To all types and tested for both low-risk (e.g., music, movie) and high-risk products (e.g., laptops, digital cameras)	Guideline 19: explain why the system recommends the suggested items to increase transparency and explain tradeoffs to increase persuasion
	Sufficiency of information	Sinha and Swearingen (2002)	To all types and tested for books and songs	Guideline 20: provide sufficient information related to the recommendation items

on how to design and implement an effective RA. Therefore, the survey provided in this paper can be seen as a complimentary work.

7 Conclusion

User evaluation of recommender systems is a crucial subject of study. In this paper, we have provided a broad and in-depth review of existing research work in this area on three key interactions between the user and the system: the initial preference elicitation process, the preference refinement process, and the presentation of the system's recommendation results. We further synthesized the survey results into 20 usability and user interface design guidelines that can potentially help designers develop more effective and satisfying recommender systems. Most of the guidelines were derived based on their ability to balance the intrinsic tradeoffs between increasing user satisfaction, while taking into account their information processing effort, need to be in control, and concerns for privacy. Table 1 lists all of the surveyed works and the corresponding guidelines, as well as the types of recommenders to which these guidelines are most likely to apply.

From a conceptual point of view, this paper has advanced our understanding of the critical issues in the recommender technology field, especially relating to how users perceive and evaluate recommender systems. For the first time, various pieces of existing research work, concerning concrete design and development issues of RS adoption, have been brought together and examined to answer two important research questions: how users interact with recommender systems, and how their perceived system qualities motivate them in adopting such systems.

From a practical point of view, our design guidelines could enhance the usability engineering process of recommender systems and hopefully help practitioners achieve a wider adoption of their systems. Should additional validations be desired, this survey paper also provides detailed characteristics of evaluation methods on how to conduct such experiments.

Acknowledgments We would like to thank Jennifer Lott Graetzel for her valuable suggestions on the various drafts of this manuscript. We also like to thank our funding agencies, the Swiss National Science Foundation, the Chinese Ministry of Education, the Hong Kong Baptist University, and EPFL, for supporting this work.

References

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)

Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: The 2nd ACM Conference on Recommender Systems (RecSys '08), pp. 335–336. ACM, New York (2008)

Basartan, Y.: Amazon versus the shopbot: an experiment about how to improve the shopbots. Unpublished Ph.D. Summer Paper, Carnegie Mellon University, Pittsburgh, PA (2001)

Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., Kraut, R.E.: Using social psychology to motivate contributions to online communities. In: 2004 ACM Conference on Computer Supported Cooperative Work (CSCW '04), pp. 212–221. ACM, New York (2004)

Berger, H., Denk, M., Dittenbach, M., Merkl, D., Pesenhofer, A.: Quo Vadis Homo Turisticus? Towards a picture-based tourist profiler. Inf. Commun. Technol. Tour. **2**, 87–96 (2007)

Bollen, D.G.F.M., Knijnenburg, B.P., Willemsen, M.C., Graus, M.P.: Understanding choice overload in recommender systems. In: The 4th ACM Conference on Recommender Systems (RecSys'10), pp. 63–70. ACM, New York (2010)

Brodie, C., Karat, C.M., Karat, J.: Creating an E-commerce environment where consumers are willing to share personal information. In: Karat, C., Blom, J.O., Karat, J. (eds.) Designing Personalized User Experiences in eCommerce, pp. 185–206. Springer, Netherlands (2004)

Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User Adapt. Interact. **12**(4), 331–370 (2002)

Burke, R., Hammond, K., Young, B.: The FindMe approach to assisted browsing. IEEE Expert Intell. Syst. Appl. **12**(4), 32–40 (1997)

Chen, L., Pu, P.: Preference-based organization interface: aiding user critiques in recommender systems. In: International Conference on User Modeling (UM'07), Corfu, Greece, 25–29 June, pp. 77–86 (2007)

