A User-Centric Evaluation Framework of Recommender Systems

Article · January 2011

CITATIONS
45

READS
506

2 authors:

Pearl Pu
École Polytechnique Fédérale de Lausanne
175 PUBLICATIONS **4,586** CITATIONS
SEE PROFILE

Li Chen
Hong Kong Baptist University
90 PUBLICATIONS **3,845** CITATIONS
SEE PROFILE

User Evaluation Framework of Recommender Systems

Li Chen
Department of Computer Science
Hong Kong Baptist University, Hong Kong
lichen@comp.hkbu.edu.hk

Pearl Pu
Human Computer Interaction Group
Swiss Federal Institute of Technology in Lausanne
pearl.pu@epfl.ch

ABSTRACT
This paper explores the evaluation issues of recommender systems particularly from users' perspective. We first show results of literature surveys on human psychological decision theory and trust building in online environments. Based on the results, we propose an evaluation framework aimed at assessing a recommender's practical ability in providing decision support benefits to end-users from various aspects. It includes both accuracy/effort measures and a user-trust model of subjective constructs, and a corresponding sample questionnaire design.

Author Keywords
Recommender systems, user evaluation, adaptive decision theory, trust building, decision accuracy and effort.

ACM Classification
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Recommender systems emerged as an independent research area since the appearance of papers on "collaborative filtering" in the mid-1990s to resolve the recommendation problem [54]. The automated collaborative filtering (ACF) originated as an information filtering technique that used group opinions to recommend information items to individuals. For instance, the user will be recommended items that people with similar tastes and preferences liked in the past. Various collaborative algorithms based on data mining and machine learning techniques (e.g. K-nearest neighbor, clustering, classifier learning) have been developed to reach the goal. A typical application is MovieLens that predicts the attractiveness of an unseen movie for a given user based on a combination of the rating scores derived from her nearest neighbors [46]. At Amazon.com, the "people who bought this book also bought" was also one example of the commercial adoptions of this technology. Recently, Bonhard et al. showed ways to improve the user-user collaborative filtering techniques by including information on the demographics similarity [8].

Workshop SRS'10, February 7, 2010, Hong Kong, China.

In the case that relationship among products is stronger than among customers, content-based recommender methods have been often used to compute the set of items that are similar to what the user has preferred in the past [1]. For example, Pandora, an online music recommender tool, can suggest a sequence of music the user would probably like according to the features (e.g. genre, musician) of ones that she has indicated her preferences on.

Another branch of recommender systems, called preference-based or knowledge-based systems, has been mainly oriented for high-involvement products with well-defined features (such as computers, houses, cars), for which selection a user is willing to spend considerable effort in order to avoid any financial damage [61, 52]. In such systems, a preference model is usually explicitly established for each user.

Researchers have previously indicated the challenges for different types of recommenders. For example, as for the collaborative system, its main limitations are new user problem (i.e. a new user having very few ratings would not be able to get accurate recommendations), new item problem (i.e. until the new item is rated by a substantial number of users, the system would not be able to recommend it), and sparsity (i.e. the number of ratings is very small compared to the number of ratings that need to be predicted) [1]. In order to address these problems, the *hybrid* recommendation approach combining two or more techniques (the combination of content-based and collaborative filtering) has been increasingly explored [9]. Recently, advanced techniques that involve more types of social resources such as tags and social ties (e.g., friendship or membership) have also emerged in order to improve the similarity accuracy between users or items, and classified into a new branch called social recommender systems [27, 63,65].

However, few studies have stood from users' angles to consider their cognitive acceptance of recommendations. Moreover, the question is how to evaluate a recommender in terms of its actual impacts on empowering users to make better decisions, except mathematical algorithm accuracy. In the following, we will first show literature reviews on decision theory from the psychology domain to understand users' decision making heuristics, given that the recommender is inherently a decision support to assist users in making choices. Furthermore, the user-trust building issues that have been promoted in online environments will be discussed and related to specific research questions to recommenders.

ADAPTIVE DECISION MAKER

The goal of a decision support system is to aid the user in making an informed decision consistent with her objectives. The elicitation of user preferences is fundamental for the recommender system to generate products or services that may interest its users. Most of preference elicitation procedures in recent recommender systems can be classified into two main technologies: *implicit preference elicitation* which has aimed to infer user preferences according to her demographic data, personality, past navigation and purchase behavior, tags, and so on [38, 8]; and *explicit preference elicitation* that has emphasized on explicitly asking for the user's preferences during interaction, such as her rating on an item (in collaborative filtering systems) or stating value functions over item features (in utility-based systems). However, recommender systems, that simply depend on initially obtained user preferences to predict recommendations, may not help the user make an accurate decision.

According to the adaptive decision theory [50], user preferences are inherently adaptive and constructive depending on the current decision task and environment, and hence their initial preferences can be uncertain and erroneous. They may lack the motivation to answer demanding initial elicitation questions prior to any perceived benefits [59], and they may not have the domain knowledge to answer the questions correctly.

As a matter of fact, in the last four decades, the classical decision theory has evolved into two conceptual shifts. One shift is the discovery of adaptive and constructive nature of human decision making. Individuals have several decision strategies at their disposal and when faced with a decision they select a strategy depending on a variety of factors related to the task, the context, and individual differences. Additional studies indicated that individuals often do not possess well-defined preferences on many objects and situations, but construct them in a highly context-dependent fashion during the decision process [62,51].

