User Model User-Adap Inter (2009) 19:167–206
DOI 10.1007/s11257-008-9057-x

ORIGINAL PAPER

Interaction design guidelines on critiquing-based recommender systems

Li Chen · Pearl Pu

Received: 19 September 2007 / Accepted in revised form: 25 August 2008 /
Published online: 3 October 2008

Abstract A critiquing-based recommender system acts like an artificial salesperson. It engages users in a conversational dialog where users can provide feedback in the form of critiques to the sample items that were shown to them. The feedback, in turn, enables the system to refine its understanding of the user's preferences and prediction of what the user truly wants. The system is then able to recommend products that may better stimulate the user's interest in the next interaction cycle. In this paper, we report our extensive investigation of comparing various approaches in devising critiquing opportunities designed in these recommender systems. More specifically, we have investigated two major design elements which are necessary for a critiquing-based recommender system: *critiquing coverage*—one vs. multiple items that are returned during each recommendation cycle to be critiqued; and *critiquing aid*—system-suggested critiques (i.e., a set of critique suggestions for users to select) vs. user-initiated critiquing facility (i.e., facilitating users to create critiques on their own). Through a series of three user trials, we have measured how real-users reacted to systems with varied setups of the two elements. In particular, it was found that giving users the choice of critiquing one of multiple items (as opposed to just one) has significantly positive impacts on increasing users' decision accuracy (particularly in the first recommendation cycle) and saving their objective effort (in the later critiquing cycles). As for critiquing aids, the hybrid design with both system-suggested critiques and user-initiated critiquing support exhibits the best performance in inspiring users' decision confidence and increasing their intention to return, in comparison with the uncombined exclusive approaches. Therefore, the results from our studies shed light

L. Chen (✉) · P. Pu
Human Computer Interaction Group, School of Computer and Communication Sciences,
Swiss Federal Institute of Technology in Lausanne (EPFL), 1015 Lausanne, Switzerland
e-mail: li.chen@epfl.ch

P. Pu
e-mail: pearl.pu@epfl.ch

on the design guidelines for determining the sweetspot balancing user initiative and system support in the development of an effective and user-centric critiquing-based recommender system.

Keywords Critiquing-based recommender systems · Decision support · Preference revision · User control · Example critiquing · Dynamic critiquing · Hybrid critiquing · User evaluation · Usability · Human–computer interaction

1 Introduction

According to adaptive decision theory (Payne et al. 1993), the human decision process is inherently highly constructive and adaptive to the current decision task and decision environment. In particular, when users are confronted with an unfamiliar product domain or a complex decision situation with overwhelming information, such as the current e-commerce environment, they are usually unable to accurately state their preferences at the outset (Viappiani et al. 2007) but likely construct them in a highly context-dependent fashion during their decision process (Tversky and Simonson 1993; Payne et al. 1999; Carenini and Poole 2002).

In order to assist people in making accurate as well as confident decisions, especially in the complex decision setting, critiquing-based recommender systems have emerged in the form of both natural language models (Shimazu 2001; Thompson et al. 2004) and graphical user interfaces (Burke et al. 1996, 1997; Reilly et al. 2004; Pu and Kumar 2004). This type of system has been broadly recognized as an effective feedback mechanism that may guide users to efficiently target at their ideal products, which is particularly meaningful when users are searching for high-involvement products (e.g., computers, houses and cars) with the primary goal of avoiding any financial damage. Other terms for these systems are conversational recommender systems (Smyth and McGinty 2003), conversational case-based reasoning systems (Shimazu 2001), and knowledge-based recommender systems (Burke et al. 1997; Burke 2000).

More specifically, the critiquing-based recommender system mainly acts like an artificial salesperson that engages users in a conversational dialog where users can provide feedback in form of critiques (e.g., "I like this laptop, but prefer something cheaper" or "with faster processor speed") to one of currently recommended items. The feedback, in turn, enables the system to more accurately predict what the user truly wants and then return some products that may better interest the user in the next conversational cycle. The main component of this interaction model is therefore that of recommendation-and-critiquing, which is also called tweaking (Burke et al. 1997), critiquing feedback (Smyth and McGinty 2003), candidate/critiquing (Linden et al. 1997), and navigation by proposing (Shimazu 2001).

To our knowledge, the 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 has been aiming to pro-actively 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 this paper). This mechanism has been adopted in FindMe systems (Burke et al. 1997)

and more recent DynamicCritiquing agents (Reilly et al. 2004; McCarthy et al. 2005c). The main advantage, as detailed in related literatures (Reilly et al. 2004; McCarthy et al. 2004b; McSherry 2004), is that system-suggested critiques can not only expose the knowledge of remaining recommendation opportunities, but also potentially accelerate the user's critiquing process if they can 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 (so called *user-initiated critiquing support* in this paper). As a typical application, the ExampleCritiquing agent has been developed for this goal, and its focus is showing examples and facilitating users to compose their self-initiated critiques (Pu and Kumar 2004). In essence, the ExampleCritiquing agent is capable of allowing users to choose which feature(s) to be critiqued and how to critique it (or them) under their own control. Previous work proved that it enabled users to obtain significantly higher decision accuracy and preference certainty, compared to non critiquing-based systems such as a ranked list (Pu and Kumar 2004; Pu and Chen 2005).

In addition to characterizing the critiquing-based recommender system in terms of its nature of critiquing support (i.e., *system-suggested critiques* or *user-initiated critiquing support*), another important factor is the number of items that the system returns during each recommendation cycle for users to critique. For example, FindMe and DynamicCritiquing systems return one item, whereas ExampleCritiquing agents show multiple k items (e.g., $k = 7$) at a cycle. Multi-item display provides users a chance to choose the product to be critiqued after making a comparison between several options.

Thus, there are in nature two crucial design components contained in a critiquing-based recommender system. One is its *critiquing aid*: suggesting critiques for users to select or aiding them to construct their own critiques. Another is the number of recommended items (called *critiquing coverage* in this paper): suggesting a single vs. multiple products for users to critique.

The options are inherently related to different levels of user control in either the process of identifying the critiqued reference or the process of specifying concrete critiquing criteria. As a matter of fact, perceived behavioral control has been regarded as an important determinant of user beliefs and actual behavior (Ajzen 1991). In the context of e-commerce, it has been found to have a positive effect on customers' attitudes including their perceived ease of use, perceived usefulness and trust (Novak et al. 2000; Koufaris and Hampton-Sosa 2002). User control has been also determined as one of the fundamental principles for general user interface design (Shneiderman 1997) and Web usability (Nielsen 1994).

However, there are few works having studied the effect of locus of user initiative in critiquing-based recommender systems. There is indeed a complex tradeoff that underlies the successful design: giving users too much control may cause them to perform an unnecessary complex critiquing, whereas giving little or no control may force users to accept system-suggested items even though they do not match users' truly-intended choices. The goal of this paper is therefore to investigate the different degrees of user control vs. system support in both *critiquing aid* and *critiquing coverage*, so as to identify the optimal combination of components that could positively influence users' actual decision performance and subjective attitudes.

To achieve our goal, we have conducted a series of three trials. In our first user trial, we compared two well-known critiquing-based recommender agents which respectively represent a typical setup combination of *critiquing coverage* and *critiquing aid*. Concretely, one is the DynamicCritiquing system that shows one recommended product during each interaction cycle, accompanied by a user-initiated unit critiquing area and a list of system-suggested compound critiques. Another is the ExampleCritiquing system that returns multiple products in a display and stimulates users in building and composing critiques to one of the shown products in their self-motivated way. The experimental results show that the ExampleCritiquing agent achieved significantly higher decision accuracy (in terms of both objective and subjective measures) and users' behavioral intentions (i.e., intention to purchase and return), while requiring lower level of interaction and cognitive effort.

In the second trial, we modified ExampleCritiquing and DynamicCritiquing to make their critiquing coverage (i.e., the number of recommended items during each cycle) constant and keep the difference only on their critiquing aids. The results surprisingly showed that there is no significant difference between the two modified versions in terms of both objective and subjective measures. Further analysis of participants' comments revealed the pros and cons of system-suggested critiques and user-initiated critiquing support. Additionally, combining the results with the first trial's, we found that giving users the choice of critiquing one of multiple items (as opposed to just one) has significantly positive impacts on increasing their decision accuracy and confidence particularly in the first recommendation cycle and saving objective effort in the later critiquing rounds.

The third user trial was conducted to measure users' performance in a hybrid critiquing system where system-suggested critiques and user-initiated critiquing aid was combined on one screen. Analyzing users' critiquing application frequency in such system shows that the application of user-initiated critiquing support in creating users' own critiques is relatively higher than picking suggested critique options. Moreover, the respective practical effects of *user-initiated* and *system-suggested* critiquing facilities were identified. That is, they are both significantly contributive to improve users' decision confidence and return intention, and system-suggested critiques are even effective in saving effort perception.

Therefore, all of our trial results infer that giving users multiple recommended products as critiqued options and providing them both system-suggested and user-initiated critiquing aids for specifying concrete critiquing criteria can obtain substantial benefits.

Another contribution of our work is that we have established a user-evaluation framework. It contains both objective variables such as decision accuracy, task completion time and interaction effort, and subjective measures like perceived cognitive effort, decision confidence and *trusting intentions*. All of these factors are fundamentally important, given that a recommender system's ultimate goal should be to allow its users to achieve high decision accuracy and build high trust in it, and require them to expend a minimal amount of effort to obtain these benefits (Häubl and Trifts 2000; Chen and Pu 2005; Pu and Chen 2005).

The rest of this paper is organized as follows. We first introduce existing critiquing-based recommender systems, with DynamicCritiquing and ExampleCritiquing as two

representatives. According to their respective characteristics, we summarize two main elements that can be varied to reflect different degrees of user control. We then introduce a user evaluation framework with major dependent variables measured in our experiments. Detailed descriptions of three user-trials then follow, including their materials, recruited participants, experimental procedures, results analyses and discussions. Finally, we conclude our work and indicate its practical implications and future directions.

2 Critiquing-based recommender systems

Our investigation of existing critiquing-based recommender systems revealed that they basically follow a similar interaction model (see Fig. 1). The user first specifies her initial preferences on product attributes. The system then returns one or multiple recommended items. Either the user selects an item as her final choice and terminates the interaction, or she makes critiques by picking system-suggested critiques or defining critiques herself. If critiques were made, the system updates the recommendation(s) and the list of suggested critiques (if provided) in the next interaction cycle. This process continues until the user decides that she has found her most preferred product.

Most of existing systems fall into two specific branches: one is called *single-item system-suggested critiquing* since it recommends one item at a time and guides users to provide feedback by selecting a system-suggested critique; another is called *k-item user-initiated critiquing*, because it provides multiple items during each recommendation cycle and a critiquing aid that assists users in choosing one product to be critiqued and creating their self-specified critiquing criteria to the product. In the following, we will introduce both approaches in detail with two typical applications as examples.

2.1 Single-item system-suggested critiquing

The FindMe system was the first known single-item system-suggested critiquing system (Burke et al. 1996, 1997). It uses knowledge about the product domain to help users

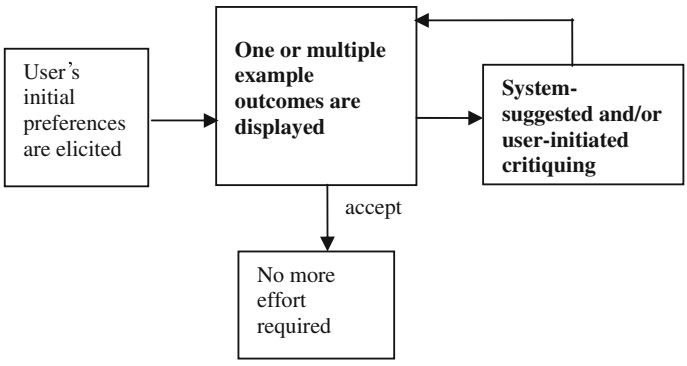

Fig. 1 The typical interaction model of a critiquing-based recommender system

navigate through the multi-dimensional space. An important interface component in FindMe is called *tweaking*, which allows users to critique the current recommendation by selecting one of the proposed simple tweaks (e.g., "cheaper", "bigger" and "nicer"). When a user finds the current recommendation short of her expectations and responds to a tweak, the remaining candidates will be filtered to leave only those candidates satisfying the tweak.