Chen, L., Pu, P.: Interaction design guidelines on critiquing-based recommender systems. User Model. User Adapt. Interact. **19**(3), 167–206 (2009)

Chen, L., Pu, P.: Experiments on the preference-based organization interface in recommender systems. ACM Trans. Comput. Hum. Interact. **17**(1), 1–33 (2010)

Cosley, D., Lam, S.K., Albert, I., Konstan, J.A., Riedl, J.: Is seeing believing? How recommender system interfaces affect users' opinions. In: SIGCHI Conference on Human Factors in Computing Systems (CHI '03), pp. 585–592. ACM, New York (2003)

Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. **13**, 319–339 (1989)

Drenner, S., Sen, S., Terveen, L.: Crafting the initial user experience to achieve community goals. In: 2008 ACM Conference on Recommender Systems (RecSys '08), pp. 187–194. ACM, New York (2008)

Dunn, G., Wiersema, J., Ham, J., Aroyo, L.: Evaluating interface variants on personality acquisition for recommender systems. In: Houben, G.-J., McCalla, G., Pianesi, F., Zancanaro, M. (eds.) User Modeling, Adaptation, and Personalization, vol. 5535, pp. 259–270. Springer-Verlag, Berlin (2009)

Einhorn, H., Hogarth, R.: Confidence in judgment: persistence of the illusion of validity. Psychol. Rev. **85**, 395–416 (1978)

Guttman, R.H.: Merchant differentiation through integrative negotiation in agent-mediated electronic commerce. Master's Thesis, School of Architecture and Planning, Program in Media Arts and Sciences, Massachusetts Institute of Technology (1998)

Haubl, G., Trifts, V.: Consumer decision making in online shopping environments: the effects of interactive decision aids. Mark. Sci. **19**, 4–21 (2000)

Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00), pp. 241–250. ACM, New York (2000)

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

Hu, R., Pu, P.: A comparative user study on rating vs. personality quiz based preference elicitation methods. In: The 14th International Conference on Intelligent User Interfaces (IUI '09), 8–11 February, pp. 367–372. ACM, Sanibel Island (2009a)

Hu, R., Pu, P.: Acceptance issues of personality-based recommender systems. In: The 3rd ACM Conference on Recommender Systems (RecSys 2009), 22–25 October, pp. 221–224. ACM, New York (2009b)

Hu, R., Pu, P.: A study on user perception of personality-based recommender systems. In: De Bra, P., Kobsa, A., Chin, D. (eds.) User Modeling, Adaptation, and Personalization, LNCS 6075, pp. 291–302. Springer, Heidelberg (2010)

Hu, R., Pu, P.: Enhancing recommendation diversity with organization interfaces. In: the 16th International Conference on Intelligent user Interfaces (IUI '11), pp. 347–350. ACM, New York (2011)

Jones, N., Pu, P.: User technology adoption issues in recommender systems. In: Networking and Electronic Commerce Research Conference (NAEC '07), pp. 379–394 (2007)

Karau, S.J., Williams, K.D.: Understanding individual motivation in groups: the collective effort model. In: Turner, M.E. (ed.) Groups at Work: Theory and Research, pp. 113–141. LEA, Mahwah (2001)

Kirakowski, J.: SUMI: the software usability measurement inventory. Br. J. Educ. Technol. **24**(3), 210–214 (1993)

Kleinmuntz, D.N., Schkade, D.A.: Information displays and decision processes. Psychol. Sci. **4**, 221–227 (1993)

Knijnenburg, B.P., Willemsen, M.C., Hirtbach, S.: Receiving recommendations and providing feedback: the user-experience of a recommender system. In: The 11th International Conference on Electronic Commerce and Web Technologies, pp. 207–216 (2010)

Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C.: Explaining the user experience of recommender systems. User Model. User Adapt. Interact. **22** (2012). doi:10.1007/s11257-011-9118-4