Another shift has occurred in the field of prescriptive decision making and it is called *value-focused thinking* [35], different from the traditional attribute-focused thinking. In this approach, once a decision problem is recognized, fundamental and relevant values are first identified to creatively identify possible alternatives and to carefully assess their desirability [10].

Based on the two shifts, researchers in areas of decision theory have identified the following typical phenomena that may occur in a person's adaptive decision process.

Context-dependent preferences. An important implication of the constructive nature of preferences is that decisions and decision processes are highly contingent upon a variety of factors characterizing decision problems. First, choice among options is context (or menu) dependent. The relative value of an option depends not only on the characteristics of that option, but also upon characteristics of other options in the choice set. For example, the relative attractiveness of *x* compared to *y* often depends on the presence or absence of a third option *z* [62]. Second, preference among options also depends upon how the valuation question is asked. Strategically equivalent methods for eliciting preferences can lead to systematically different preference orderings. Third, choice among options depends upon how the choice set is represented (framed) or displayed. Finally, the process used to make a choice depends on the complexity of the decision tasks: the use of simple decision heuristics increases with task complexity [51].

Four decision metagoals. Evidence from behavioral studies indicates four main metagoals driving human decision making. Although individuals clearly aim at *maximizing the accuracy* of their decisions, they are often willing to tradeoff accuracy to *reduce cognitive effort*. Also, because of their social and emotional nature, when making a decision, people try to *minimize/maximize negative/positive emotions* and *maximize the ease of justifying a decision* [5]. When faced with a decision, people make critical assessments of the four metagoals contingent on the decision task (e.g. number of alternatives) and the decision environment (e.g. how information is presented to the decision maker). Especially in unfamiliar and complex decision conditions, decision makers reassess the metagoals and switch from one strategy to another as they learn more about the task structure and the environment during the course of decision making [50].

Anchoring effect. Researchers also suggested that people use an anchor-and-adjust strategy to solve a variety of estimation problems. For example, when asked questions about information that people do not know, they may spontaneously anchor on information that comes to mind and adjust their responses in a direction that seems appropriate [34]. This heuristic is helpful, but the final estimate might be biased toward the initial anchor value [19].

Tradeoff avoidance. Decision problems often involve conflict among values, because no one option is best on all attributes of values, and conflict has long been recognized as a major source of decision difficulty [56]. Thus, many researchers argued that making tradeoffs between more of one thing and less of another is a crucial aspect of high-quality and rational decision making [21]. However, decision makers often avoid explicit tradeoffs, relying instead on an array of non-compensatory decision strategies [49]. The explanation for tradeoff avoidance is that tradeoffs can be difficult for emotional as well as cognitive reasons [31, 40].

Means objectives. According to value-focused thinking (VFT), the decision maker should qualitatively distinguish between *fundamental* and *means* objectives. Fundamental objectives should reflect what the decision maker really wants to accomplish with a decision, while means objectives simply help to achieve other objectives [36].

However, inadequate elicitation questions can easily circumscribe a user in thinking about means objectives rather than fundamental objectives. For example, a traveler lives near Geneva and wants to be in Malaga by 3:00 pm (her fundamental objective), but if she was asked to state departure time first, she would have to formulate a means objective (i.e. departure at 10:00 am), even though there is a direct flight that leaves at 2:00 pm.

Therefore, as suggested in [51], metaphorically speaking, preference elicitation is best viewed as architecture (building a set of values) rather than archeology (uncovering existing values). In order to avoid human decision biases, preference elicitation tools must attempt to quickly collect as much preference data as possible so that users can begin working towards their goals. Furthermore, they must also be able to resolve potential conflicting preferences, discover hidden preferences, and make reasonable decisions about tradeoffs with competing user goals.

Unfortunately, most of current recommender system designs did not recognize the importance of these implications. In order to help the user make an accurate and confident decision, we have been mainly engaged to realize a decision aid that can embody all of the requirements. In addition, by means of user experience research, we have attempted to derive more useful principles for the development of an intelligent and adaptive preference-based recommender system.

TRUST BUILDING IN ONLINE ENVIRONMENTS

The second challenge is about how to build user trust in recommender systems. Less attention has been paid in related work to evaluating and improving the recommender system from the aspect of users' subjective attitudes. Among the many factors, the perception of the recommender's trustworthiness would be most prominent as it facilitates long-term relationship and encourages potential repeat interactions and purchases [22, 17].

Trust has been in nature regarded as a key factor to the success of e-commerce [23]. Due to the lack of face-to-face interaction with consumers in online environments, users' actions undertake a higher degree of uncertainty and risk than in traditional settings. As a result, trust is indeed difficult to build and easy to lose with the virtual store, which has impeded customers from actively participating in e-commerce environments [33].