The critique suggestions in FindMe are called unit critiques since each of them constrains on a single feature at a time. More recently, a so-called *dynamic critiquing* method (Reilly et al. 2004; McCarthy et al. 2004a) has been developed with the objective of automatically generating a set of compound critiques, each of which can operate over multiple features simultaneously (e.g., "Different Manufacture, Lower Resolution and Cheaper"). A live-user trial showed that the integration of the *dynamic critiquing* method can effectively reduce users' intention cycles from an average of 29 in purely applying unit critiques to 6 (McCarthy et al. 2005c). The compound critiques can also perform as explanations, revealing the remaining recommendation opportunities except for the current product (Reilly et al. 2005). Therefore, we use the DynamicCritiquing system as the representative to illustrate the main components that a *single-item system-suggested critiquing system* may comprise.

2.1.1 DynamicCritiquing

Figure 2 shows a sample DynamicCritiquing interface where both unit and compound critiques are available to users as feedback options (Reilly et al. 2004; McCarthy et al. 2005c). It can be seen that the DynamicCritiquing interface mainly contains three components: a single item as the current recommendation, a unit critiquing area and a list of compound critiques. In the first recommendation cycle, an item that best matches the user's initially stated preferences is returned, and then after each critiquing action, a new item that satisfies the user's critique as well as being most similar to the previous recommended product will be shown as the current recommendation.

In the unit critiquing area, the system determines a set of main features, one of which users can choose to critique at a time. For each numerical feature (e.g., price), two critiquing directions are provided: increasing the value (e.g., more expensive) or decreasing it (e.g., cheaper), and for discrete features (e.g., manufacture) all of the relevant options are displayed under a drop-down menu. Therefore, this area performs more like a user-initiated unit critiquing support, rather than a limited small set of unit critique suggestions as in FindMe systems.

The list of three compound critiques are automatically computed by discovering the recurring subsets of unit differences between the current recommended item and the remaining products using a data mining algorithm called Apriori (Agrawal et al. 1993). More concretely, each remaining product, except the current recommendation, is first converted into a critique pattern indicating its differences from the current recommended product in terms of all main features (e.g., {(manufacture, =), (price, $<$), (weight, $>$),...}). Since there will be a number of critique patterns representing all of the remaining products, the Apriori algorithm is employed to discover the frequent association rules among features within these patterns. A set of compound

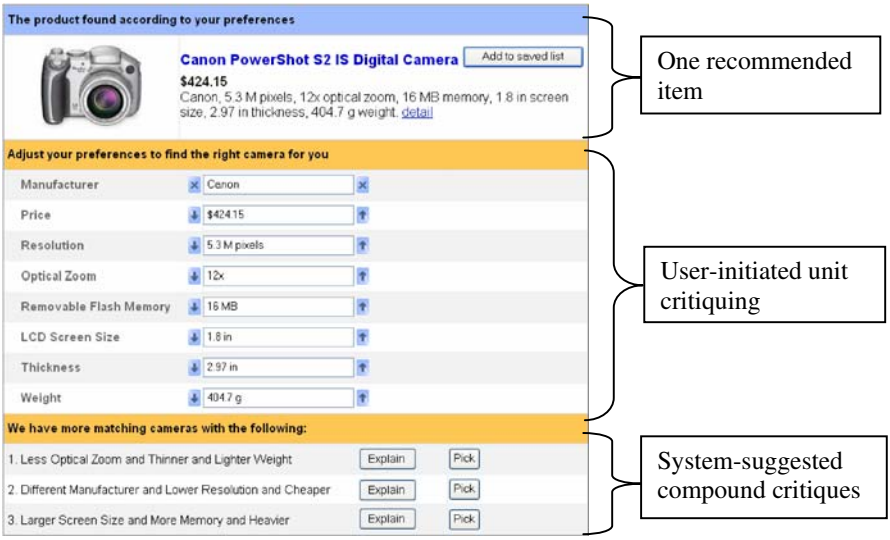

Fig. 2 The DynamicCritiquing interface

critique options (as the frequent association rules) will be then produced. For example, supposing the occurrence of heavier laptops is highly frequently associated with the occurrence of cheaper prices in the remaining items, a compound critique with the form of {[weight >], [price <]} (i.e., heavier and cheaper) will be generated. Thus, the DynamicCritiquing agent uses the Apriori algorithm to discover the highest recurring compound critiques representative of a given data set. It then favors those candidates with the lowest support values ("support value" refers to the percentage of products that satisfy the critique). Such selection criterion was motivated by the fact that presenting critiques with lower support values provides a good balance between their likely applicability to the user and their ability to narrow the search (McCarthy et al. 2004a, 2005b,c).

In addition to functioning as critique suggestions, the dynamically generated compound critiques have been also regarded as explanations exposing the recommendation opportunities that exist in the available products (McCarthy et al. 2004b; Reilly et al. 2005). They may help users be familiar with the product domain and understand the relationship among different features within the alternatives. Users can be then stimulated to express more preferences or be prevented from making retrieval failures (Reilly et al. 2005).

2.2 K-item user-initiated critiquing

Instead of suggesting pre-computed critiques for users to select, the purely user-initiated critiquing approach focuses on showing examples and stimulating users to define critiques themselves. It does not limit the size of critiques a user can manipulate during each cycle, so that the user can post either unit or compound critiques

over any combination of features with freedom. In fact, the purpose of this type of critiquing support is to assist users in freely executing tradeoff navigation, which is a process shown to improve users' decision accuracy and confidence (Pu and Kumar 2004; Pu and Chen 2005). The ExpertClerk (Shimazu 2001), ATA (Automated Travel Assistant) (Linden et al. 1997) and SmartClient (Pu and Faltings 2000) were all examples of such systems. Nguyen et al.2004 realized the idea mainly to support on-tour recommendations for mobile users.

Such system is mainly composed of two components: a recommender agent that computes a set of k items that best match the user's current preference model, and a critiquing component that allows the user to actively create critiquing criteria and then examine a new set of k tradeoff alternatives. ExpertClerk and ATA display three items at a time, whereas SmartClient returned seven items in its recent versions. Users can select any of the displayed items and navigate to products that offer tradeoff potentials. As for the critiquing aid, ExpertClerk provides a natural language dialog to request for users' feedback, ATA stated that it developed a graphical interface but without detailed description, and SmartClient has constantly improved the usability of its critiquing facility through user evaluations. We have chosen a latest version of SmartClient, called ExampleCritiquing, to explain the typical constructs of a *k-item user-initiated critiquing system.*

2.2.1 ExampleCritiquing

SmartClient was originally developed as an online preference-based search tool for finding flights (Pu and Faltings 2000; Torrens et al. 2002). Its elementary model is the example-and-critiquing interaction, which was subsequently applied to product catalogs of vacation packages, insurance policies, apartments, and more recent commercial products such as tablet PCs and digital cameras (Pu and Faltings 2004; Pu and Kumar 2004; Chen and Pu 2006).

In the latest ExampleCritiquing system, the recommendation part can be further divided into two sub-components: the first set of recommendations computed according to the user's initial preferences, and the set of tradeoff alternatives recommended after each critiquing process. For example, for product catalogs of digital cameras and tablet PCs, k items (e.g., $k = 7$) are displayed in both cases. The number k was determined according to (Faltings et al. 2004) that discussed the optimal number of displayed solutions based on catalog sizes.

In the critiquing panel (see Fig. 3), three radio buttons are next to each main feature, facilitating users to choose to "keep" its value, "improve" it, or accept a compromised value suggested by the system (i.e., via "Take any suggestion"). In particular, users can freely compose compound critiques by combining criteria on any set of multiple features. The interface also supports users to perform simple similarity-based critiquing (e.g., "show similar products with this one") by just keeping all current values, or define concrete value improvements on features (for example, under the "Improve" dropdown menu of price, there are options "$100 cheaper", "$200 cheaper", etc.).

This kind of critiquing support has been also named as tradeoff assistance in some related literatures (Pu and Kumar 2004; Chen and Pu 2006), since it is in nature to

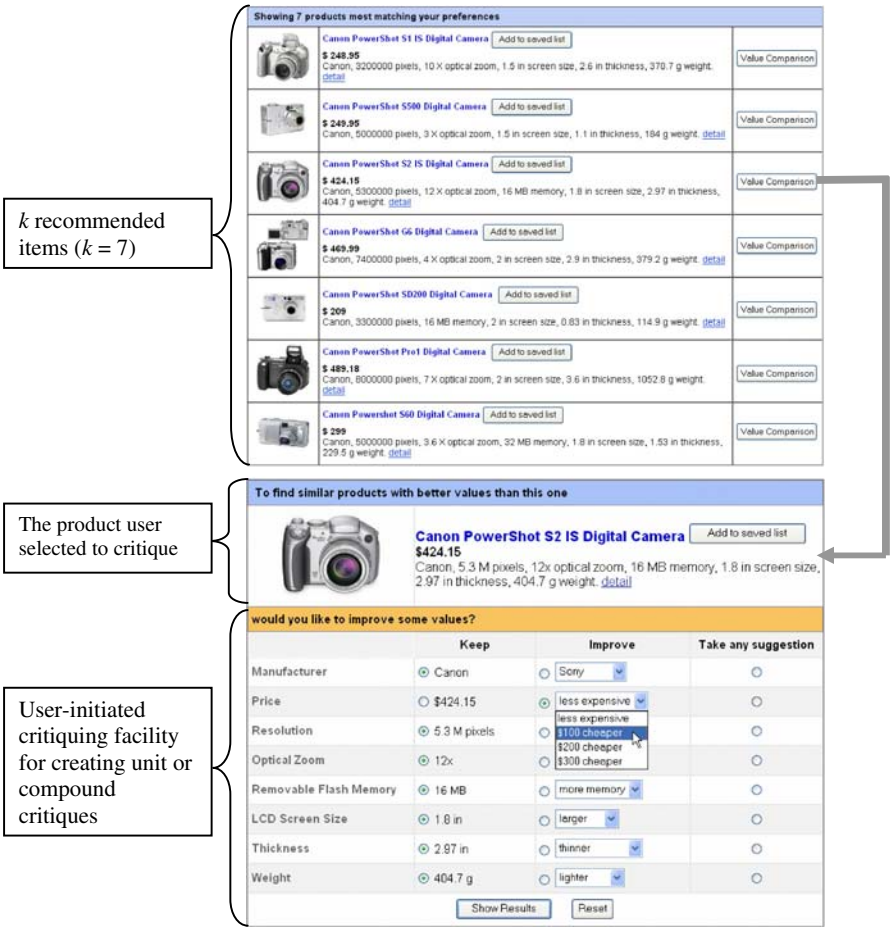

Fig. 3 The ExampleCritiquing interfaces

facilitate a user to specify tradeoff criteria: improving on one or several attributes that are important to her, while accepting compromised values on less important ones. Tradeoff process involving only one feature (unit critique) or multiple features (compound critique) are respectively termed as simple and complex tradeoffs by Pu and Kumar (2004).

The search engine of computing recommended alternatives is adjusted for different decision environments. For configurable products, it employs sophisticated constraint satisfaction algorithms and models user preferences as soft constraints (Torrens et al. 2002). For multi-attribute products, it is in theory grounded on the Weighted Additive sum rule (WADD), a compensatory decision strategy for explicitly resolving conflicting values (Payne et al. 1993). As required by WADD, the user's preferences are structured as a set of (*attribute's acceptable value, relative importance*) pairs.

After a user specifies her initial preferences, all alternatives will be ranked by their weighted utilities, and the top k items best matching the user's stated requirements will

be returned. Among the initial set of recommendations, the user either accepts a result, or takes a near solution to activate the critiquing panel (by clicking on the button "Value Comparison" along with the product, see Fig. 3). Once the critiquing criteria have been built in the critiquing panel, the system will refine the user's preference model and adjust the relative importance of all critiqued attributes (i.e., the weight of improved attribute(s) will be increased and that of compromised attribute(s) will be decreased). The search engine will then apply a combination of elimination-by-aspect (EBA) and WADD strategy (Payne et al. 1993). The combined strategy begins with EBA to first eliminate products that do not reach the minimal acceptable value (i.e., cutoff) of the improved attribute(s), and WADD is then applied to examine the remaining alternatives in more detail to select ones that best satisfy all of the user's tradeoff criteria. This example-and-critiquing process completes one cycle of interaction, and it continues as long as the user wants to refine the results.

3 Control variables

In a summary, the components contained by both DynamicCritiquing and E xample-Critiquing can be categorized into two independent variables: the number of recommendations that users could examine at a time based on which to perform critiquing, and the critiquing aid by which users could specify specific feedback criteria. As introduced before, two typical combinations of the two variables are *single-item system-suggested critiquing* and *k-item user-initiated critiquing*, but there should be more combination possibilities. In this section, we mainly discuss each variable's possible values.