Kobsa, A., Schreck, J.: Privacy through pseudonymity in user-adaptive systems. ACM Trans. Internet Technol. **3**(2), 149–183 (2003)

Lam, S.K., Frankowski, D., Riedl, J.: Do you trust your recommendations? An exploration of security and privacy issues in recommender systems. In: 2006 International Conference on Emerging Trends in Information and Communication Security (ETRICS), Freiburg, Germany, pp. 14–29 (2006)

Linden, G., Hanks, S., Lesh, N.: Interactive assessment of user preference models: the automated travel assistant. In: The Sixth International Conference on User Modeling, pp. 67–78. Springer, Chia Laguna (1997)

Locke, E.A., Latham, G.P.: Building a practically useful theory of goal setting and task motivation: a 35 year odyssey. Am. Psychol. **57**(9), 705–717 (2002)

Mahmood, T., Ricci, F., Venturini, A.: Improving recommendation effectiveness by adapting the dialogue strategy in online travel planning. Int. J. Inf. Technol. Tour. **11**(4), 285–302 (2010)

McCarthy, K., Reilly, J., McGinty, L., Smyth, B.: Thinking positively—explanatory feedback for conversational recommender systems. In: European Conference on Case-Based Reasoning (ECCBR-04) Explanation Workshop, Madrid, Spain, pp. 115–124 (2004)

McCarthy, K., Reilly, J., McGinty, L., Smyth, B.: Experiments in dynamic critiquing. In: The 10th International Conference on Intelligent User Interfaces (IUI '05), pp. 175–182. ACM, New York (2005)

McGinty, L., Smyth, B.: On the role of diversity in conversational recommender systems. In: The Fifth International Conference on Case-Based Reasoning, pp. 276–290. Springer, Berlin (2003)

McNee, M.S., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A., Riedl, J.: On the recommending of citations for research papers. In: 2002 ACM Conference on Computer Supported Cooperative Work (CSCW '02), pp. 116–125. ACM, New York (2002)

McNee, S.M., Lam, S.K., Konstan, J.A., Riedal, J.: 2003. Interfaces for eliciting new user preferences in recommender systems. In: User Modeling 2003, pp. 178–187 (2003)

McNee, S.M., Riedl, J., Konstan, J.A.: Being accurate is not enough: how accuracy metrics have hurt recommender systems. In: CHI Extended Abstracts, pp. 1097–1101 (2006a)

McNee, S.M., Riedl, J., Konstan, J.A.: Making recommendations better: an analytic model for human-recommender interaction. In: CHI Extended Abstracts, pp. 1103–1108 (2006b)

McSherry, D.: Explanation in recommender systems. Artif. Intell. Rev. **24**(2), 179–197 (2005)

Ochi, P., Rao, S., Takayama, L., Nass, C.: Predictors of user perceptions of web recommender systems: How the basis for generating experience and search product recommendations affects user responses. Int. J. Hum. Comput. Stud. **68**(8), 472–482 (2010)

Paramythis, A., Weibelzahl, S., Masthoff, J.: Layered evaluation of interactive adaptive systems: framework and formative methods. User Model. User Adapt. Interact. **20**(5), 383–453 (2010)

Payne, J.W., Bettman, J.R., Schkade, D.A.: Measuring constructed preference: towards a building code. J. Risk Uncertain. **19**(1–3), 243–270 (1999)

Pu, P., Chen, L.: Integrating tradeoff support in product search tools for e-commerce sites. In: The 6th ACM Conference on Electronic Commerce (EC '05), pp. 269–278. ACM, New York (2005)

Pu, P., Chen, L.: Trust building with explanation interfaces. In: The 11th International Conference on Intelligent User Interfaces (IUI '06), pp. 93–100. ACM, New York (2006)

Pu, P., Chen, L.: Trust-inspiring explanation interfaces for recommender systems. Knowl-Based. Syst. **20**(6), 542–556 (2007)

Pu, P., Chen, L.: A user-centric evaluation framework of recommender systems. In: Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces (UCERSTI' 10), ACM Conference on Recommender Systems (RecSys' 10), Barcelona, Spain, pp. 14–21 (2010)

Pu, P., Faltings, B.: Decision tradeoff using example-critiquing and constraint programming. Constraints Int. J. **9**(4), 289–310 (2004)

Pu, P., Kumar, P.: Evaluating example-based search tools. In: The 5th ACM Conference on Electronic Commerce (EC '04), pp. 208–217. ACM, New York (2004)

Pu, P., Viappiani, P., Faltings, B.: Increasing user decision accuracy using suggestions. In: Grinter, R., Rodden, T., Aoki, P., Cutrell, E., Jeffries, R., Olson, G. (eds.) The SIGCHI Conference on Human Factors in Computing Systems (CHI '06), ACM, New York (2006)

Pu, P., Zhou, M., Castagnos, S.: Critiquing recommenders for public taste products. In: The Third ACM Conference on Recommender Systems (RecSys '09), pp. 249–252. ACM, New York (2009)

Pu, P., Faltings, B., Chen, L., Zhang, J.Y., Viappiani, P.: Usability guidelines for product recommenders based on example critiquing research. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, Chapter 16, pp. 511–546. Springer (2010)

Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In: The 5th ACM Conference on Recommender Systems (RecSys'11), Chicago, IL, USA, 23–27 October (2011)

Rashid, A.M., Albert, I., Cosley, D., Lam, S.K., McNee, S.M., Konstan, J.A., Riedl, J.: Getting to know you: learning new user preferences in recommender systems. In: The 7th International Conference on Intelligent User Interfaces (IUI '02), pp. 127–134. ACM, New York (2002)

Reilly, J., McCarthy, K., McGinty, L., Smyth, B.: Dynamic critiquing. In: Funk, P., González Calero, P.A. (eds.) Advances in Case-Based Reasoning (ECCBR 2004), LNAI 3155, pp. 763–777. Springer, Heidelberg (2004)

Resnick, P., Varian, H.R.: Recommender systems. Commun. ACM **40**, 56–58 (1997)

Riedl, J.: Personalization and privacy. IEEE Internet Comput. **5**(6), 29–31 (2001)

Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Analysis of recommendation algorithms for e commerce. In: The 2nd ACM Conference on Electronic Commerce (EC '00), pp. 158–167. ACM, New York (2000)

Sarwar, B., Karypis, G., Konstan, J., Riedl, J. Item-based collaborative filtering recommendation algorithms. In: WWW'01, pp. 285–295 (2001)

Shearin, S., Lieberman, H.: Intelligent profiling by example. In: The 6th International Conference on Intelligent User Interfaces (IUI '01), pp. 145–151. ACM, New York (2001)

Simonson, I.: Determinants of customers' responses to customized offers: conceptual framework and research propositions. J. Mark. **69**, 32–45 (2005)

Sinha, R., Swearingen, K.: Comparing recommendations made by online systems and friends. In: DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries (2001)

Sinha, R., Swearingen, K.: The role of transparency in recommender systems. In: CHI Extended Abstracts, pp. 830–831(2002)

Smyth, B.: Case-based recommendation. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web: Methods and Strategies of Web Personalization, Lecture Notes in Computer Science, vol. 4321. Springer-Verlag, Berlin (2007)

Smyth, B., McClave, P.: Similarity vs. diversity. In: Aha, D.W., Watson, I. (eds.) The 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development (ICCBR '01), pp. 347–361. Springer-Verlag, London (2001)

Spiekermann, S., Parachiv, C.: Motivating human-agent interaction: transferring insights from behavioral marketing to interface design. J. Electron. Commer. Res. **2**(3), 255–285 (2002)

Spiekermann, S., Grossklags, J., Berendt, B.: E-privacy in 2nd generation E-commerce: privacy preferences versus actual behavior. In: The 3rd ACM Conference on Electronic Commerce, pp. 38–47. ACM, New York (2001)