The definition of trust has varied from study to study. The most frequently cited definition of trust in various contexts is the "willingness to be vulnerable" proposed by Mayer et al. [42]. Adapting from this definition, Chopra and Wallace defined trust in the electronic environment as the "willingness to rely on a specific other, based on confidence that one's trust will lead to positive outcomes." [15] More specifically, consumer trust in online shopping was defined as "the willingness of a consumer to expose himself/herself to the possibility of loss during an Internet shopping transaction, based on the expectation that the merchant will engage in generally acceptable practices, and will be able to deliver the promised products or services." [39]

As these definitions indicate, consumer trust is essentially leading to behavioral intentions [24], referred as "trusting intentions" by McKnight et al. [45]. Consistent with the Theory of Planned Behavior [2], consumer trust (as a belief) will influence customer intentions. Empirical studies have shown that trust in an e-commerce website increases customer intention to purchase a product from the website, as well as intention to return to it for future use. Other potential trusting intentions include providing personal information (email, phone number and credit card number) and continuing to transact with the website [26].

Many researchers have also experimentally investigated the antecedents of on-line trust. For example, Pavlou and Chellappa explained how perceived privacy and perceived security promote trust in e-commerce transactions [48]. De Ruyter et al. examined the impact of organizational reputation, relative advantage and perceived risk on trust in e-service and customer behavior intentions [55]. Jarvenpaa et al. validated that the perceived size of an Internet store and its perceived reputation are positively related to consumers' initial trust in the store [33].

The effect of experience with website interface on trust formation has been also investigated based on the Technology Acceptance Model (TAM) [16,68]. TAM has long been considered a robust framework for understanding how users develop attributes towards technology and when they decide to adopt it. It posits that intention to voluntarily accept and use a new information technology (IT) is determined by two beliefs: the perceives usefulness of using the new IT, and the perceived ease of use of the new IT. According to TAM, Koufaris and Hampton-Sosa established a trust model and demonstrated that both the perceived usefulness and the perceived ease of use of the website are positively associated with customer trust in the online company and customer' intentions to purchase and return [37]. Gefen et al. expanded TAM to include a familiarity and trust aspect of e-commerce adoption, and found that repeat customers' purchase intentions were influenced by both their trust in the e-vendor and their perceived usefulness of the website, whereas potential customers were only influenced by their trust [25]. Hassanein and Head identified the positive influence of social presence on customers' perceived usefulness of an e-commerce website and their trust in the online vendor [29].

In the domain of recommender systems, trust value has been also noticed but it has been mainly used to empower the prediction of user interests, especially for the collaborative filtering (CF) systems [32]. For instance, O'Donovan and Smyth have proposed a method to incorporate the trustworthiness of partners into the standard computation process in CF frameworks in order to increase the predictive accuracy of recommendations [47]. Similarly,

Massa and Bhattacharjee developed a trust-aware technique taking into account the "web of trust" provided by each user to estimate the relevance of users' tastes in addition to similarity measure [41]. Few literatures have highlighted the importance of user trust in recommender systems and proposed effective techniques to achieve it. The studies done by Swearingen and Sinha showed the positive role of transparency, familiarity of the recommended items and the process for receiving recommendations in trust achievement [60]. Zimmerman and Kurapati described a method of exposing the reflective history in user interface to increase user trust in TV recommender [66].

However, the limitations are that there is still lack of in-depth investigations of the concrete system design features that could be developed to promote user trust, and lack of empirical studies to measure real-users' trust formation and the influential constructs that could be most contributive to users' behavioral intentions in a recommender system.

Considering these limitations, our main objective is to explore the crucial antecedents of trustworthiness for recommender systems and their exact nature in providing benefits to users. Concretely, driven by the above decision theory findings and trust issues, we have built an evaluation framework aimed at including all of crucial standards to assess a recommender's true ability.

EVALUATION FRAMEWORK

As a matter of fact, identifying the appropriate criteria for evaluating the true benefits of a recommender system is a challenging issue. Most of related user studies purely focused on users' objective performance such as their interaction cycles and task completion time [43], less on decision accuracy that the user can eventually achieve, and subjective effort that the user cognitively perceived in processing information. Moreover, as mentioned above, the consumer trust should be also included as a key standard, such as whether the recommender could significantly help to increase users' competence-inspired trust and furthermore their behavioral intention to purchase a product or intention to return to it for repeated uses.

Decision Accuracy and Decision Effort

According to [50], two key considerations underlying a user's decision strategy selection are: the *accuracy* of a strategy in yielding a "good" decision, and the "*cognitive effort*" required of a strategy in making a decision. All else being equal, decision makers prefer more accurate choices and less effortful choices. Unfortunately, strategies yielding more accurate choices are often more effortful (such as weighted additive rule), and easy strategies can sometimes yield lower levels of accuracy (e.g. elimination-by-aspects). Therefore, they view strategy selection to be the result of a compromise between the desire to make the most correct decision and the desire to minimize effort. Typically, when alternatives are numerous and difficult to compare, like when the complexity of the decision environment is high, decision makers are usually willing to settle for imperfect accuracy of their decisions in return for a reduction in effort. The observation is well supported by [6, 57] and consistent with the idea of bounded rationality [58].

A standard assumption in past research on decision support systems is that decision makers who are provided with decisions aids that have adequate information processing capabilities will use these tools to analyze problems in greater depth and, as a result, make better decisions [30, 28]. However, empirical studies also showed that because feedback on effort expenditure tends to be immediate while feedback on accuracy is subject to delay and ambiguity, the use of decision aids does not necessarily enhance decision making quality, but merely leads individuals to reduce effort [18, 4].