3.1 Critiquing coverage (the number of recommendations)

Here we refer the *critiquing coverage* to the number of example products that are recommended to users for them to choose the final choice or critiqued object. In the ExampleCritiquing system, multiple examples are displayed during each recommendation cycle, because its objective is to stimulate users to make self-initiated critiques. On the contrary, the FindMe and DynamicCritiquing agent only returns one product based on which system-suggested critiques are generated. This simple display strategy has the advantage of not overwhelming users with too much information, but it deprives users of the right of choosing their own interested critiquing product, and potentially brings them the risk of engaging in a longer interaction session.

The critiquing coverage can be further separated into two sub-variables: the number of the first round's recommendations right after users' initial preference specification (called *NIR*), and the number of items (i.e., tradeoff alternatives) in the later cycle after each critiquing action (called *NCR*). The two numbers can be equal or different. For example, in DynamicCritiquing and ExampleCritiquing, they are both equal to 1 or 7. It is also possible to set them differently, for example, *NIR* as 1 and *NCR* as 7 if users are only interested in one best matching product according to their initial preferences, but would like to see multiple alternatives comparable with their critiqued reference once critiquing a product.

3.2 Critiquing aid

After recommended items are computed and displayed to the user, the critical concern now should be how to aid users in providing critiques to the item.

As introduced before, there are principally two types of critiquing aids: the system-suggested critiquing approach that generates and proposes a limited set of critiques for users to select, and the user-initiated critiquing approach that does not offer pre-computed critiques, but allows users to create and compose critiques on their own. The user-initiated method is more flexible to support various critique forms. For example, in the ExampleCritiquing interface, users can choose to make similarity-based critiquing (e.g., "find some cameras similar to this one"), quality-based (e.g., "find a similar camera, but cheaper") or even quantity-based (e.g., "find something similar to this camera, but at least $100 cheaper"). However, the system-suggested critiquing approach is limited in this respect given that it is the system to determine the form, not the user. In fact, FindMe and DynamicCritiquing only suggest quality-based critiques (e.g., "cheaper," "bigger," or "Different Manufacture, Lower Resolution and Cheaper") which were viewed as a compromise between the detail provided by value elicitation and the ease of feedback associated with preference-based methods (Smyth and McGinty 2003; McCarthy et al. 2005c).

In reference to the DynamicCritiquing interface, the critiquing aid can contain two sub-components: unit critiquing (on a single feature) and compound critiquing (on multiple features simultaneously) which are respectively termed UC and CC in the following content. Each sub-component can be in either system-suggested or user-initiated style. For example, the UC in FindMe (Burke et al. 1997) is system-suggested (e.g., "cheaper", "bigger"), whereas in DynamicCritiquing, it is more user-initiated since users can choose which feature to critique and how to critique it. The CC support in DynamicCritiquing, however, is purely system-suggested because a limited set of compound critiques is proposed for users to select (usually three suggestions as shown in Fig. 2).

In the ExampleCritiquing interface, both UC and CC are supported in the user-initiated way. Specifically, the user can improve or compromise one feature at a time and leave the others unchanged (i.e., unit critique), or combine any set of unit critiques into a compound critique.

Therefore, considering the degree of user control, the user-initiated method should allow for a higher level given that the control is largely in the hands of users, relative to the system-suggested critiquing approach where users can only "select", not "create". However, it is hard to assert which method would certainly perform better in improving on real-users' decision performance and subjective attitudes.

Table 1 lists all of the discussed variables, with DynamicCritiquing and ExampleCritiquing as examples to see their typical values.

4 User evaluation framework

We have conducted a series of three user trials, in order to understand the effect of these variables on users' actual decision behavior and subjective perceptions. The first trial

Table 1 Summary of control variables in a critiquing-based recommender system and the main differences between DynamicCritiquing and ExampleCritiquing in respect of these aspects

	Critiquing coverage		Critiquing aid	
	Number of initial recommendations (NIR)	Number of recommended items after each critiquing (NCR)	Unit critiquing (UC)	Compound critiquing (CC)
DynamicCritiquing (McCarthy et al. 2005c)	Single item	Single item	User-initiated	System-suggested
ExampleCritiquing (Chen and Pu 2006)	k items ($k = 7$)	k items ($k = 7$)	User-initiated	User-initiated

was a comparative user study of the two typical applications: DynamicCritiquing and ExampleCritiquing, with the purpose of identifying which one would perform more effectively. In the second trial, we made some changes on the two systems to make them different only on one dimension, the critiquing aid, in order to observe the single element's influence. The third trial measured users' performance in a hybrid critiquing system where the two types of critiquing aids: system-suggested and user-initiated, were combined on the same screen. Combining the results from these three trials, we expected to reveal the effects of different independent variables on users' decision performance and quality.

Therefore, before carrying out these experiments, it was necessary to first define concrete dependent variables that we were to measure. We have established an evaluation framework aimed to contain all of key standards. In fact, identifying the appropriate criteria for evaluating the true benefits of a recommender system is a challenging issue. Related work has primarily focused on users' objective interaction effort, such as their interaction sessions (McCarthy et al. 2005a,b,c) and task completion time, while placing less emphasis on what actual decision accuracy users can eventually achieve and how much cognitive effort users perceive to exert. In fact, the accuracy-effort model has long been studied in the domain of classical decision theories (Payne et al. 1993; Spiekermann and Parachiv 2002), and it has been broadly accepted that they are both important to determine the fundamental user benefits of a decision support, since the system's ideal goal should be to enable its users to obtain high level of decision accuracy with low amount of effort (Häubl and Trifts 2000).

In addition, a recommender system's ability in increasing user trust and convincing them of its recommendations, such as which camera to purchase, is also a crucial factor, particularly meaningful when the system is applied in the e-commerce environment. Two main trust-inspired behavioral intentions (called *trusting intentions*) include *intention to purchase* indicating whether the system could stimulate its users to purchase a product, and *intention to return* referring whether the system could prompt users to return to it for future use so that a long-term relationship is established (Grabner-Kräuter and Kaluscha 2003).

Therefore, motivated by these requirements, we have classified them into three categories of dependent variables in our evaluation framework: decision accuracy, decision effort and *trusting intentions* (see Fig. 4).

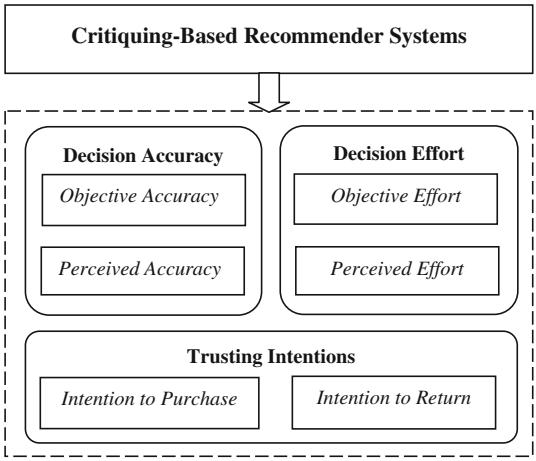

Fig. 4 User evaluation framework for critiquing-based recommender systems

4.1 Decision accuracy

The foremost criterion of evaluating a recommender system should be the decision accuracy that it enables users to eventually achieve. If a user can target her ideal choice with the system, it means that the system assisted her in reaching 100% decision accuracy. In our experiments, we not only measured the objective accuracy that a participant may obtain, but also her subjectively perceived accuracy (i.e., confidence in choice).

Objective accuracy

The objective accuracy was quantitatively measured by the fraction of participants who switched to a different, better option than the one chosen with the system, when they were asked to view all alternatives in the database. This procedure is known as the switching task, and has been practically applied by researchers in marketing science to measure consumers' decision quality (Häubl and Trifts 2000). A lower switching fraction means that the system supports higher decision accuracy since most of users stood by their choice with it. On the contrary, a higher switching fraction implies that the recommender is not very capable of guiding users to locate what they truly want. For expensive products, inaccurate tools may cause both financial damage and emotional burden to a decision maker.

Perceived accuracy

Besides objective accuracy, we also measured the degree of accuracy users subjectively perceived while using the system, which is also called decision confidence (Pu and Kumar 2004). The confidence judgment may potentially impact on users' perception of the system's competence and even their intention to purchase the chosen product. This variable was quantitatively assessed by asking subjects to respond to a statement (e.g., "I am confident that the product I just 'purchased' is really the best

Table 2 Questions to measure subjective perceptions

Measured subjective variables	Questions each responded on a 5-point Likert scale from "strongly disagree" to "strongly agree"
Perceived decision accuracy	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*).
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*).

choice for me") on a 5-point Likert scale ranging from "strongly disagree" to "strongly agree" see Table 2.

4.2 Decision effort

According to the accuracy-effort framework (Payne et al. 1993), another important criterion is the amount of decision effort users expended in making their choice with the system. Similar to decision accuracy, we not only measured how much objective effort users actually consumed, but also their perceived cognitive effort which we hope would indicate the amount of subjective effort people exerted.

Objective effort

The objective effort further includes two dimensions: task time and interaction effort. The task time is the total time a subject spent from she started using the system till she made her final choice. The interaction effort mainly considers the amount of interaction cycles (e.g., critiquing cycles) that a user was involved. The two variables have been widely used as main measurements in related work to evaluate their recommender systems (McCarthy et al. 2005b,c).

Perceived effort

Perceived effort refers to the psychological cognitive cost of information-processing. It represents the ease with which the subject can perform the task of obtaining and processing the relevant information in order to arrive at her decision. Since it is a subjective variable, two unified scale items (e.g., "I easily found the information I was looking for") were used to quantify its value (see Table 2 of concrete questions).

4.3 Accuracy and effort

The objective and subjective assessments of both decision accuracy and decision effort can not only show their respective values, but also allow us to understand how the concepts are interrelated.

According to (Bettman et al. 1990), individuals typically settle for imperfect accuracy of their decisions in return for a reduction in effort used. Empirical evidence also shows 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 (Einhorn and Hogarth 1978; Benbasat and Nault 1990; Häubl and Trifts 2000). On the other hand, recent research claimed that online users will be willing to make more effort if they perceive more benefits from the decision aid (Spiekermann and Parachiv 2002).

These statements drove us to address a critical question: what is the tradeoff relationship between accuracy and effort in our evaluated critiquing-based recommender systems? In particular, we were interested in identifying how much accuracy users could obtain with the system and the corresponding effort they were willing to expend.

4.4 Trusting intentions

Lack of trust has been demonstrated as one of the most frequently cited reasons for consumers not purchasing from Internet vendors (Grabner-Kräuter et al. 2006). Therefore, we included both *intention to purchase* and *intention to return* (McKnight and Chervany 2002; Grabner-Kräuter and Kaluscha 2003), to measure whether the system could contribute to building user trust regarding inspiring them to purchase the chosen product or return to the system for repeated uses. All associated questions came from existing literatures (Grabner-Kräuter and Kaluscha 2003), where they had been repeatedly shown to exhibit strong content validity (see Table 2).

5 User evaluations

5.1 Experiment design

In Sect. 3, we have discussed four independent variables (NIR, NCR, UC, and CC) configurable in a critiquing-based recommender system (see Table 1). Each of them may have two or more options. In order to reduce the complexity of our experiment setup but still be capable of revealing these variables' respective impacts via user studies, we have conducted three trials (see Table 3).

User Trial 1

This trial was designed to identify the general performance difference between DynamicCritiquing (DC) and ExampleCritiquing (EC), so as to understand which typical system design would be more effective regarding the measured objective and subjective variables. The results would predict some underlying benefits of giving user control over one or more design variables.

Table 3 The experiment design of three user-trials

User Trial 1	DynamicCritiquing (DC: NIR=1, NCR=1, user-initiated UC, system-suggested CC)	vs.	ExampleCritiquing (EC: NIR=7, NCR=7, user-initiated UC and CC)
	vs.		vs.
User Trial 2	Modified DynamicCritiquing (MDC: NIR=1, NCR=7, user-initiated UC, system-suggested CC)	vs.	Modified ExampleCritiquing (MEC: NIR=1, NCR=7, user-initiated UC and CC)
	vs.		vs.
User Trial 3	Hybrid critiquing (HC: NIR=1, NCR=7, system-suggested CC, user-initiated UC and CC)		

User Trial 2

In the second trial, we modified DC and EC (respectively termed as MDC and MEC in this trial) to make them different only on their critiquing aids, in order to distinguish this single element's impact. Thus, modifications were made on their critiquing coverage (i.e., NIR and NCR), which were changed constant between the two systems. Moreover, each system was modified on one sub-variable (either NIR or NCR) so that it was feasible to compare the modified version with its original one in respect of the influence of the single sub-variable's change.