Swearingen, K., Sinha, R.: Interaction design for recommender systems. In: Designing Interactive Systems (DIS'02), London, 25–28 June (2002)

Tintarev, N., Masthoff, J.: Effective explanations of recommendations: user-centered design. In: 2007 ACM Conference on Recommender Systems (RecSys '07), pp. 153–156. ACM, New York (2007a)

Tintarev, N., Masthoff, J.: A survey of explanations in recommender systems. In: The 23rd IEEE International Conference on Data Engineering Workshop, pp. 801–810 (2007b)

Viappiani, P., Faltings, B., Pu, P.: Evaluating preference-based search tools: a tale of two approaches. In: The Twenty-First National Conference on Artificial Intelligence (AAAI-06), Boston, USA, 16–20 July, pp. 205–210 (2006)

Viappiani, P., Faltings, B., Pu, P.: Preference-based search using example-critiquing with suggestions. J. Artif. Intell. Res. **27**, 465–503 (2007)

Vig, J., Sen, S., Riedl, J.: Tagsplanations: explaining recommendations using tags. In: The 13th International Conference on Intelligent User Interfaces (IUI'09), pp. 47–56. ACM, New York (2009)

Williams, M.D., Tou, F.N.: RABBIT: an interface for database access. In: ACM '82 Conference (ACM '82), pp. 83–87. ACM, New York (1982)

Xiao, B., Benbasat, I.: Ecommerce product recommendation agents: use, characteristics, and impact. MIS Q. **31**(1), 137–209 (2007)

Zanker, M., Jessenitschnig, M.: Case-studies on exploiting explicit customer requirements in recommender systems. User Model. User Adapt. Interact. **19**(1–2), 133–166 (2009)

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

Zukerman, I., Albrecht, D.W.: Predictive statistical models for user modeling. User Model. User Adapt. Interact. **11**(1–2), 5–18 (2001)

Author Biographies

Pearl Pu currently leads the HCI Group in the School of Computer and Communication Sciences at the Swiss Federal Institute of Technology in Lausanne (EPFL). She obtained her Master and Ph.D. degrees from the University of Pennsylvania in artificial intelligence and computer graphics. Her research is multi-disciplinary and focuses on issues in the intersection of human computer interaction, artificial intelligence, and behavioral decision theories. Her work was primarily concerned with the design and validation of user interface and interaction techniques for decision support and product recommender systems in e-commerce and social media environments. Together with her colleagues and students, she has developed the EC systems, semantic fisheye views, and organized interfaces for product recommenders. Her current research interests include understanding consumer decision behaviors in new media environments, designing trust-inspiring interfaces for recommender agents, and developing forecasting models for user technology adoption. She is Associate Editor for the IEEE Transactions on Multimedia and UMUAI, the personalization journal.

Li Chen is an Assistant Professor at Hong Kong Baptist University. She obtained her Ph.D. degree in Human Computer Interaction at Swiss Federal Institute of Technology in Lausanne (EPFL), and Bachelor and Master Degrees in Computer Science at Peking University, China. Her research interests are mainly in the areas of human–computer interaction, user-centered development of recommender systems and e-commerce decision supports. Her co-authored papers have been published in journals and conferences on e-commerce, artificial intelligence, intelligent user interfaces, user modeling, and recommender systems.

Rong Hu is a Ph.D. candidate in Human Computer Interaction at Swiss Federal Institute of Technology in Lausanne (EPFL). She received her B.A. in Computer Science from Shanghai Jiaotong University, China, and her M.S. degree in the same field from Fudan University, China. Her primary interests lie in the areas of human–computer interaction, recommender systems, data mining and artificial intelligence. The joint research with Dr. Pearl Pu and Dr. Li Chen described in this article reflects an interest in developing a framework of evaluating recommender systems from the user's perspective.