Given this mixed evidence, it cannot be assumed that the use of interaction decision aids will definitely enhance users' decision quality. Thus, an open question to recommender systems is that whether they could enable users to reach the optimal level of accuracy under the acceptable amount of effort users are willing to exert during their interaction with the system. In the following, we introduce our accuracy-effort measurement model, derived from the ACE (Accuracy, Confidence, Effort) framework that we have previously built for preference-based product recommenders [67]. Decision accuracy and decision effort are respectively evaluated from both objective and subjective dimensions and their tradeoff relationship is also included as shown in Figure 1.

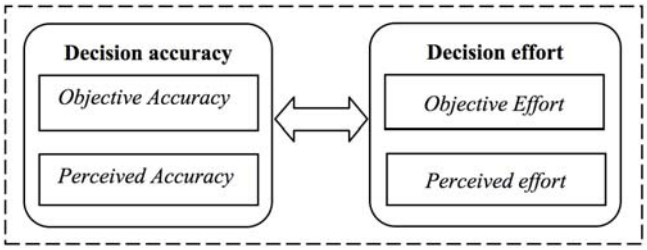

Figure 1. The accuracy and effort measurement model.

Objective and Perceived Decision Accuracy

In related work, decision accuracy has been measured adaptive to different experimental situations or purposes. In Payne et al.'s simulations, the accuracy of a particular heuristic strategy was defined by comparing its produced choice against the standard of a normative model like the weighted additive rule (WADD) [50]. The performance measures of *precision* and *recall* have been commonly applied to test an information retrieval system's accuracy based on a set of ground truths (previously collected items that are relevant the user's information need) [7]. In the condition of user experience researches, Haubl and Trifts suggested three indicators of a user's decision quality: increased probability of a non-dominated alternative selected for purchase, reduced probability of switching to another alternative after making the initial purchase decision, and a higher degree of confidence in purchase decisions [28]. In our case, we considered two facets: objective decision accuracy and perceived accuracy.

Objective Decision Accuracy. It is defined as the quantitative accuracy a user can eventually achieve by using the assigned decision system to make a choice. More specifically, it can be measured by the fraction of participants whose final option found with the decision tool agrees with the target option that they find after reviewing all available options in an offline setting. This procedure is known as the switching task. Switching refers to whether a user switches to another choice of product after reviewing all products instead of standing by the choice made with the tool. In our previous experiments [11,53], the "switching" task was supported by both sorting and comparison facilities. Subjects were encouraged to switch whenever they saw an alternative they preferred over their initial choice.

A lower switching fraction, thus, means that the decision system allows higher decision accuracy since most users are able to find their best choice with it. On the contrary, a higher switching fraction implies that the system is not very capable of guiding users to obtain what they truly want. For expensive products, such inaccurate tools may cause both financial damage and emotional burden to a decision maker.

Perceived Accuracy. Besides objective accuracy, it is also valuable to measure the degree of accuracy users subjectively perceived while using the system, which is also called *decision confidence* in some literatures [52]. The confidence judgment is important since it would be likely associated with users' competence perception of the system or even their intention to purchase the chosen product. The variable is concretely assessed either by asking subjects to express any opinions on the interface or directly requiring them to rate a statement like "I am confident that the product I just 'purchased' is really the best choice for me" on a Likert scale ranging from "strongly disagree" to "strongly agree".

Objective and Perceived Decision Effort

According to the accuracy-effort framework [50], another important criterion of evaluating a decision system's benefit is the amount of decision effort users expend to make their choice. So far, the most common measure appearing in related literatures is the number of interaction cycles or task time that the user actually took while using the tool to reach an option that she believes to be the target option. For example, *session length* (the number of recommendation cycles) was regarded as an importance factor of distinguishing the Dynamic Critiquing system with its compared work like FindMe interfaces [43]. In our model, we not only care about how much objective effort users actually consumed, but also their perceived cognitive effort, which we hope would indicate the amount of subjective effort people exert.

Objective Effort. The objective effort is concretely reflected by two dimensions: the task completion time and the interaction effort. The interaction effort was either simply defined as the total interaction cycles users were involved, or divided into more detailed constructs if they were necessary to indicate an average participant's effort distribution. For instance, in an online shopping setting, the interaction effort may be consumed in browsing alternatives, specifying filtering criteria, viewing products' detailed information, putting multiple products into a consideration set, and so on. Such effort components were also referred to Elementary Information Processes (EIPs) for a decision strategy's effort decomposition [50,64].

Perceived Cognitive Effort. Cognitive decision effort indicates the psychological cost of processing information. It represents the ease with which the subject can perform the task of obtaining and processing the relevant information in order to enable her to arrive at her decision. Normally, two or more scale items (e.g. "I easily found the information I was looking for") can be used to measure the construct *perceived effort*. The respondents were told to mark each of items on a Likert scale ranging from "Strongly Disagree" to "Strongly Agree".