User Trial 3

A hybrid critiquing aid (abbreviated as HC) was evaluated in this trial to measure users' critiquing behavior when both types of aids (system-suggested and user-initiated) were presented to them on the same screen. Additionally, two between-subjects analyses, in combination with second trial's results, were done to show the respective effects of system-suggested compound critiques and user-initiated critiquing facility on users' decision quality and subjective perceptions.

Therefore, three elements were controlled in these user-trials: NIR (1 vs. k), NCR (1 vs. k) ($k = 7$) and the critiquing aid (user-initiated UC plus system-suggested CC, user-initiated for both UC and CC, or hybrid of system-suggested CC and user-initiated critiquing facility). Note that system-suggested unit critiques (from FindMe approach) were not included because we mainly emphasized on compound-critique suggestions owing to their dynamic and explanatory strengths. The analysis of each trial's results and their combination would likely help us to realize the most effective design direction for each element.

5.1.1 Experiment procedure

The three user-trials basically obeyed the same experiment procedure. An online experiment framework was implemented, by which users would easily follow the trial and all of their actions could be automatically recorded for data analysis. Except for the fact that the evaluated systems were different between the three trials, user tasks were practically identical.

More concretely, for each participant, s/he was first asked to complete a demographic questionnaire (her/his age, gender, education, profession, online shopping experience, etc.), followed by a brief reading of the user study's objective. The participant was then pointed to the assigned system's entry and instructed to begin. The main user task was to "*find a product you would purchase if given the opportunity*". After the choice was made, the participant was asked to fill in a post-study questionnaire about her/his perceived cognitive effort, decision confidence and *trusting intentions* (see Table 2). Then the system's decision accuracy was measured by revealing all alternatives in the product catalog to the participant to see whether s/he prefers another product or stands by the choice just made using the system. If s/he was involved in a within-subjects experiment setup, the participant was further required to evaluate another system with same tasks, and finally a post-question was asked about her/his preference over which critiquing system s/he would like to use for future search and why s/he preferred it to another.

5.1.2 Participants and product catalogs

Groups of subjects participating in these user-trials were randomly recruited from the same population range (Master and PhD students in our university), so they represented a similar demographical distribution (see Table 4). As to product catalogs, all evaluated critiquing systems were developed with two product datasets: a tablet PC catalog comprising 55 products each described by 10 main features (manufacturer, price, processor speed, weight, etc.), and a digital camera catalog of 64 products each characterized by 8 main features (manufacturer, price, resolution, optical zoom, etc.). All products were extracted from a real e-commerce website.

In the following, we describe each trial's detailed setup and main findings.

5.2 User Trial 1: DynamicCritiquing vs. ExampleCritiquing

DynamicCritiquing and ExampleCritiquing are the names of the two compared typical applications. We could term them respectively as *single-item system-suggested*

Table 4 Demographical distributions of three trials' participants

	User trial 1 (36 subjects)	User trial 2 (36 subjects)	User trial 3 (18 subjects)
Gender	Male (86%), Female (14%)	Male (78%), Female (22%)	Male (94%), Female (6%)
Average age	25.69	22.92	25
Nationality	12 countries (Switzerland, Romania, Spain, etc.)	10 countries (Switzerland, France, Italy, etc.)	5 countries (Switzerland, Spain, Indian, etc.)
Current education	Master (75%), Ph.D. (25%)	Master (81%), Ph.D. (19%)	Master (72%), Ph.D. (28%)
Online shopping experience	Yes (81%), No (19%)	Yes (89%), No (11%)	Yes (78%), No (22%)

compound critiquing system and *k-item user-initiated critiquing system*, but for the sake of simplicity, the abbreviations of their names are used henceforth (i.e., DC and EC).

5.2.1 Setup

As introduced in Sect. 2 the initial interaction with both DC and EC is identical with a preference specification page to obtain users' initial preferences. Then in DC, a single item that best satisfies the stated preferences is shown at the top, accompanied by a user-initiated unit critiquing area and three system-suggested compound critiques (see Fig. 2). Once a critique is posted, a new item will be returned with updated critique suggestions. In EC, seven products that best match users' initially specified preferences will be returned. If a user finds her target choice among the seven items, she can proceed to check out. However, if she likes one product (called the reference product) but wants some of its aspects improved, she can proceed to the critiquing interface to create her critiquing criteria (see Fig. 3). Subsequently, a new set of seven items will be recommended for the user to compare with the reference product.

In both systems' interfaces, users can view the product's detailed specifications with a "detail" link. Users can also save all satisfactory solutions in a "saved list" to facilitate comparing them before checking out with a final choice.

The user study was conducted in a within-subject design. Each participant evaluated the two applications one after the other. In order to avoid any carryover effect, we developed four (2 x 2) experiment conditions. The manipulated factors are systems' order (DC first or EC first) and product catalogs' order (tablet PC first or digital camera first). About 36 participants were evenly distributed into the four experiment conditions, resulting in a sample size of 9 subjects per condition cell. The same administrator supervised the experiment for all of the participants.

5.2.2 Results analysis

The result analysis would let us know which typical combination of control variables (DC: NIR = 1, NCR = 1, user-initiated unit critiquing and system-suggested compound critiques; and EC: NIR = 7, NCR = 7, user-initiated unit and compound critiquing) could achieve a better result in terms of the measured variables: decision accuracy, decision effort and *trusting intentions*. The analysis tool is paired samples *t*-test, with estimated power of 83%[1] inferring that 83% chance would be expected to yield a significant difference for each dependent variable if it exists.

Table 5 shows all of the mean values with standard deviations and degrees of freedom, and Fig. 5 illustrates differences respecting subjective perceptions which were rated on the same scale.

[1] The power was computed given the recruited sample size (i.e., 36 subjects) and assumptions of "medium" effect size and two-tailed 0.05 Alpha (see http://www.power-analysis.com/power_analysis.htm).

Table 5 Comparison of DC and EC regarding all of the measured variables

	Mean (SD)		p value (df = 35)
	DC (NIR = 1, NCR = 1, user-initiated UC, system-suggested CC)	EC (NIR = 7, NCR = 7, user-initiated UC and CC)	
Decision accuracy	47.2%(0.51)	86.1%(0.35)	**0.002**
Perceived accuracy	3.36(0.96)	3.97(0.65)	**0.004**
Time consumption	3.91(2.46)	4.25(2.10)	0.404
Critiquing cycles	7.64(8.58)	2.08(1.89)	**0.000**
Perceived effort	2.47(0.93)	2.14(0.76)	**0.053**
Intention to purchase	3.31(0.89)	3.78(0.72)	**0.005**
Intention to return	3.43(1.07)	4.11(0.93)	**0.001**

The bold values indicate that they are significant at the 0.1 level

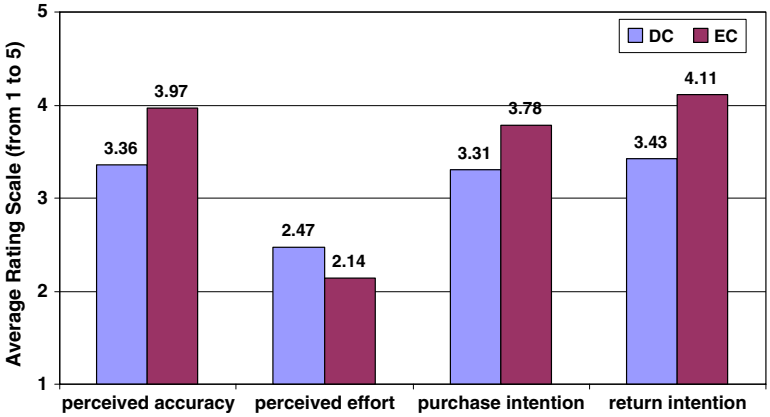

Fig. 5 Subjective perceptions with DC and EC

Decision accuracy and decision effort

The decision accuracy of EC was shown to be significantly different ($p < 0.01$, $t = 3.39$) from that of DC. Actually, 86.1% of participants found their target choice using EC. DC allowed a relatively lower decision accuracy of 47.2%, given that the remaining 52.8% users switched to a different, better choice when they were provided the opportunity to view all products in the catalog.

Moreover, participants were more confident that the products they "purchased" with EC were really the best choice for them (3.97 against 3.36 with DC on the 5-point Likert scale, $p < 0.01$, $t = 3.11$), inferring that they truly perceived EC to provide a higher level of decision accuracy.

It was then interesting to know how much effort users expended in achieving the corresponding accuracy. As introduced in Sect. 4, the decision effort was measured by two aspects: the objective effort including task completion time and interaction effort, and the subjective effort psychologically perceived by users.

The average task completion time was 4.25 min with EC versus 3.91 min with DC, but this slight difference is not significant ($p = 0.4$, $t = 0.84$). As to the objective interaction effort, we mainly measured the *critiquing cycles* referring to how many times users consulted with the critiquing aid to refine their preferences. The results indicate that the participant was on average involved in 2.08 critiquing cycles with EC, compared to 7.64 cycles with DC ($p < 0.001$, $t = -4.21$).

On the other hand, users perceived EC easier to use in finding information and more efficient in looking for a product, resulting in a significantly lower cognitive effort of 2.14 versus 2.47 with DC ($p = 0.05$, $t = 2$).

Computation of the correlation between perceived accuracy and perceived effort indicated that they are significantly negatively associated (correlation $= -0.464$, $p < 0.01$), implying that once users experienced more accuracy benefited from the critiquing system, they may perceive less cognitive effort even though more objective effort was actually spent in making the choice.

Trusting intentions

As for two *trusting intentions*, although both systems obtained positive appraises, the mean rates for EC are all significantly higher.

Concretely, participants on average indicated higher level of intention to purchase the product that they chose in EC, had they been given the opportunity (3.78 against 3.31 in DC, $p < 0.01$, $t = 3.01$), and higher level of intention to return to EC for future use (4.11 versus 3.43, $p < 0.001$, $t = 3.68$; see Fig. 5). The results infer that EC is more likely to convince its users to purchase products and to establish a strong long-term relationship with the users.

5.2.3 User comments

Participants' responses to the final post-question (about their preference over which system they would like to use in the future) show that most participants (63.9%) subjectively preferred EC to DC. Each participant was further required to write her/his brief voting reasons, so it was possible to analyze these written protocols to reveal EC and DC's respective advantages.

Each written comment was broken into episodes and each episode contained at most one concept. For EC, there are in total 24 episodes, among which 45.8% (11/24) were favorable arguments about EC's critiquing coverage. That is, it returned more results during each recommendation cycle than DC, making participants feel "easier to get an overview of all the different products", "easier to compare between products", and "easier to find a product that suits my needs". In the remaining episodes, 20.8% (5/24) were that EC overall gave users a feeling of "having more control" and "freedom". 12.5% (3/24) were particularly related to the critiquing aid (e.g., "there were more choices and options for optimizing my choices", "the value comparison[2] is nice"). 12.5% (3/24) were attributed to the higher decision confidence with EC, and 8.3%

[2] The "value comparison" is the button clickable to activate the critiquing aid.

Table 6 Participants' favorable arguments for DC and EC

Main reasons of voting for DC (13 votes)	Main reasons of voting for EC (23 votes)
Favoring system-suggested compound critiques (5/13)	More items were displayed at a time (11/24)
Easier to use and more intuitive (5/13)	More freedom and control (5/24)
Higher decision confidence (2/13)	Favoring user-initiated critiquing aid (3/24)
Faster (1/13)	Higher decision confidence (3/24)
	Missing product features in DC (2/24)

(2/24) blamed DC on its missing product features (e.g., digital camera lacks memory card information).

As for the main reasons behind favoring DC, 13 episodes were collected and 38.5% (5/13) were associated with its system-suggested compound critiques (e.g., "I liked the option to refine searches with the three proposed criteria at the bottom of the page"). Another 38.5% (5/13) appreciated the ease of use of DC ("more intuitive", "less overwhelming", "more clear", etc.), and remaining 15.4% (2/13) and 7.7% (1/13) were respectively related to the feeling of higher decision confidence ("I really find what I wanted") and "faster" accessing speed.

Table 6 summarized all of the mentioned aspects and their contributions to each system's success. It can be seen that the advantages of EC were mainly placed on its critiquing coverage (k-item display strategy) and user-initiated critiquing aid and those of DC were grounded on its suggested compound critiques and simple interface design.

5.2.4 Discussion

Thus, this user-trial revealed the performance difference of two typical critiquing systems (DC and EC) which are respectively of varied values on control variables. Results show that EC (k-item user-initiated critiquing) outperformed DC (single-item system-suggested compound critiquing) on most measured variables: objective/subjective accuracy, objective interaction effort and perceived effort, and two *trusting intentions*.