Trust Model for Recommender Systems

As indicated before, trust is seen as a long term relationship between a user and the organization that the recommender system represents. Therefore, trust issues are critical to study especially for recommender systems used in e-commerce where the traditional salesperson, and subsequent relationship, is replaced by a product recommender agent. Studies showed that customer trust is positively associated with customers' intention to transact, purchase a product, and return to the website [33]. These results have mainly been derived from online shops' ability to ensure security, privacy and reputation, i.e., the integrity and benevolence aspects of trust formation, and less from a system's competence such as a recommender system's ability to explain its result.

These open issues led us to develop a trust model for building user trust in recommender systems, especially focusing on the role of competence constructs. The term "trust" is theoretically defined by a combination of trusting beliefs and trusting intentions, in accordance with the Theory of Planned Behavior (TPB) asserting that behavior is influenced by behavior intention and that intention is determined by attitudes and beliefs [2]. So we first introduce TPB and Technology Acceptance Model, based on which our trust model has been established.

Theory of Planned Behavior

In psychology, the theory of planned behavior (TPB) is a theory about the link between attitudes and behavior. It was proposed by Icek Ajzen as an extension of the theory of reasoned action (TRA) [20,2]. It is one of the most predictive persuasion theories. It has been applied to studies of the relations among beliefs, attitudes, behavioral intentions and behaviors in various fields such as advertising, public relations, campaigns, healthcare, etc.

TPB posits that individual behavior is driven by behavioral intentions where behavioral intentions are a function of an individual's attitude toward the behavior, the subjective norms surrounding the performance of the behavior, and the

individual's perception of the ease with which the behavior can be performed (behavioral control) (see Figure 2).

Attitude toward the behavior is defined as the individual's positive or negative feeling about performing a behavior. It is determined through an assessment of one's beliefs regarding the consequences arising from a behavior and an evaluation of the desirability of these consequences. Subjective norm is defined as an individual's perception of whether people think their significant others wanted them to perform the behavior. The contribution of the opinion of any given referent is weighted by the motivation that an individual has to comply with the wishes of that referent. Behavioral control is defined as one's perception of the difficulty of performing a behavior. TPB views the control that people have over their behavior as lying on a continuum from behaviors that are easily performed to those requiring considerable effort, resources, etc.

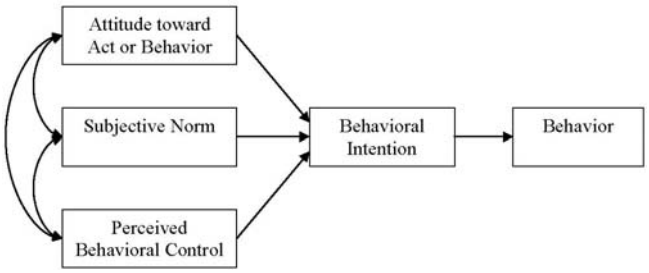

Figure 2. The model of Theory of Planned Behavior [2].

Technology acceptance model is another influential extension of Ajzen and Fishbein's theory of reasoned action (TRA) [20]. Some online trust models were built based on it especially when they examined user experience with Web technologies.

It was developed by Fred Davis and Richard Bagozzi to model how users come to accept and use a technology [3,16]. The model suggests that when users are presented with a new software package, a number of factors (replacing many of TRA's attitude measures) influence their decision about how and when they will use it.

TAM posits that *perceived usefulness* and *perceived ease of use* determine an individual's intention to use a system, with *intention to use* serving as a mediator of actual system use. Perceived usefulness is also seen as being directly impacted by perceived ease of use. Formally, perceived usefulness (PU) was defined as "the degree to which a person believes that using a particular system would enhance his or her job performance", and perceived ease-of-use (PEOU) is "the degree to which a person believes that using a particular system would be free from effort" [16].

Trust Model

Inspired by the two theories, our trust model consists of four main components specific to recommender systems: system design features, competence constructs, trustworthiness of the system and trusting intentions (see Figure 3 of the model)

System Design Features. The system features basically deal with all design aspects of a recommender system that will probably contribute to the promotion of its overall competence perceptions. Concretely, they include the interface display techniques such as the explanation-based interface to give system *transparency*, the *recommendation quality* to reflect users' perception of the recommender algorithm's accuracy and the user-system interaction models like the allowed degree of *user control*.

Competence Constructs. It is widely accepted that competence, benevolence and integrity explain a major portion of a trustee's trustworthiness [23]. Among them, we believe that the competence perception would be most reflective of system design qualities of the recommender. Based on TAM and related works [16,37], we include typical constructs of *perceived ease of use*, *perceived usefulness*, and an additional capacity assessment *enjoyment*. Moreover the two subjective measurements of decision accuracy and decision effort are also involved to represent the system's decision support quality.

Trustworthiness. The "trustworthiness" (or called credibility) [42] is the main positive influence on trusting intentions [23,44]. In our model, it is generally assessed by two major constructs: *trust in recommendations* and users' overall *satisfactory* degree with the recommender interface.

Trusting Intentions. Trusting intention is the extent to which the user is willing to depend on the technical party in a given situation [45]. We include in our model the *intention to purchase* (i.e. purchase a product from the website where the recommender is found) and the *intention to return* (i.e. return to the recommender system for more products information), as most of e-commerce based trust models emphasize. In addition, we added the *intention to save effort* to address whether the recommender system could allow its users to benefit from the built trust. That is, whether upon establishing a certain trust level with the recommender at the first visit, users will more readily accept the recommended items, rather than exerting extra effort to process all information themselves, once returning to use it.

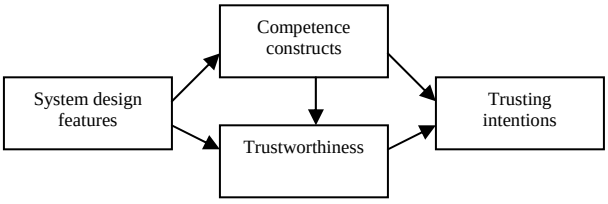

Figure 3. The user-trust model for recommender systems.

Therefore, as shown in Figure 3, all subjective variables are grouped into four categories. During the user evaluation of a system, in addition to analyzing each single variable, it will be also interesting to identify the relationships between different variables through correlation analysis. For instance, how the perceptions of system-design features are associated

with specific constructs of competence assessments, and how the competence constructs influence trust promotions, which would furthermore affect trusting intentions.

Table 1 lists all of the questions that can be adopted to measure these subjective variables. Most of them came from existing literatures where they have been repeatedly shown to exhibit strong content validity and reliability. Each question is required to respond on a 5-point Likert Scale ranging from "strongly disagree" to "strongly agree".

Table 1. Questions to measure subjective constructs in our trust model.

Measured variable	Question responded on a 5-point Likert scale from "strongly disagree" to "strongly agree"
Subjective perceptions of system-design features	
Transparency	I understand why the products were returned through the explanations in the interface.
Recommendation quality	This interface gave me some really good recommendations.
User control	I felt in control of specifying and changing my preferences in this interface.
Overall competence perceptions	
Perceived ease of use	I find this interface easy to use.
Perceived usefulness	This interface is competent to help me effectively find products I really like.
	I find this interface is useful to improve my "shopping" performance.
Enjoyment	I found my visit to this interface enjoyable.
Decision confidence	I am confident that the product I just "purchased" is really the best choice for me.
Perceived effort	I easily found the information I was looking for.
	Looking for a product using this interface required too much effort (*reverse scale*).
Trustworthiness	
Trust in recommendations	I feel that this interface is trustworthy.
	I trust the recommended products since they were consistent with my preferences.
Satisfaction	My overall satisfaction with the interface is high.
Trusting intentions	
Intention to purchase	I would purchase the product I just chose if given the opportunity.
Intention to return	If I had to search for a product online in the future and an interface like this was available, I would be very likely to use it.
	I don't like this interface, so I would not use it again (*reverse scale*).
Intention to save effort in next visit	If I had a chance to use this interface again, I would likely make my choice more quickly.

CONCLUSION

Thus, as a summary, our evaluation framework is mainly composed of two important components: the accuracy-effort measures and the user-trust model. The objective accuracy and effort are respectively measured by observing users' switching rate, recording their interaction effort and time consumed to accomplish their search tasks. Regarding subjective measures such as perceived accuracy, perceived effort and trust-related constructs, a post-study questionnaire is designed to ask for users' subjective opinions and comments after they finished their decision tasks with the assigned recommender system.

We have previously conducted a series of user studies with the goal of consolidating the evaluation framework. For example, *example-critiquing systems*, that support users to provide explicit feedbacks to recommendations, have been mainly tested regarding users' decision accuracy and decision effort [52,53,11]. The user evaluations found that the system can significantly improve decision accuracy compared to non-critiquing systems, while demanding similar level of effort. Another user study on *organization-based explanation interface* identified the explanation's positive role in increasing users' competence perception and return intention as focused in the trust model [12]. The user study on *hybrid critiquing* interface (integrated with system-suggested critiques) took both of users' decision performance and subjective perceptions into consideration [13,14].

Given the importance of these evaluation criteria as implied by researches on adaptive decision theory and online-trust building, we believe that this evaluation framework will be useful and scalable to the evaluation of recommender systems in a broad domain including recent social recommender systems [27]. In fact, we have started to extend the subjective constructs with more aspects, such as diversity, novelty, attractiveness, and test their practicability in domains of public tastes (e.g. movie, music). In the future, we will continually validate and refine the framework in different system scenarios in order to generalize its applicable value.

REFERENCES

1. G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734-749, June 2005.
2. I. Ajzen. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179-211, December 1991.
3. R. P. Bagozzi, H. Baumgartner, and Y. Yi. State versus action orientation and the theory of reasoned action: An application to coupon usage. *Journal of Consumer Research*, 18(4):505-518, 1992.
4. I. Benbasat and B. R. Nault. An evaluation of empirical research in managerial support systems. *T.H.E. Journal (Technological Horizons in Education)*, 6(3):203-226, 1990.
5. J. R. Bettman, M. F. Luce, and J. W Payne. Constructive consumer choice processes. *Journal of Consumer Research: An Interdisciplinary Quarterly*, 25(3):187-217, December 1998.
6. J. R. Bettman, E. J. Johnson, and J. W. Payne. A componential analysis of cognitive effort in choice. *Organizational Behavior and Human Decision Processes*, 45(1):111-139, February 1990.
7. P. Bollmann. A comparison of evaluation measures for document-retrieval systems. *Journal of Informatics*, 1:97-116, 1977.