Further analysis of users' written protocols uncovered their respective advantages. In particular, the primary factor leading to EC's success would be its combination of both k-item strategy and user-initiated critiquing aid, which gave users a higher degree of control in comparing products and composing critiquing criteria. On the other hand, DC's compound critique suggestions and simple interface design were also favored by a certain percentage (around 1/3) of participants.

In the next two trials, we have aimed at identifying the exact role of each independent element and exploring the best way of combining the two systems' strengths so as to allow for further improvements on users' decision performance.

5.3 User Trial 2: Modified DC vs. Modified EC

In the second trial, both DC and EC were modified to return the same amounts of NIR (i.e., the number of recommendations in the first round) and NCR (i.e., the number

of recommendations in each of later critiquing cycles), so that the two systems were kept different only on their critiquing aids. Specifically, each system was changed on one sub-variable of critiquing coverage: DC was modified to return seven items after each critiquing process (NCR = 7), and EC was modified to show one item during the first recommendation cycle (NIR = 1). The modified versions (MDC and MEC) were hence assigned with equal NIR (i.e., 1) and NCR (i.e., 7).

As to the critiquing aid, no change was made, so MDC provides user-initiated unit critiquing plus system-suggested compound critiques (as in the original DC), and MEC supports purely user-initiated critiquing (as in EC).

Therefore, given this experiment design, we could not only reveal the single impact of critiquing aid through the comparative measurement of MDC and MEC, but also identify effective designs respectively for NIR and NCR by comparing the modified versions with their original ones which were evaluated in the first trial.

Moreover, the reason of normalizing NIR on 1 and NCR on 7 (rather than in the opposite way) was essentially driven by the assumption that users may attach more importance and attraction to tradeoff alternatives (according to their critiquing criteria) than the initial recommendation (according to their initial preferences), since their initial preference can be uncertain and erroneous (Pu and Kumar 2004). In addition, because the emphasis of this study is the "critiquing aid", it makes sense to introduce its function to users as early as possible (such as in the first recommendation cycle to be right along with one recommended item), so that users would likely be well informed and motivated to apply it whenever they think it is necessary.

5.3.1 Setup

The entry to both MDC and MEC is still a preference specification page to get the user's initial preferences, and then one product that best matches the stated preferences will be returned. In MDC, this product is accompanied by a user-initiated unit critiquing area and a list of three compound critique suggestions (like Fig. 2), and in MEC, it is followed by a user-initiated critiquing panel for the user to freely create her own critiques (like Fig. 3). The user can either choose this recommended product and "check out", or make critiques in the accompanying critiquing area. In the latter condition, MDC and MEC will then both return a set of seven items as tradeoff alternatives best satisfying the user's critiquing criteria. The user could continue to perform critiques based on one product selected from these items (by clicking the button "Value Comparison" to evoke the critiquing aid).

The second trial followed the same experiment design as the first one: a within-subject design. About 36 new participants were recruited (see Table 4), and they were evenly assigned to one of the four experiment conditions: (MDC first or MEC first) x (tablet PC first or digital camera first).

5.3.2 Results analysis

The second user trial was targeted to identify which specific critiquing aid design could be more effective in positively affecting users' decision accuracy, decision effort and trusting intentions. We also measured participants' actual critiquing application

Table 7 Comparison of MDC and MEC regarding all of the measured variables

	Mean (SD)		p value (df = 35)
	MDC (NIR = 1, NCR = 7, user-initiated UC, system-suggested CC)	MEC (NIR = 1, NCR = 7, user-initiated UC and CC)	
Decision accuracy	52.8%(0.51)	47.2%(0.51)	0.535
Perceived accuracy	3.67(0.83)	3.50(0.74)	0.350
Time consumption	2.68(1.93)	3.14(3.18)	0.202
Critiquing cycles	1.44(1.70)	1.58(1.15)	0.576
Perceived effort	2.38(0.97)	2.57(0.91)	0.274
Intention to purchase	3.5(0.77)	3.28(0.78)	0.186
Intention to return	3.54(0.96)	3.40(1.01)	0.360

respectively with the two compared critiquing aids. The paired samples t-test was still used to analyze the user data (see Table 7 for measured variables' mean values, standard deviations and degrees of freedom).

Similar to the trial 1, this study was also estimated with power 83% (conditional on its sample size and assumed "medium" effect size), which indicates a high likelihood that it would detect a significant effect provided one exists.

Critiquing application

In MEC, around 88.9% of participants consulted with the user-initiated critiquing support to specify their tradeoff criteria, and the remaining 11.1% participants chose the first recommended product as their choice (without any critiquing action). In MDC, 72.2% participants performed critiquing at least once.

Moreover, the in-depth analysis of unit and compound critiquing application in MDC shows that users were more frequently self-initiated to build unit critiques than selecting suggested compound critiques (the average application time of UC is 0.86 vs. 0.58 of CC, $t = 1.19$, $p = 0.24$). In MEC, the application frequency of the two types of critiques, however, is much closer (0.64 vs. 0.58, $t = 0.25$, $p = 0.80$), and there were some participants just searching for "similar products" without concrete critiquing criteria (average application time = 0.34).

Decision accuracy, decision effort and trusting intentions

In terms of all the dependent variables contained in our evaluation framework, the experimental results, surprisingly, indicated that there is no significant difference between MDC and MEC. More specifically, regarding objective and subjective decision accuracy, the two systems reached similar levels. The objective accuracy in MDC is 52.8%, against 47.2% in MEC ($t = 0.63$, $p = 0.53$), and the perceived decision accuracy is respectively 3.67 and 3.5 ($t = 0.95$, $p = 0.35$).

Participants in MDC and MEC also consumed nearly equal amount of objective and subjective effort. For instance, the average difference of task time consumption between the two systems is only 0.45 seconds (2.68 with MDC vs. 3.14 with MEC,

$t = -1.3$, $p = 0.20$), and the difference respecting critiquing cycles is 0.14 (1.44 vs. 1.58, $t = -0.56$, $p = 0.58$). Perceived effort is slightly higher with MEC but still not at a significant level (2.57 vs. 2.38 with MDC, $t = 1.11$, $p = 0.27$).

As for two *trusting intentions*, both systems obtained positive responses. That is, the user on average intended to purchase the chosen product in MDC and MEC (3.5 vs. 3.28, $t = 1.35$, $p = 0.19$), and to return to the system for repeated uses (3.54 to MDC vs. 3.40 to MEC, $t = 0.93$, $p = 0.36$). The rates on MDC are all slightly higher but without significant phenomena.

5.3.3 User comments

At the end of the trial, each participant was asked about her/his preference over the critiquing interface design ("comparing the two interfaces you just used, which interface design do you relatively prefer to use?"), given that it is the only difference between the two compared systems.

It was shown that 21 out of 36 (58.3%) participants voted MDC, and the remaining 41.7% preferred MEC. Analysis of users' written protocols showed that the major reason (9/21 = 42.9%) behind favoring MDC was due to its compound critique suggestions (see Table 8), which made the interface "interesting", "more useful", "easier to use" and helped users "access to what they want quickly". In the remaining favorable episodes, five (out of 21) were general opinions on the interface's ease of use and usability, five were attributed to the product domain (e.g., "because I am more interested in a computer than a digital camera") and two were negative impressions of MEC ("it was not practical" and "it did not give me exactly the kind of product I wanted").

The reason behind voting for MEC was largely placed on its user-initiated critiquing facility (10/15 = 66.7%). Subjects felt that "it allowed for very detailed refinements", "gave the chance to refine search in a more intuitive way", enabled them to "have more control over the new search terms" and was "quicker to go through many products". The remaining four episodes were respectively about the interface's ease of use (2/15), the product domain (2/15) and the negative impression of MDC (1/15).

Table 8 Participants' favorable arguments for MDC and MEC

Main reasons of voting for MDC (21 votes)	Main reasons of voting for MEC (15 votes)
Favoring system-suggested compound critiques (9/21): more options, global view of products' characteristics, useful, enable to access to products more quickly, etc.	Favoring user-initiated critiquing aid (10/15): support detailed refinement, more intuitive to refine, give more control over search, easier to adjust parameters, with the "improve" option, etc.
Easier to compare products and easier to understand (5/21)	Easier to use and easier to find the product's information (2/15)
Familiar with the product domain (5/21)	Familiar with the product domain (2/15)
Negative impression of MEC (2/21): not practical, inaccurate recommendations	Negative impression of MDC (1/15): take long to change preferences

Therefore, users' qualitative comments imply that system-suggested critiques and user-initiated critiquing aid both provide substantial advantages, which should be why the corresponding two systems (MDC and MEC) performed equally as actively in influencing users' decision quality and subjective assessments. In addition, since MDC also contains user-initiated unit critiquing and for most measures it performed slightly (but not significantly) better than MEC, it infers that the combination of both the *user-initiated* and *system-suggested* critiquing options would potentially obtain more benefits. We experimentally explored this issue in the third trial.

5.3.4 MDC vs. DC and MEC vs. EC

As mentioned in the beginning of this section, another goal of this user-trial was to identify the effect of single change on critiquing coverage through the comparison of the modified version with its original one (e.g., MDC vs. DC). Since participants in trials 1 and 2 were recruited from a similar population range and they followed the same experiment procedure, it was feasible to do two between-group analyses (MDC vs. DC, and MEC vs. EC) (a statistical application of "two trials plus between-subjects effects" as described by Hopkins (1997)).

Tables 9 and 10 respectively show the comparison results of MDC and DC, and of MEC and EC. For each system, 18 participants who used it at their first order were considered in order to avoid any carryover biases. All of the significant values (p) were computed by Student t-test assuming unequal variances.

The only difference between MDC and DC is on their NCR (the number of recommended items after each critiquing action). MDC increased it from one to seven. The results indicate that due to this change, participants expended significantly less time and effort in making their final choice. Regarding the other variables such as objective accuracy, decision confidence and *trusting intentions*, there was no significant influence.

The change from EC to MEC was the decrease of the number of the first round's recommendations (NIR) from seven to one. Comparison analysis shows that this decrease significantly impacted subjects' objective/subjective decision accuracy, perceived effort and two *trusting intentions* in a negative manner, while the task time was reduced. Therefore, it implies that the first set of items recommended according to the user's initial preferences should be a very important factor positively influential to the user's subjective perceptions of the system, which is contrary to our assumption when setting NIR as 1 in the experiment design.

5.3.5 Discussion

The second trial mainly showed that when both systems (DC and EC) were different only on their critiquing aids, users on average performed nearly identically in both conditions. More specifically, the *user-initiated unit critiquing plus system-suggested compound critiques* (MDC) and *user-initiated unit and compound critiquing support* (MEC) enabled participants to reach similar levels in terms of decision accuracy, decision effort and *trusting intentions*. Users' written protocols qualitatively revealed their respective strengths: MDC provides suggestions that accelerated users' decision

Table 9 The comparison of DC and MDC (mean and SD for each dependent variable)

	Decision accuracy		Decision effort			Trusting intentions	
	Objective accuracy	Perceived accuracy	Task time (mins)	Critiquing cycles	Perceived effort	Purchase intention	Return intention
DC (NCR=1)	33.3% (0.49)	3.5 (0.62)	0.5 (2.75)	9.89 (9.86)	2.67 (0.94)	3.17 (0.86)	3.36 (0.97)
MDC (NCR=7)	50% (0.51)	3.5 (0.92)	3.22 (2.2)	1.5 (1.5)	2.39 (1.06)	3.22 (0.88)	3.44 (1.11)
p value (df)	0.324 (34)	1 (30)	**0.039** (32)	**0.002** (18)	0.413 (33)	0.849 (34)	0.812 (33)

The bold values indicate that they are significant at the 0.1 level

Table 10 The comparison of EC and MEC (mean and SD for each dependent variable)

	Decision accuracy		Decision effort			Trusting intentions	
	Objective accuracy	Perceived accuracy	Task time (mins)	Critiquing cycles	Perceived effort	Purchase intention	Return intention
EC (NIR=7)	77.8% (0.43)	4.06 (0.42)	4.33 (2.2)	1.44 (1.42)	1.86 (0.7)	3.89 (0.68)	4.42 (0.86)
MEC (NIR=1)	38.9% (0.50)	3.33 (0.69)	2.88 (1.28)	1.56 (0.98)	2.69 (1.02)	3.11 (0.76)	3.39 (1.11)
p value (df)	**0.017** (33)	**0.001** (28)	**0.023** (27)	0.787 (30)	**0.008** (30)	**0.003** (34)	**0.004** (32)

The bold values indicate that they are significant at the 0.1 level

process and made the critiquing action easier, and MEC allows for higher user-control and detailed preference refinement.