8. P. Bonhard, C. Harries, J. McCarthy, and M. A. Sasse. Accounting for taste: Using profile similarity to improve recommender systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'06)*, pages 1057-1066, New York, NY, USA, 2006. ACM.
9. R. D. Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331-370, 2002.
10. G. Carenini and D. Poole. Constructed preferences and value-focused thinking: Implications for AI research on preference elicitation. In *AAAI'02 Workshop on Preferences in AI and CP: Symbolic Approaches*, Edmonton, Canada, 2002.
11. L. Chen and P. Pu. Evaluating critiquing-based recommender agents. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI'06)*, pages 157-162, Boston, USA, 2006. AAAI.
12. L. Chen and P. Pu. Trust building with explanation interfaces. In *Proceedings of International Conference on Intelligent User Interface (IUI'06)*, pages 93-100, Sydney, Australia, January 2006.
13. L. Chen and P. Pu. Hybrid critiquing-based recommender systems. In *Proceedings of International Conference on Intelligent User Interfaces (IUI'07)*, pages 22-31, Hawaii, USA, January 2007.
14. L. Chen and P. Pu. A Cross-Cultural User Evaluation of Product Recommender Interfaces. In *Proceedings of ACM Conference on Recommender Systems (RecSys'08)*, pages 75-82, Lausanne, Switzerland, October 23-25, 2008.
15. K. Chopra and W. A. Wallace. Trust in electronic environments. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03)*, Washington, DC, USA, 2003. IEEE Computer Society.
16. F. D. Davis. Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS Quarterly*, 13(3):319-340, September 1989.
17. P. M. Doney and J. P. Cannon. An examination of the nature of trust in buyer-seller relationships. *Journal of Marketing*, 61:35-51, 1997.
18. H. Einhorn and R. Hogarth. Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, 85:395-416, 1978.
19. N. Epley and T. Gilovich. Putting adjustment back into the anchoring-and-adjustment heuristic: Self-generated versus experimenter provided anchors. *Psychological Science*, 12(5):391-396, 2001.
20. M. Fishbein and I. Ajzen. *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Reading, MA: Addison-Wesley, 1975.
21. D. Frisch and R. T. Clemen. Beyond expected utility: Rethinking behavioral decision research. *Psychological Bulletin*, 116:46-54, 1994.
22. S. Ganesan. Determinants of long-term orientation in buyer-seller relationships *Journal of Marketing*, 58:1-19, 1994.
23. D. Gefen. E-commerce: the role of familiarity and trust. *International Journal of Management Science*, 28:725-737, 2000.
24. D. Gefen, E. Karahanna, and D. W. Straub. Inexperience and experience with online stores: the importance of tam and trust. *IEEE Transactions on Engineering Management*, 50(3):307-321, 2003.
25. D. Gefen, V. S. Rao, and N. Tractinsky. The conceptualization of trust, risk and their relationship in electronic commerce: The need for clarifications. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03)*, page 192, 2003.
26. S. Grabner-Krauter and E. A. Kaluscha. Empirical research in on-line trust: A review and critical assessment. *International Journal of Human-Computer Studies*, 58(6):783-812, 2003.
27. I. Guy, N. Zwerdling, D. Carmel, I. Ronen, E. Uziel, S. Yogev, and S. Ofek-Koifman. Personalized recommendation of social software items based on social relations. In *Proceedings of the Third ACM Conference on Recommender Systems (RecSys'09)*, pages 53-60, New York, NY, USA, 2009. ACM.
28. G. Haubl and V. Trifts. Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science*, 19(1):4-21, 2000.
29. K. S. Hassanein and M. M. Head. Building online trust through socially rich web interfaces. In *Proceedings of the Second Annual Conference on Privacy, Security, and Trust (PST'2004)*, pages 15-22, Fredericton, Canada, 2004.
30. S. J. Hoch and D. A. Schkade. A psychological approach to decision support systems. *Management Science*, 42(1):51-64, 1996.
31. R. Hogarth. *Judgment and Choice: The Psychology of Decision*. J. Wiley, New York, 1987.
32. M. Jamali and M. Ester. Using a trust network to improve top-N recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems (RecSys'09)*, pages 181-188, New York, NY, USA, 2009. ACM.
33. S. L. Jarvenpaa, N. Tractinsky, and M. Vitale. Consumer trust in an internet store. *Information Technology and Management*, 1(1-2):45-71, 2000.
34. D. Kahneman, P. Slovic, and A. Tversky. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, 1981.
35. R. L. Keeney. *Value-Focused Thinking*. Harvard University Press, Cambridge, MA, 1992.