Combining with the first trial's results, we become clearer about the key factor for EC's success in the first trial. That is, it should be largely attributed to its multi-item display strategy against single-item in DC, given the fact that the difference on their critiquing aids did not produce any significant impacts as shown in the second study. Actually, with the two trials' user data, we found that the increase of NCR can significantly reduce users' objective effort including time consumption and critiquing cycles, and the decrease of NIR can significantly impair decision accuracy and all of measured subjective perceptions. Therefore, it infers that both NCR and NIR should be ideally kept at k ($k > 1$) as in the original EC.

5.4 User Trial 3: hybrid critiquing aid

The final user trial investigated how to further improve on the critiquing interface, in consideration of the respective advantages of *system-suggested critiques* and *user-initiated critiquing* derived from above two user-trials' results. Analysis of participants' comments revealed that the best approach would be to synthesize them in a single system, so that the hybrid critiquing aid would support an optimal level of user-control: users can have the freedom to choose whether specifying their own critiquing criteria, or selecting the suggested critiques if one of them matches their desires.

Therefore, in this trial, we measured users' critiquing behavior in such hybrid critiquing system (HC) that combines the critiquing aids from both DC and EC on the same screen. The hypothesis was that the hybrid system would outperform the uncombined exclusive approaches, since it enables users to have more freedom in choosing the type of critiquing support they are willing to use in a certain situation.

5.4.1 Setup

Figure 6 shows the sample of the hybrid critiquing interface. The proposed critiques are listed under the critiqued product and the bottom is the user-initiated critiquing area with functions to facilitate creating either unit or compound critiques by users themselves. Once a critique was posted, the recommender algorithm is run adaptive to the type of critiques users made. Specifically, it applies similarity and compatibility selection strategy if the *dynamic-critiquing* based critique suggestion was picked (McCarthy et al. 2005c), and employs EBA+WADD (elimination-by-aspect plus weighted additive sum rule) ranking mechanism if the critique was composed in the user-initiated critiquing panel.

The system returns one initial recommendation (NIR = 1) and seven items after each critiquing (NCR = 7) as MDC and MEC did, so that it is feasible to compare it with both MDC and MEC only in respect of their critiquing aids' difference. Among the recommended item(s), if the user finds her target choice, she can proceed to check out. Otherwise, if she likes one product but wants some values improved, she can resume a new critiquing cycle. Similar to previously implemented systems, the hybrid

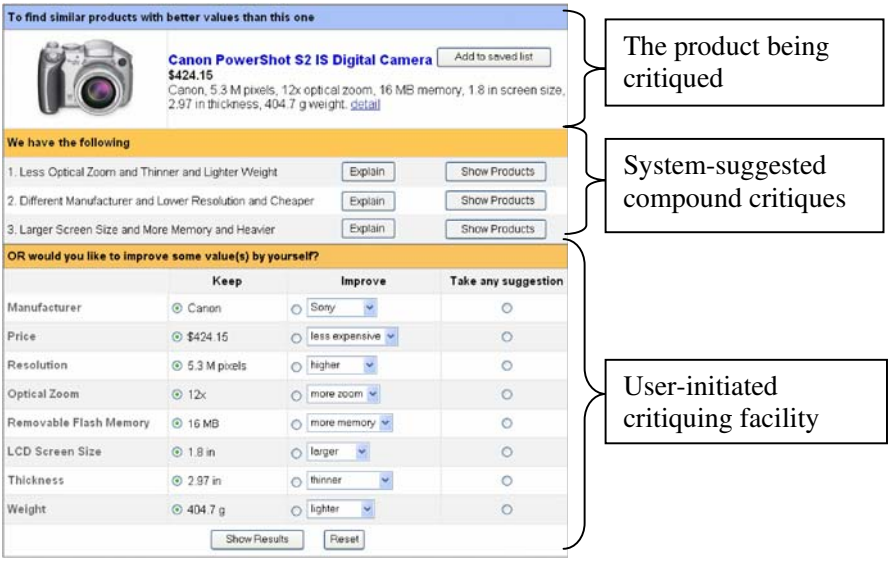

Fig. 6 A hybrid critiquing interface with both system-suggested compound critiques and user-initiated critiquing facility

one also provides the product's detailed specifications accessible by a "detail" link and a "save list" for the user to record products that interest her.

We randomly recruited 18 new volunteers from the same population range as in trial 1 and 2 (see Table 4). Each of them was required to evaluate one system: the *hybrid critiquing* with the main task of "*find a product you would purchase if given the opportunity*". Each participant was randomly assigned one product domain (tablet PC or digital camera) to search. After the choice was made, the participant was asked to fill in a post-study questionnaire about her/his perceived cognitive effort, decision confidence, and *trusting intentions*. Then her/his objective decision accuracy was measured by revealing all products to the participant to determine whether s/he prefers another product in the catalog or stands by the choice that was just made with the hybrid critiquing system.

5.4.2 Results analysis

Critiquing application

Among the 18 participants, 88.9% conducted self-initiated critiquing and 44.4% picked the compound critique suggestions at least once. On average, the application time of user-initiated critiquing per user is 2.5 against 1.1 of system-suggested compound critiques ($t = 2.11$, $p < 0.01$ by paired-samples t-test). In addition, around 36% of user-initiated critiques were compound critiques that involved maximal 7 features at a time, 55.6% were unit critiques (one feature to be improved or compromised) and 8.9% were similarity-based critiquing without exact criteria.

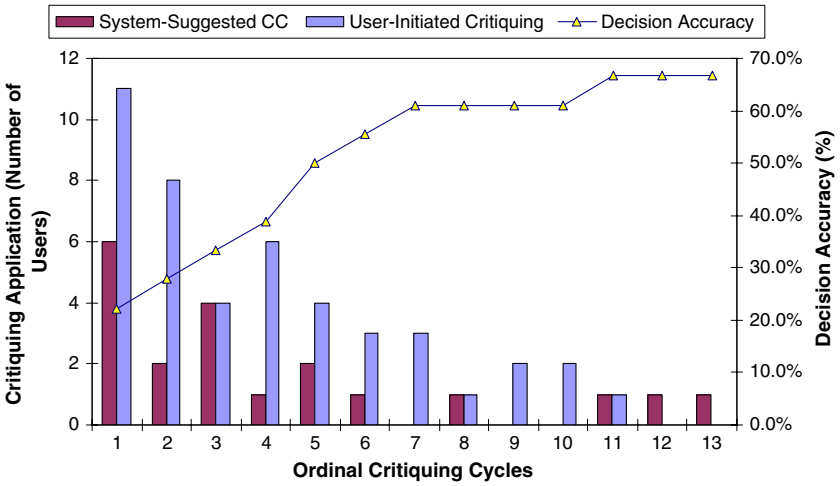

Fig. 7 Critiquing application in the hybrid critiquing interface on a per cycle basis

Figure 7 illustrates the critiquing application frequency on a per cycle basis. The left vertical axis is the number of users who applied system-suggested compound critiques or user-initiated critiquing facility in the corresponding cycle. It refers to those people who did not stop before that cycle and continued making critiques. The right vertical axis is the aggregated decision accuracy. It can be observed that during 84.6% (11/13) of maximal critiquing cycles, the number of users who created critiques on their own is all greater than (during 8 cycles) or equal to (3 cycles) the number of ones picking suggested critiques.

Another finding is that 83.3% of participants ended their session by utilizing the self-initiated critiquing feature. It infers that system-suggested critiques may be more useful in the earlier cycles when users are less certain about their preferences or have a superficial understanding of the product domain. Later on, once users obtain a certain degree of product knowledge and what they truly want, they will be more likely self-motivated to make their own critiques that ultimately lead to their final choice.

Decision accuracy, decision effort and trusting intentions

The objective decision accuracy in the hybrid system was 66.7%, because 12 participants (out of 18) stuck with their choice by using the system when they had a chance to view all of the alternatives (see Fig. 7 for the aggregated of decision accuracy with the increase of critiquing cycles). We further examined the accuracy distribution corresponding to users' critiquing application. The results indicate that 50% of decision accuracy was contributed from participants who performed both system-suggested CC and self-initiated critiquing, 41.67% from ones only applying self-initiated critiquing and 8.33% from those who did not make any critiquing (the initial recommended item

was their choice). This distribution exhibits a significant phenomenon ($p = 0.03$ by Chi-square test).

As for the perceived decision accuracy, the average rate is above 3 indicating that most of users (i.e., 88.9%) were confident that their choice was the best with the hybrid critiquing system (mean = 4, median = 4).

In terms of the objective decision effort, the participant on average consumed 5.52 min and 2.83 critiquing cycles. Moreover, responses to questions related to perceived effort show that they subjectively perceived a low level of cognitive effort (mean = 2.06, median = 2).

Analysis of users' answers to *trusting intentions* indicated that most of participants expressed positive intention to purchase their chosen products (61.1% users, mean = 3.44, median = 4) and positive intention to return to the system for future use (77.8% users, mean = 4.06, median = 4.5).

5.4.3 HC vs. MDC and HC vs. MEC

The hybrid system was further compared with MDC and MEC due to their exclusive differences on the critiquing aid (NIR = 1, NCR = 7 in these three systems). Concretely, 18 subjects who used MDC and 18 who used MEC at their first order were respectively compared with the 18 participants using the hybrid critiquing system (HC). Two between-group analyses were conducted using Student t-test assuming unequal variances.

HC's only defining difference from MDC is that it provides user-initiated critiquing facility for creating compound critiques (U-CC) while MDC does not, and the only difference from MEC is that it contains system-suggested compound critiques (S-CC) but MEC does not. Therefore, by comparing HC with MDC and MEC respectively, we could reveal the respective role of U-CC and S-CC in the hybrid system, and more importantly understand whether HC could perform better than both MDC and MEC since it provides a combination of their critiquing aids.

Table 11 lists the comparison results of HC and MDC, which show that owing to the add-on element U-CC in HC, users exhibited significantly higher decision confidence and return intention, although they spent more time and critiquing cycles. The application frequency of critiquing facilities that are provided by both systems did not significantly vary (system-suggested CC: 1.11 in HC vs. 0.61 in MDC, $p = 0.12$; user-initiated UC: 0.78 in HC and 0.89 in MDC, $p = 0.74$), inferring that participants did take extra time and critiquing effort with U-CC while using HC, which may directly lead to their increased accuracy perception and intention to return.

The comparison between HC and MEC (see Table 12) also shows similar results regarding S-CC. That is, its appearance stimulated users to reach significantly higher subjective perceptions including decision confidence and return intention, although more time and critiquing effort were expended. The extra objective effort was also found mostly consumed with S-CC, since the user-initiated critiquing that is supported by both systems was applied at around equal frequency (1.72 in HC vs. 1.56 in MEC, $p = 0.64$). The objectively consumed effort, however, did not affect users' subjective effort perception. The perceived effort was in fact significantly lower in HC

Table 11 The comparison of MDC and HC (mean and SD for each dependent variable)

	Decision accuracy		Decision effort			Trusting intentions	
	Objective accuracy	Perceived accuracy	Task time (mins)	Critiquing cycles	Perceived effort	Purchase intention	Return intention
MDC (without U-CC)	50%(0.51)	3.5(0.92)	3.22(2.2)	1.5(1.5)	2.39(1.06)	3.22(0.88)	3.44(1.11)
HC (with U-CC)	66.7% (0.49)	4 (0.49)	5.52 (3.67)	2.83 (2.28)	2.06 (0.68)	3.44 (0.86)	4.06 (0.92)
p value (df)	0.324 (34)	**0.052** (26)	**0.030** (28)	**0.048** (29)	0.273 (29)	0.447 (34)	**0.081** (33)

The bold values indicate that they are significant at the 0.1 level

Table 12 The comparison of MEC and HC (mean and SD for each dependent variable)

	Decision accuracy		Decision effort			Trusting intentions	
	Objective accuracy	Perceived accuracy	Task time (mins)	Critiquing cycles	Perceived effort	Purchase intention	Return intention
MEC (without S-CC)	38.9% (0.50)	3.33 (0.69)	2.88 (1.28)	1.56 (0.98)	2.69 (1.02)	3.11 (0.76)	3.39 (1.11)
HC (with S-CC)	66.7% (0.49)	4 (0.49)	5.52 (3.67)	2.83 (2.28)	2.06 (0.68)	3.44 (0.86)	4.06 (0.92)
p value (df)	0.100 (34)	**0.002** (31)	**0.009** (21)	**0.040** (23)	**0.035** (30)	0.225 (34)	**0.058** (33)

The bold values indicate that they are significant at the 0.1 level

Table 13 Correlations between objective and subjective variables (by Pearson's Correlation)

	Objective accuracy	Perceived accuracy	Task time	Critiquing cycles	Perceived effort	Purchase intention	Return intention
Objective accuracy	1	0.361***(0.000)	0.138*(0.081)	−0.094(0.233)	−0.310***(0.000)	0.319***(0.000)	0.293***(0.000)
Perceived accuracy	0.361***(0.000)	1	0.087(0.270)	−0.032(0.682)	−0.533***(0.000)	0.476***(0.000)	0.521***(0.000)
Task time	0.138*(0.081)	0.087(0.270)	1	0.392***(0.000)	0.057(0.472)	0.033(0.678)	0.157**(0.047)
Critiquing cycles	−0.094(0.233)	−0.032(0.682)	0.392***(0.000)	1	0.146*(0.064)	−0.032(0.686)	−0.050(0.523)
Perceived effort	−0.310***(0.000)	−0.533***(0.000)	0.057(0.472)	0.146*(0.064)	1	−0.405***(0.000)	−0.620***(0.000)
Purchase intention	0.319***(0.000)	0.476***(0.000)	0.033(0.678)	−0.032(0.686)	−0.405***(0.000)	1	0.336***(0.000)
Return intention	0.293***(0.000)	0.521***(0.000)	0.157**(0.047)	−0.050(0.523)	−0.620***(0.000)	0.336***(0.000)	1

***Correlation is significant at the 0.01 level (2-tailed); ** at the 0.05 level (2-tailed); * at the 0.1 level (2-tailed)

than in MEC, inferring that the integration of system-suggested critiques will likely save users' cognitive critiquing effort.

5.4.4 Discussion

The final user-trial studied users' decision behavior in a hybrid critiquing system that combines critiquing aids from both DC and EC on the same screen. It was observed that users behaved more actively in creating their own criteria with the self-initiated critiquing aid, relative to their application of system-suggested critiques. Eventually, the hybrid critiquing system enabled its users to obtain high level of decision accuracy and subjective perceptions.

Furthermore, by respectively comparing the hybrid critiquing interface with MDC and MEC, the respective roles of user-initiated compound critiquing (U-CC) and system-suggested compound critiques (S-CC) were empirically validated. Both of them were shown to significantly contribute to enhancing users' decision confidence and return intention and enabling the hybrid system to outperform MDC and MEC with respect to the two important subjective aspects.

5.5 Other results: relationships between objective and subjective variables

We collected totally 90 users' data from the three trials, including their objective/subjective decision accuracy, objective/subjective decision effort, and *trusting intentions* with the evaluated systems. Based on the collection, we calculated the correlations between the objective and subjective variables contained in our evaluation framework, with the aim to see whether they were significantly related to each other, and especially whether the improvement on users' objective performance could have significant impact on their subjective perceptions. In particular, it would be interesting to know whether objective decision accuracy and effort are respectively positively associated with users' subjectively perceived accuracy and cognitive effort, and how subjective accuracy/effort further influences users' *trusting intentions* with the system.

Table 13 gives the coefficient values by Pearson's Correlation. It shows that most of the variables were significantly positively or negatively correlated, except the relationship between objective decision effort and some subjective perceptions. Specifically, both task time and critiquing cycles did not show significant correlations with perceived accuracy and purchase intention. Moreover, there is no significant relationship between task time and perceived effort, between critiquing cycles and objective accuracy, and between critiquing cycles and return intention. These results imply that the decrease of objective decision effort is not likely to result in increases in uses' subjective perceptions, and vice versa.

We further calculated standardized path coefficients to reveal these variables' causal relationships (Fig. 8). The path regression coefficient measures the extent of the effect of one variable on another in the path model using a correlation matrix as the input.

Our results indicate that the objective decision accuracy is highly significantly associated with users' perceived accuracy ($b = 0.38$, $p < 0.01$), implying that the increased level of a system's recommendation accuracy will likely have a significantly

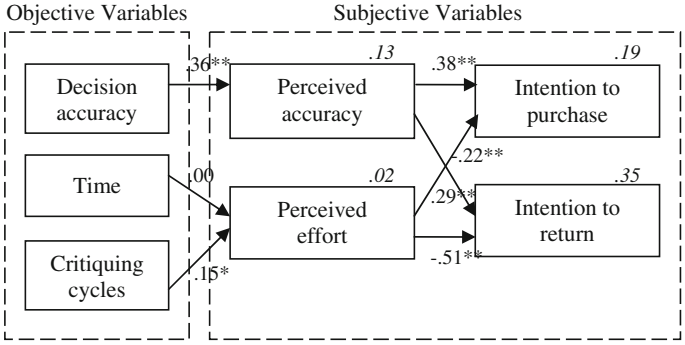

Fig. 8 Standardized path coefficients and explained variances for the measured variables (** indicating the coefficient is at the $p < 0.01$ significant level, * at the $p < 0.1$ level; explained variance R^2 appearing in italics over the box)

positive effect on users' decision confidence. Perceived decision accuracy was further found significantly positively related to users' intention to purchase ($b = 0.38$, $p < 0.01$) and intention to return ($b = 0.29$, $p < 0.01$), which suggests that if a user is more confident that she made the best choice, she will more likely purchase the chosen product and return to the recommender system for future search.

The two *trusting intentions* are also significantly influenced by the user's perceived cognitive effort ($b = -0.22$, $p < 0.01$ for intention to purchase, and $b = -0.51$, $p < 0.01$ for intention to return), indicating that the decrease of subjective effort in the decision process will also likely lead to increases in purchase and return intentions. In fact, both perceived accuracy and perceived decision effort account for approximately 19% and 35% respectively of the variance in intention to purchase ($R^2 = 0.19$) and intention to return ($R^2 = 0.35$) (both exceeding the 10% benchmark recommended by Falk and Miller (1992)). 13% of the variance in perceived accuracy ($R^2 = 0.13$) can be further explained by objective decision accuracy.

The path coefficient from actual task time to perceived effort does not show a significant relationship ($b = 0$, $p = 0.996$), and the number of critiquing cycles is marginally significantly associated with the perceived effort ($b = 0.15$, $p = 0.085$). Therefore, even though less task time is spent on the interface, it may not predict that users perceive the interface to be less demanding, whereas the saving of interaction cycles will be relatively effective to affect effort perception.

6 Practical implications

To our knowledge, our study is the first one to address user-control issues for critiquing-based recommender systems. We mainly investigated two crucial elements: the number of items that are returned during each recommendation cycle to be critiqued (i.e., critiquing coverage), and the critiquing aid by which users can specify concrete feedback criteria. Three user-trials were conducted to evaluate systems with different combinations of the two elements' possible configurations, and according to these trials' results,

we revealed their respective impacts on users' decision performance and subjective perceptions.

Other researchers have also studied control elements for some preference elicitation systems. For instance, Ariely (2000) studied the role of information flow control in consumers' decision-making and preferences. The information flow control means users can be free to choose which pieces of information (e.g., which camera and which attribute of the camera) to view and for how long. Experimental results show that controlling the information flow can help consumers better match their preferences, have better memory and knowledge about the domain and be more confident in their judgment, but it takes the risk of creating demands on processing resources and having detrimental effects on consumers' ability to utilize information. As for collaborative filtering based recommender systems, McNee et al. (2003) found that asking users to rate items out of their own selection in the sign up process resulted in more accurate user preference models and user loyalty to the system, compared to asking them to rate the system-proposed items.

However, few works have researched the role of user control in the conversational recommender system, which kind of system indeed demands high involvement from users to participate. The degree of control over the process of specifying preferences and providing feedback should be essentially associated with the levels of accuracy and subjective perceptions users can reach.

The findings of our empirical studies strongly support our overall suggestions for improving the current critiquing-based recommender system, in terms of how to design the *critiquing coverage* and the *critiquing aid*. In addition, we contribute an evaluation framework applicable for the precise measurement of a recommender system's true benefits to its users.

Critiquing coverage

Combining the results from the first and the second trials showed that recommending multiple k items ($k = 7$ in our experiments) for users to select a critiqued reference product performed more effectively against showing just one. Specifically, multiple NCR (the number of recommended items after each time users posted critiquing criteria) was found to significantly save users' task time and critiquing effort, and multiple NIR (the number of recommendations in the first cycle based on users' initial preferences) significantly improved users' objective/perceived decision accuracy, perceived effort and *trusting intentions*. Subjects also qualitatively commented that the multi-item display strategy made them feel to be of more freedom in comparing different products, choosing critiqued object and speeding up the decision process, relative to the single-item display.

Critiquing aid

As to the critiquing aid that supports the specification of concrete feedback criteria, a hybrid critiquing interface, that combines both system-suggested compound critiques and a user-initiated critiquing facility, was found to outperform the uncombined exclusive approaches particularly in enhancing users' subjective perceptions. Actually, users

reacted actively to both types of critiquing aids in the hybrid interface, and consumed a certain amount of objective effort with each of them, with the resulting benefit of obtaining a higher level of decision confidence. Moreover, they expressed stronger intention to return to the hybrid system for future use. The respective advantages of the two types of critiquing supports were also revealed: system-suggested compound critiques provide a global view of available products and make the critiquing process simpler and quicker, and the user-initiated critiquing support allows for detailed refinement of preferences and more user control over specifying users' own critiquing criteria. Therefore, the hybrid critiquing interface, where users can have much more freedom in choosing which support they would like to apply at a time, should be ideally achievable. The user-trials experimentally proved its significantly positive effect on improving two important subjective standards: perceived decision accuracy and return intention.

User evaluation framework

Another contribution of our work is the evaluation framework we have established containing key criteria for assessing a critiquing-based recommender system. It was grounded on the accuracy-effort model from the classical decision theory (Payne et al. 1993). We extended it to include more subjective variables for measuring accuracy/effort perceptions and trust-inspired behavioral intentions. More concretely, decision accuracy and decision effort are not only measured by traditional objective manners, but also subjectively defined regarding how users perceive the accuracy of their choice and the cognitive effort of information-processing. Two subjective intentions induced by user trust were also contained. One is the *intention to purchase*, measuring whether the system could convince its users to purchase the chosen product, and another is the *intention to return* about whether a long-term relationship could be potentially built between the user and the system. We believe that this evaluation framework will be certainly helpful for the evaluation of other types of online recommender systems. Related researchers may also benefit from the assessment procedures and questions we have developed through our practical experiences.

The causal relationships between objective and subjective variables were additionally identified based on all of user data collected from the three trials. Users' decision confidence was found to be positively affected by objective decision accuracy, and perceived cognitive effort was marginally influenced by the number of critiquing cycles users underwent in locating their final choice. Furthermore, increased perceived accuracy or decreased cognitive effort will likely lead to increases in intention to purchase and intention to return. Sharing these results and methods may potentially stimulate more researches in this respect.

7 Limitations & future work

In the third trial, we found that participants more actively created critiques on their own relative to their application frequency of system-suggested compound critiques, inferring that the critique suggestion may have a poor prediction on users' true needs.

Therefore, in the future, we are interested in exploring how to improve the prediction accuracy of system-proposed critiques to make them more accurately match users' desired feedback criteria, and investigating whether this improvement will in practice have significant impact on saving users' critiquing effort.

Another issue arising from our experiment setup is whether it was valid to perform comparisons (i.e., between-group analyses) among different user-trials given that they were conducted at different time (with average four or five months elapsed between). As a matter of fact, it is that the results from one trial did motivate us to conduct a following-up study in order to verify the significance of some of its interesting observations. In order to ensure the whole experiment's validity, we have obeyed standard requirements for the comparison of multiple trials, which include recruiting participants from the same range of population and asking them to follow uniform experimental procedure. In the future, it will be desirable to conduct a validation study where all of the varied conditions are comprised in a single trial.

Moreover, the role of system-suggested unit critiques has not been addressed in our studies, since our focus is mainly compound critique suggestions due to their dynamic and explanatory features. For the future, it should be of interest to perform more user evaluations investigating whether unit critique suggestions will be also helpful to be integrated in the hybrid critiquing interface.

It is also meaningful to study NIR (i.e., the number of initial recommendations) and NCR (i.e., the number of displayed items after each critiquing) for different circumstances. Future work includes further investigation of their optimal setups being adaptive to the product catalog as well as the screen size. For example, Nguyen et al.2004 suggested changing the amount of returned items dependent on the mobile device's screen length in order to reduce users' scrolling effort.

As for the user-scale of our experiments, we recruited in total 90 university students, and each of them was required to seriously imagine s/he was about to "purchase" a product with the assigned system. The demographic variety may be limited, such as the low percentage of female participants, but the experimental results provided promising design guidelines. We intend to further verify the scalability and generality of our findings by recruiting more females, and also making the source of participants as diverse as possible in terms of the age, cultural background, education and profession. We also want to integrate our technologies into a real e-commerce platform on which users could make genuine purchasing decisions.