36. R. L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-offs*. Cambridge University Press, Cambridge, 1993.
37. M. Koufaris and W. Hampton-Sosa. Customer trust online: Examining the role of the experience with the web-site. Working paper series, Zicklin School of Business, Baruch College, New York, NY, 2002.
38. B. Krulwich. Lifestyle finder: Intelligent user profiling using large-scale demographic data. *AI Magazine*, 18(2):37-45, 1997.
39. K. Lim, C. Sia, M. Lee, and I. Benbasat. Do I trust you online, and if so, will I buy? An empirical study of two trust-building strategies. *Journal of Management Information Systems*, 23(2):233-266, 2006.
40. M. F. Luce, J. W. Payne, and J. R. Bettman. Emotional trade-off difficulty and choice. *Journal of Marketing Research*, 36:143-159, May 1999.
41. P. Massa and B. Bhattacharjee. Using trust in recommender systems: an experimental analysis. In *Proceedings of 2nd International Conference on Trust Management*, 2004.
42. R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. *Academy of Management Review*, 20(3):709-734, 2005.
43. K. McCarthy, J. Reilly, L. McGinty, and B. Smyth. Experiments in dynamic critiquing. In *Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI'05)*, pages 175-182, New York, NY, USA, 2005. ACM.
44. D. H. McKnight and N. L. Chervany. What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. *International Journal of Electronic Commerce*, 6(2):35-59, 2002.
45. D. H. McKnight, L. Cummings, and N. L. Chervany. Initial trust formation in new organizational relationship. *Academy of Management Review*, 23(3):473-490, 1998.
46. B. Miller, I. Albert, S. K. Lam, J. Konstan, and J. Riedl. Movielens unplugged: Experiences with a recommender system on four mobile devices. In *Proceedings of the 17th Annual Human-Computer Interaction Conference*, September 2003.
47. J. O'Donovan and B. Smyth. Trust in recommender systems. In *Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI'05)*, pages 167-174, New York, NY, USA, 2005. ACM.
48. P. A. Pavlou and R. K. Chelllappa. The role of perceived privacy and perceived security in the development of trust in electronic commerce transactions. Working paper, 2001.
49. J. W. Payne. Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Processes*, 16:366-387, 1976.
50. J. W. Payne, J. R. Bettman, and E. J. Johnson. *The Adaptive Decision Maker*. Cambridge University Press, Cambridge, UK, 1993.
51. J. W. Payne, J. R. Bettman, and D. A. Schkade. Measuring constructed preferences: Towards a building code. *Journal of Risk and Uncertainty*, 19(1-3):243-270, December 1999.
52. P. Pu and P. Kumar. Evaluating example-based search tools. In *Proceedings of the 5th ACM Conference on Electronic Commerce (EC'04)*, pages 208-217, New York, NY, USA, 2004. ACM.
53. P. Pu and L. Chen. Integrating tradeoff support in product search tools for e-commerce sites. In *Proceedings of ACM Conference on Electronic Commerce (EC'05)*, pages 269-278, Vancouver, Canada, June 2005.
54. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work (CSCW'94)*, pages 175-186, New York, NY, USA, 1994. ACM.
55. K. De Ruyter, M. Wetzels, and M. Kleijnen. Customer adoption of eservices: an experimental study. *International Journal of Service Industry Management*, 12(2):184-207, 2001.
56. R. N. Shepard. On subjectively optimum selection among multiattribute alternatives. *Human Judgment and Optimality*, 1964.
57. S. M Shugan. The cost of thinking. *Journal of Consumer Research: An Interdisciplinary Quarterly*, 7(2):99-111, Se 1980.
58. H. A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99-118, 1955.
59. S. Spiekermann and C. Paraschiv. Motivating human-agent interaction: Transferring insights from behavioral marketing to interface design. *Electronic Commerce Research*, 2(3):255-285, 2002.
60. K. Swearingen and S. Rashmi. Interaction design for recommender systems. In *Proceedings of the Conference on Designing Interactive Systems (DIS'02)*, London, England, 2002. ACM Press.
61. M. Torrens, B. Faltings, and P. Pu. Smartclients: Constraint satisfaction as a paradigm for scalable intelligent information systems. *Constraints*, 7(1):49-69, 2002.
62. A. Tversky and I. Simonson. *Context-dependent preferences*. Management Science, 39(10):1179-1189, 1993.
63. Q. Yuan, S. Zhao, L. Chen, S. Ding, X. Zhang and W. Zheng. Augmenting collaborative recommender by fusing explicit social relationships. In *ACM Conference on Recommender Systems (RecSys'09), workshop on*

Recommender Systems and the Social Web, New York City, NY, USA, October 22-25, 2009.

64. J. Zhang and P. Pu. Effort and accuracy analysis of choice strategies for electronic product catalogs. In *Proceedings of the 2005 ACM Symposium on Applied Computing (SAC'05)*, pages 808-814, New York, NY, USA, 2005. ACM.
65. Y. Zhen, W. Li, and D. Yeung. TagiCoFi: tag informed collaborative filtering. In *Proceedings of the Third ACM Conference on Recommender Systems (RecSys'09)*, pages 69-76, New York, NY, USA, 2009. ACM.
66. J. Zimmerman and K. Kurapati. Exposing profiles to build trust in a recommender. In *Extended Abstracts on Human Factors in Computing Systems (CHI'02)*, pages 608-609, New York, NY, USA, 2002. ACM.
67. P. Pu, B. Faltings, L. Chen and J. Zhang. Usability Guidelines for Preference-based Product Recommenders. *Recommender Systems Handbook*, 2010 (to appear).
68. Flavián, C., Guinalíu, M., and Gurrea, R. The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Inf. Manage.* 43 (1), Jan. 2006, 1-14.