8 Conclusion

In conclusion, in this paper, we studied how to design effective components for critiquing-based recommender systems in order to assist users in providing feedback to recommended items and guide them to efficiently target at their best choice. Through a series of three user-trials, we have investigated two crucial control elements: *critiquing coverage* and *critiquing aid*, and investigated their respective influences on users' decision accuracy, decision effort, and *trusting intentions*.

The first trial compared two typical existing applications: DynamicCritiquing that suggests one item during each recommendation cycle and a list of system-suggested

compound critiques for users to select as improvement to the recommended item, and ExampleCritiquing which returns multiple items at a time and provides a user-initiated critiquing facility to assist users in freely creating their own tradeoff criteria. Results show that ExampleCritiquing (EC) performed significantly better than DynamicCritiquing (DC) in improving users' objective/subjective accuracy, reducing their interaction effort and cognitive effort, and increasing their trust-inspired purchase and return intentions.

In the second trial, both EC and DC were modified so that the only differing factor was their critiquing aids. Although there is no significant measurement difference between the two modified versions, participants' written comments revealed their respective strengths: *system-suggested compound critiques* made users feel of obtaining more knowledge of remaining recommendation opportunities and being easier to make critiques; *user-initiated critiquing aid* allowed for detailed preference refinement and higher user-control over composing users' own searching criteria. Moreover, the significant effects of k-NIR (k recommendations in the first round) and k-NCR (k recommended items after each critiquing action) were identified ($k = 7$ in our experiments). That is, k-NIR was significantly contributive 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 should be the key factor leading to EC's success in the first trial.

The third trial evaluated user performance in a hybrid critiquing interface that combines both system-suggested critiques and user-initiated critiquing facility on the same screen. It was shown that users acted actively to both types of critiquing supports. Furthermore, in comparison with user data in the system without system-suggested compound critiques or the one without user-initiated compound critiquing support, the hybrid system enabled users to reach significantly higher levels of decision confidence and return intention.

Practical implications were finally derived from all of experimental results regarding the breadth of recommendational choice and the locus of user-initiative. We suggest that a critiquing-based recommender system should better return multiple recommended items at a time, especially during the first recommendation cycle, for users to determine their own interested critiquing object. Once users have been given such multiple choices, it should offer them additionally a hybrid critiquing interface with two types of critiquing aids: system-suggested critiques and user-initiated critiquing, by which users can freely decide whether picking one of suggested critiques (if it matches their desires) or specifying their own feedback criteria with the self-initiated support if necessary.

Acknowledgements We thank the Swiss National Science Foundation for sponsoring the reported research work. We are grateful to all participants of our user studies for their patience and time.

References

Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, USA, pp. 207–216. (1993)

Ajzen, I.: The theory of planned behavior. Organ. Behav. Hum. Decis. Process. **50**, 179–211 (1991)
Ariely, D.: Controlling the information flow: effects on consumers. Decision making and preferences. J. Consum. Res. **27**, 233–248 (2000)
Benbasat, I., Nault, B.R.: An evaluation of empirical research in managerial support systems. Decis. Support Syst. **6**(2), 203–226 (1990)
Bettman, J.R., Johnson, E.J., Payne , J.W.: A componential analysis of cognitive effort in choice. Organ. Behav. Hum. Decis. Process. **45**, 111–139 (1990)
Burke, R.: Knowledge-based recommender systems. Encyclopedia Library Inform. Syst. **69**, Supplement 32 (2000)
Burke, R., Hammond, K., Cooper, E.: Knowledge-based navigation of complex information spaces. In: Thirteenth National Conference on Artificial Intelligence, Portland, Oregon, pp. 462–468 (1996)
Burke, R., Hammond, K., Young, B.: The FindMe approach to assisted browsing. IEEE Expert: Intell. Syst. Appl. **12**, 32–40 (1997)
Carenini, G., Poole, D.: 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)
Chen, L., Pu, P.: Trust building in recommender agents. In: Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces at the Second International Conference on E-Business and Telecommunication Networks, Reading, UK, pp. 135–145 (2005)
Chen, L., Pu, P.: Evaluating critiquing-based recommender agents. In: Twenty-first National Conference on Artificial Intelligence, Boston, USA, pp. 157–162 (2006)
Einhorn, H., Hogarth, R.: Confidence in judgment: persistence of the illusion of validity. Psychol. Rev. **85**, 395–416 (1978)
Falk, R.F., Miller, N.B.: A Primer for Soft Modeling, 1st edn. The University of Akron Press, Akron, Ohio (1992)
Faltings, B., Torrens, M., Pu, P.: Solution generation with qualitative models of preferences. Int. J. Comput. Intell. Appl. **20**, 246–264 (2004)
Grabner-Kräuter, S., Kaluscha, E.A.: Empirical research in online trust: a review and critical assessment. Int J Hum–Comput Stud **58**, 783–812 (2003)
Grabner-Kräuter, S., Kaluscha, E.A., Fladnitzer, M.: Perspectives of online trust and similar constructs: a conceptual clarification. In: Eighth International Conference on Electronic Commerce, Fredericton, New Brunswick, Canada, pp. 235–243 (2006)
Häubl, G., Trifts, V.: Consumer decision making in online shopping environments: the effects of interactive decision aids. Mark Sci **19**, 4–21 (2000)
Hopkins, W.: A New View of Statistics. http://www.sportsci.org/resource/stats/index.html (1997)
Koufaris, M., Hampton-Sosa, W.: Customer trust online: examining the role of the experience with the web-site. CIS Working Paper Series, Zicklin School of Business, Baruch College, New York, NY (2002)
Linden, G., Hanks, S., Lesh, N.: Interactive assessment of user preference models: the automated travel assistant. In: International Conference on User Modeling, Chia Laguna, Sardinia, Italy, pp. 67–78(1997)
McCarthy, K., Reilly, J., McGinty, L., Smyth, B.: On the dynamic generation of compound critiques in conversational recommender systems. In: Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Eindhoven, Netherlands, pp. 176–184 (2004a)
McCarthy, K., Reilly, J., McGinty, L.,Smyth, B.: Thinking positively $\frac{1}{\mu}$ explanatory feedback for conversational recommender systems. In: Workshop on Explanation in CBR at the Seventh European Conference on Case-Based Reasoning, Madrid, Spain, pp. 115–124 (2004b)
McCarthy, K., McGinty, L., Smyth, B., Reilly, J.: A live-user evaluation of incremental dynamic critiquing. In: Sixth International Conference on Case-based Reasoning, Chicago, IL, USA, pp. 339–352 (2005a)
McCarthy, K., McGinty, L., Smyth,B., Reilly J.: On the evaluation of dynamic critiquing: a large-scale user study. In: Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, Pittsburgh, Pennsylvania, USA, pp. 535–540 (2005b)
McCarthy, K., Reilly, J., McGinty, L., Smyth, B.: Experiments in dynamic critiquing. In: Tenth International Conference on Intelligent User Interfaces, San Diego, California, USA, pp. 175–182 (2005c)
McKnight, D.H., Chervany, N.L.: What trust means in E-commerce customer relationships: conceptual typology. Int J Electron Commer **6**(2), 35–59 (2002)

McNee, S.M., Lam, S.K., Konstan, J.A., Riedl, J.: Interfaces for eliciting new user preferences in recommender systems. In: Ninth International Conference on User Modeling, Johnstown, Pennsylvania, USA, pp. 178–188 (2003)
McSherry, D.: Explanation in recommender systems. In: Workshop Proceedings of the Seventh European Conference on Case-Based Reasoning, Madrid, Spain, pp. 125–134 (2004)
Nielsen, J.: Enhancing the explanatory power of usability heuristics. In: SIGCHI Conference on Human factors in Computing Systems, Boston, USA, pp. 152–158 (1994)
Nguyen, Q.N., Ricci, F., Cavada, D.: Critique-based recommendations for mobile users: GUI design and evaluation. In: Third Workshop on "HCI in Mobile Guides" in Conjunction with Sixth International Conference on Human Computer Interaction with Mobile Devices and Services, Glasgow, Scotland (2004)
Novak, T.P., Hoffman, D.L., Yung, Y.-F.: Measuring the customer experience in online environments: a structural modelling approach. Mark Sci **19**(1), 22–42 (2000)
Payne, J.W., Bettman, J.R., Johnson, E.J.: The Adaptive Decision Maker. Cambridge University Press (1993)
Payne, J.W., Bettman, J.R., Schkade, D.A.: Measuring constructed preference: towards a building code. J Risk Uncertainty **19**(1–3), 243–270 (1999)
Pu, P., Chen, L.: Integrating tradeoff support in product search tools for E-commerce sites. In: Sixth ACM Conference on Electronic Commerce, Vancouver, BC, Canada, pp. 269–278 (2005)
Pu, P., Faltings, B.: Enriching buyers' experiences: the SmartClient approach. In: SIGCHI Conference on Human Factors in Computing Systems, Hague, Netherlands, pp. 289–296 (2000)
Pu, P., Faltings, B.: Decision tradeoff using example critiquing and constraint programming. Special Issue User-Interact Constraint Satisfaction, CONSTRAINT: an Int. J. **9**(4), 289–310 (2004)
Pu, P., Kumar, P.: Evaluating example-based search tools. In: Fifth ACM Conference on Electronic Commerce, New York, NY, USA, pp. 208–217 (2004)
Reilly, J., McCarthy, K., McGinty, L., Smyth, B.: Dynamic critiquing. In: Seventh European Conference on Case-based Reasoning, Madrid, Spain, pp. 763–777 (2004)
Reilly, J., McCarthy, K., McGinty, L., Smyth, B.: Explaining compound critiques.. Artif. Intell. Rev. **24**(2), 199–220 (2005)
Shimazu, H.: ExpertClerk: navigating shoppers' buying process with the combination of asking and proposing. In: Seventeenth International Joint Conference on Artificial Intelligence, Seattle, Washington, USA, pp. 1443–1450 (2001)
Smyth, B., McGinty L.: An analysis of feedback strategies in conversational recommenders. In: Fourteenth Irish Artificial Intelligence and Cognitive Science Conference, Dublin, Ireland, pp. 211–216 (2003)
Shneiderman, B.: In: Designing the User Interface: Strategies for Effective Human–Computer Interaction, 3rd edn. Addison-Wesley, Reading, MA (1997)
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)
Torrens, M., Faltings, B., Pu, P.: SmartClients: constraint satisfaction as a paradigm for scaleable intelligent information systems. Int J Constraints **7**(1), 49–69 (2002)
Thompson, C.A., Goker, M.H., Langley, P.: A personalized system for conversational recommendations. J. Artif. Intell. Res. **21**, 393–428 (2004)
Tversky, A., Simonson, I.: Context-dependent preferences. Manage. Sci. **39**(10), 1179–1189 (1993)
Viappiani, P., Faltings, B., Pu, P.: Preference-based search using example-critiquing with suggestions. J. Artif. Intell. Res. **27**, 465–503 (2007)
Williams, M.D., Tou, F.N.: RABBIT: an interface for database access. In: ACM '82 Conference, pp. 83–87 (1982)

Authors' vitae

Li Chen obtained her Bachelor and Master degrees in Computer Science from Peking University, China, and her Ph.D. degree from the School of Computer and Communication Sciences at the Swiss Federal Institute of Technology in Lausanne (EPFL). She participated in a Swiss National Science Foundation project whose goal is to design and develop intelligent and adaptive user interfaces to improve users' decision quality in e-commerce, reduce their efforts, and increase their subjective perception of, e.g., decision confidence and trust. She has authored and co-authored a number of publications in leading journals and conferences in

e-commerce, artificial intelligence, intelligent user interfaces, user modeling, and recommender systems. She won a Best Student Paper award in 2007 at the International Conference on User Modeling.

Pearl Pu obtained her Master and Ph.D. degrees from the University of Pennsylvania in artificial intelligence and computer graphics. She was a visiting scholar at Stanford University in 2001, both in the database and HCI groups. She currently leads the HCI Group in the School of Computer and Communication Sciences at the Swiss Federal Institute of Technology in Lausanne (EPFL). Her research is multidisciplinary and focuses on issues in the intersection of human computer interaction, artificial intelligence, and behavioral decision theories. She works on preference-based search in online environments, decision support systems, electronic commerce, online consumer decision behavior, product recommender systems, content-based product search, travel planning tools, trust-inspiring interfaces for recommender agent, music recommenders, and user technology adoption. Dr. Pu is associate editor for the IEEE Transactions on Multimedia, the general chair of the 2008 ACM Conference on Recommender Systems, and program co-chair of the 2008 International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems.