Hybrid Critiquing-based Recommender Systems

Li Chen and *Pearl Pu*
Human Computer Interaction Group, School of Computer and Communication Sciences
Swiss Federal Institute of Technology in Lausanne (EPFL)
CH-1015, Lausanne, Switzerland
{li.chen, pearl.pu}@epfl.ch

ABSTRACT

We propose a novel critiquing-based recommender interface, the hybrid critiquing interface that integrates the user self-motivated critiquing facility to compensate for the limitations of system-proposed critiques. The results from our user study show that the integration of such self-motivated critiquing support enables users to achieve a higher level of decision accuracy while consuming less cognitive effort. In addition, users expressed higher subjective opinions of the hybrid critiquing interface than the interface simply providing system-proposed critiques, and they would more likely return to it for future use.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces – Evaluation/methodology, Graphical user interfaces (GUI), User-centered design.

General terms: Design, Experimentation, Human Factors

Keywords: Example critiquing, dynamic critiquing, recommender systems, decision support, user study

INTRODUCTION

People are usually unable to accurately state their preferences up front [11,20], especially when confronted with an unfamiliar product domain or a complex decision situation with overwhelming information, such as the current e-commerce environments. As an effective preference construction and feedback mechanism, the critiquing-based recommender system has emerged and been broadly developed to expose domain knowledge and guide users to make accurate and confident decisions.

Specifically speaking, the critiquing-based recommender system uses users' current preferences to recommend specific options, and then elicits users' feedback in the form of critiques such as "I would like something cheaper" or "with bigger optical zoom". These critiques help the recommender refine users' preference model so as to improve the recommendation accuracy in the next cycle. It has been shown that the critiquing support allows users to obtain higher decision accuracy compared to non critiquing-based systems such as a ranked list [12,15].

To our knowledge, the critiquing idea was first mentioned in RABBIT systems [21] as a new interface paradigm for formulating queries in a database. In recent years, the system-proposed critiquing system has been developed aiming to pro-actively generate a set of knowledge-based critiques that users might accept as ways to improve the current recommendation. This approach has been adopted in FindMe systems [2] and the more recently proposed dynamic and incremental critiquing systems [9,16].

An alternative critiquing mechanism, the user self-motivated critiquing approach, does not propose pre-computed critiques, but provides a facility by which users can freely identify a single or a set of features to improve or compromise by themselves. The example critiquing agent is a purely user self-motivated critiquing system, since it focuses on showing examples and stimulating users to make such self-motivated critiques [15].

Recent research has compared these two types of critiquing-based recommender systems [4]. The results show that the example critiquing system achieved better results in terms of users' decision accuracy, cognitive effort and decision confidence. However, some users (36.1%) still preferred the system-proposed critiquing system, since they found it intuitive to use, straightforward for making critiques, and more importantly, the system-proposed critiques motivated them to think about tradeoff decisions. Further analysis of user data showed that the majority of these 36.1% users were able to accelerate their decision process due to the fact that the system accurately predicted the critiques that users were prepared to make.

We therefore decided to develop a hybrid critiquing system by combining the strengths from both critiquing approaches: system-proposed and user self-motivated. We believe that with the hybrid critiquing system, people can not only obtain knowledge of the domain and easily perform critiquing via the proposed critiques, but also have the opportunity to freely compose and combine critiques by themselves if necessary with the aid of user self-motivated critiquing support. Thus, users' decision performance and subjective perceptions can be potentially further improved to reach a high level.

IUI'07, January 28–31, 2007, Honolulu, Hawaii, USA.
Copyright 2007 ACM 1-59593-481-2/07/0001...$5.00.

Contribution of Our Work

The interface design of an intelligent system must be capable of delivering the intended user benefits. Therefore, improving the ability of a product recommender system to motivate users to make more tradeoff decisions and improve their decision accuracy is highly relevant to the field of intelligent user interfaces.

The focus of our work aims to understand whether the hybrid critiquing-based recommender system can improve users' decision performance and more importantly how our user self-motivated example critiquing facility acts in such systems. We have conducted a user study to evaluate the hybrid critiquing interface by comparing it with the system-proposed critiquing system, so as to determine whether the former has improved results due to the addition of the self-motivated critiquing support. We chose the dynamic critiquing interface with its incremental critiquing features as the representative of system-proposed critiquing systems because its actual advantages have been established through a series of simulations and real-user studies [9,10,16].

More specifically, we have established an evaluation framework to evaluate the two interfaces. It involves both users' objective performance such as decision accuracy, task completion time and interaction effort, and their subjective perceptions including perceived cognitive effort, decision confidence and trusting intentions. All of these factors are important for evaluating a recommender system, since the system's optimal goal should be to allow its users achieving high decision accuracy and building high trust in it, while requiring a minimal amount of effort for making decisions [3,7,12,13].

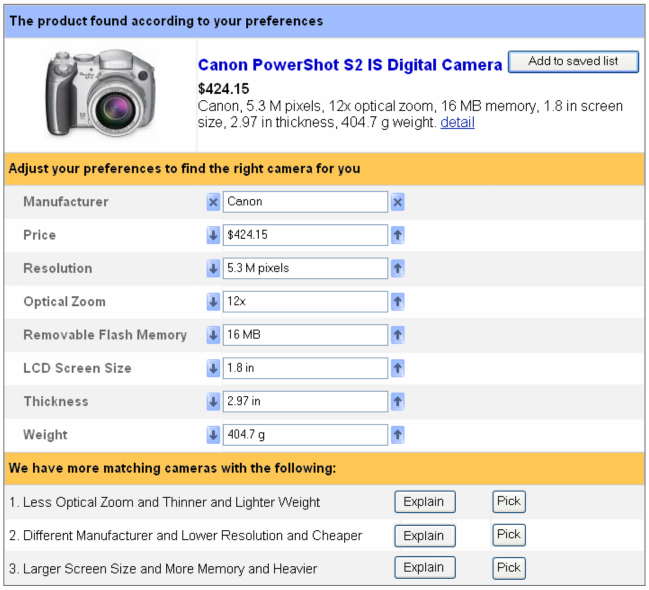
In addition, we have performed an in-depth measure on how people respectively react to the self-motivated critiquing support and system-proposed critiques in the hybrid critiquing interface (i.e. their application frequency), and how their final decision accuracy is accordingly affected by the actual application.

The rest of this paper is therefore organized as follows. We first introduce the system-proposed critiquing, focusing on the dynamic critiquing interface. We then describe our user self-motivated example critiquing support and introduce the hybrid critiquing interface that combines both critiquing approaches' advantages. We present the user evaluation in detail in terms of the evaluation framework, hypotheses, materials, participants, and experiment procedure, followed by the results analysis and discussion. Finally, we present the conclusion of our work.

SYSTEM-PROPOSED CRITIQUING

As mentioned above, the system-proposed critiquing approach generates critiques according to its knowledge about users and the product domain (also called assisted browsing in FindMe systems). The critiques were selected by users to look for products with improved values on certain attributes. For example, the tweak application developed in FindMe systems [2] allows users to critique the current recommendation by selecting one of the proposed simple tweaks (e.g. "cheaper", "bigger" and "nicer" that are along with the current suggested apartment). When a user finds the current recommendation short of their expectations and responds to a tweak, the remaining candidates will be filtered to leave only those candidates satisfying the tweak.

In addition to the use of so-called unit critiques that constrain a single feature at a time, the dynamic critiquing method [10,16] and its successor, incremental critiquing [9], automatically and dynamically generate compound critiques that can operate over multiple features (e.g. "Different Manufacture, Lower Resolution and Cheaper"). It was demonstrated that the total number of recommendation cycles can decrease from 29 to 6 when users actively selected the compound critiques [10].



The product found according to your preferences

Canon PowerShot S2 IS Digital Camera

$424.15

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

Adjust your preferences to find the right camera for you

Manufacturer: Canon

Price: $424.15

Resolution: 5.3 M pixels

Optical Zoom: 12x

Removable Flash Memory: 16 MB

LCD Screen Size: 1.8 in

Thickness: 2.97 in

Weight: 404.7 g

We have more matching cameras with the following:

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

Figure 1: The dynamic critiquing interface. The system proposes unit and compound critiques for users to select.

Dynamic and Incremental Critiquing

The dynamic critiquing interface (see Figure 1) presents both the unit and compound critiques to users as feedback options, so as to facilitate the critiquing of single feature or multiple features. The compound critiques are computed by discovering the recurring sets of unit differences between the current recommended item and the remaining products using the data mining Apriori algorithm [1].

For example, suppose the occurrence of digital cameras with cheaper price (compared with the current recommended product) is highly probably associated with the occurrence of products with lower resolution, then the Apriori algorithm will produce an association rule (i.e. a compound critique "cheaper but lower resolution"). Essentially, dynamic critiquing employs this algorithm to discover the highest recurring variations that are typical of the given data set and turn them into compound critiques. Further, it filters the potentially large number of compound critiques by using a threshold value, favoring those critiques with low support values ("support value" refers to the percentage of products that satisfy the critique). Such selection criterion was motivated by the fact that taking such low-support critiques is likely to accelerate user's navigation to the target quickly. In the dynamic critiquing system with incremental critiquing features, the products satisfying the current critique must be additionally compatible with what a user has previously critiqued. Live user evaluation showed that the incremental critiquing saves users' interaction cycles by up to 34% compared to the standard dynamic critiquing method (see details in [9]).

A typical interaction process with the incremental dynamic critiquing interface (henceforth "dynamic critiquing" for short) is therefore as follows. First the system provides a recommendation to the user, while simultaneously generating hundreds of compound critiques from the data set via the Apriori algorithm, and showing users the critiques with lower support values. The user views these critiques and picks one, and the system subsequently recommends a new product that is the most similar to the last recommendation and also the most compatible with the user's previous critiques. The list of proposed critiques is accordingly updated. This process continues until the user decides that she has found her most preferred product.

The dynamically generated critiques have been also regarded as explanations to reveal the recommendation opportunities that exist in the remaining products [17]. Especially for the user who has incomplete knowledge about the product's features and their relationships, the compound critiques may help the user better understand how the features are highly probably related within the remaining alternatives (e.g. cheaper price of digital camera is usually associated with lower resolution). This could potentially prevent the user from making further retrieval failure [17].

Limitation of System-Proposed Critiques

The main limitation of this kind of system-proposed critiquing interface is that it only allows users to make critiques by picking its proposed ones (e.g. selecting compound critiques or performing unit critiques in dynamic critiquing). Users have no chance to combine critiques on their own. Consider a user is looking for a digital camera with higher resolution and more optical zoom relative to the current recommended product. Suppose there is no suggested critique matching the intended critiquing, even though the proposed critiques can give her some knowledge of the remaining digital cameras (e.g. "larger screen size and more memory" but is not her intended criteria). At this point, she can only choose to make unit critiques on the dynamic critiquing interface by changing her preference on one feature at a time. This process, however, brings her the risk of being involved in longer interaction cycles.

In addition, as discussed in [4], the dynamic critiquing interface only allows quality-based critiquing (e.g. "cheaper," "bigger," or "Different Manufacture, Lower Resolution and Cheaper"). It does not support similarity-based (e.g. "find similar digital camera like this one, no specific critique on any feature") and quantity-based critiquing with concrete value preference (e.g. "find similar digital camera like this one, but at least $100 cheaper"). The number of recommendations during each cycle is also limited to one, so users cannot make products comparison between the current recommendation and previous ones and also cannot examine and compare more tradeoff alternatives that might satisfy her critique.

In the next sections, we introduce the user self-motivated example critiquing agent and see how it can be combined with the system-proposed critiques to overcome these limitations.

USER SELF-MOTIVATED CRITIQUING

Instead of suggesting pre-computed critiques for users to choose, the user self-motivated critiquing approach focuses on showing examples and stimulating users to make self-motivated critiques. It does not limit the critiques a user can manipulate during each cycle, so users can post unit or compound critiques over any combination of features with freedom. In fact, the focus of this interface is to assist users in executing tradeoff navigation, which is a process shown to improve users' decision accuracy and confidence [12].

More precisely, the tradeoff navigation involves finding products having more optimal values on one or several attributes that are important for the user, while accepting compromised values for other less important attributes. With the self-motivated critiquing interface, the user can conveniently start the tradeoff navigation from one item (called the reference product), specify her tradeoff criteria in terms of improvement and compromise regarding the product's features, and see a new set of products more nearly approaching to her ideal choice. The unit and compound critiques are respectively termed simple and complex tradeoffs in such systems [15].

Example Critiquing

The example critiquing agent is a purely user self-motivated critiquing system. It was initially implemented in ATP [19]. Later on, ATP became SmartClient, an online preference-based search tool for finding flights [18]. The method was subsequently applied to catalogs of vacation packages, insurance policies, apartments, and recently commercial products such as tablet PCs and digital cameras [4,14,15].

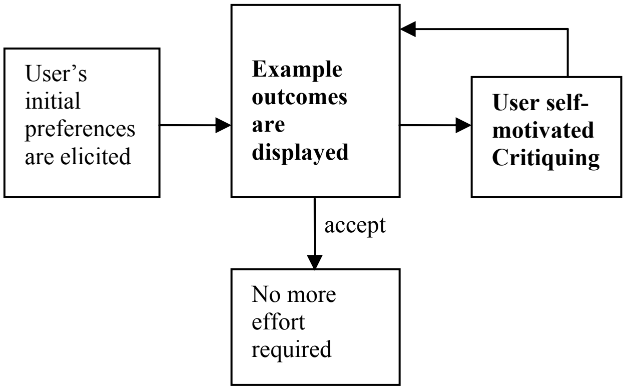
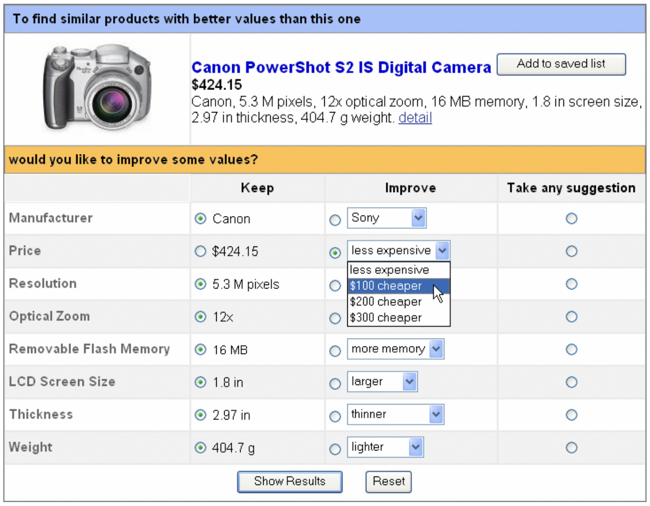

User's initial preferences are elicited
Example outcomes are displayed
User self-motivated Critiquing
accept
No more effort required

Figure 2: The example critiquing interaction model.

The example critiquing interaction model is shown in Figure 2. The general interaction process is similar to the interaction with dynamic critiquing interface, except that here the emphasis is on the computation of examples according to the user's current preferences and the stimulation of user self-motivated critiquing.

More concretely, the example critiquing implementation mainly consists of a user interface and a search engine. A user initially starts the search by specifying one or any number of preferences in the query area. Each preference is composed of one acceptable attribute value and the corresponding degree of importance (i.e. weight) of that attribute. The weight ranges over five values, from "least important" to "very important". A preference structure is hence a set of (attribute value, weight) pairs of all participating attributes.



To find similar products with better values than this one

Canon PowerShot S2 IS Digital Camera
$424.15
Canon, 5.3 M pixels, 12x optical zoom, 16 MB memory, 1.8 in screen size, 2.97 in thickness, 404.7 g weight. detail

would you like to improve some values?

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

Figure 3: The example critiquing interface. Users can freely build and combine critiques to refine their search.

Based on the initial preference model, the search engine will find and display a set of matching results (see [5] for the optimal number of solutions to display). The user either accepts a result, or takes a near solution (i.e. the reference product) and activates the tradeoff navigation panel (or called the critiquing panel, see Figure 3), where she can post critiques to the near-target solution based on her desire to trade off more of one or multiple valued attribute(s) for less of other attribute(s). Once a set of critiques has been composed, the system will refine the user's preference model and adjust the relative importance of all critiqued attributes accordingly (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 compute and return a new set of tradeoff alternatives based on the refined preference model. This query/critiquing completes one cycle of interaction, and it continues as long as the user wants to further refine the results.

In Figure 3 (i.e. the critiquing panel), three radio buttons are next to each feature, respectively under "Keep" (default), "Improve" and "Take any suggestion", thus facilitating users to critique one feature by either improving its current value (i.e. selecting "Improve") or accepting a compromised value suggested by the system (i.e. via "Take any suggestion"). More notably, users can freely compose compound critiques by combining critiques on any set of multiple features.

The interface also supports different types of critiquing. For example, users can perform the similarity-based critiquing by keeping all current values (the default option "Keep") and clicking on the "Show Results" to view the products that are the most similar to the reference product. As for the quality-based and quantity-based critiquing, users can select the respective option in the drop down menu under the "Improve" column, e.g. "less expensive" or "$100 cheaper" (see Figure 3).

The search engine to find tradeoff alternatives is adjusted for different decision environments. For configurable products, it employs sophisticated constraint satisfaction algorithms and models user preferences as soft constraints [18]. For multi-attribute products, it basically applies the weighted additive sum rule (WADD), a compensatory decision strategy to produce accurate outcome (see [8] for further details). We also implemented a combined strategy in the current system for electronic products (e.g. digital camera), that combines the elimination-by-aspect strategy (EBA) with WADD [11]. The EBA is used to retrieve alternatives that maximally match the user's improving criteria (i.e. critiques for better values such as "higher processor speed" and "more memory"), and WADD is used to rank the retrieved alternatives by their weighted utility scores relative to the user's current preferences refined by her posted critiques.

Dynamic Critiquing vs. Example Critiquing

Table 1 summarizes the main differences between the dynamic critiquing and example critiquing interfaces. In our previous user study [4], we compared these two critiquing systems and found that most users obtained more accurate decisions and higher subjective perceptions (e.g. confidence in choice) with the example critiquing interface, mainly due to their complete control of critiques building. However, approximately one third of the participants indicated that they preferred the dynamic critiquing interface to search products, since they perceived it to be less demanding by picking proposed critiques and to potentially be able to accelerate their decision making.

This study motivated us to develop a new type of critiquing interface to combine both approaches' advantages so as to maximally improve users' decision performance in terms of both accuracy and effort and their subjective preference.

HYBRID CRITIQUING INTERFACE

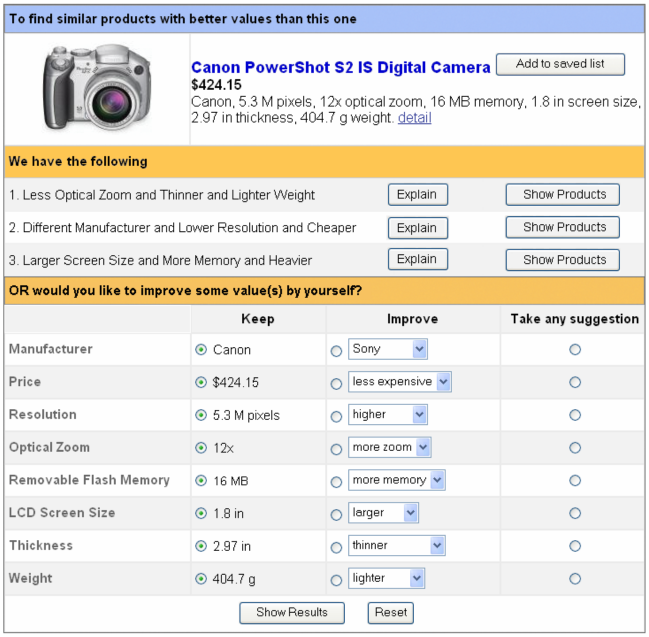
As mentioned above, the main advantage of system-proposed critiques is that it can expose to users the recommendation opportunities that exist in the remaining candidates so as to avoid retrieval failure, and potentially assist users in making a quick choice if the critiques correspond well to users' intended tradeoffs. However, it is also revealed that users are limited in making critiques and viewing tradeoff alternatives in such an interface, which would likely result in longer interaction cycles and even lower level of decision accuracy compared to the user self-motivated critiquing interface where critiques can freely be created by users themselves [4].

	Dynamic critiquing	**Example critiquing**
Critiquing generation	Users can select their own critiques, but only on unit critiques. Otherwise, the system proposes compound critiques for users to choose.	Users are able to freely create and combine critiques to simultaneously improve a number of product criteria
Critiquing modality	Only support quality-based critiquing, e.g. "cheaper", "different manufacturer, lower resolution and cheaper"	Support three types of critiquing: • Similarity-based, e.g. "similar to this one"; • Quality-based, e.g. "similar, but cheaper"; • Quantity-based, e.g. "$100 cheaper"
Critiquing unit	Unit and compound critiques	Simple and complex tradeoffs
Critiquing coverage	Critiques are made on only one current recommendation	Critiques are made on one reference product selected from multiple suggested examples
Search algorithm	Similarity and compatibility measure	Elimination by aspect (EBA) plus weighted additive sum (WADD)

Table 1: Main differences between the dynamic critiquing and example critiquing interfaces.

To keep the system-proposed critiques' advantages while compensating for their limitations, we propose a hybrid critiquing system that integrates the user self-motivated example critiquing support with the system-proposed critiquing interface. Thus, for example, when a user is looking for some products with higher resolution and more optical zoom relative to the current recommended digital camera, if one of the proposed critiques exactly matches such conditions, she can undoubtedly select it; otherwise, she can choose to specify these criteria in the self-motivated example critiquing panel (i.e. improving the resolution and optical zoom simultaneously and optionally selecting concrete value improvements). She can also choose to compromise some of other attributes that are less important for her to guarantee the intended gains.

More specifically, figure 4 shows one design of the hybrid critiquing interface that combines the system-proposed critiques based on incremental dynamic critiquing method and the user self-motivated example critiquing support. The proposed critiques are listed under the current recommendation and the bottom of the interface is the self-motivated critiquing area with functions to facilitate different types of critiquing (e.g. similarity-based, quality-based, or quantity-based) and critiquing units (e.g. unit or compound critiques). Note here that it does not display the unit critiquing part from the dynamic critiquing interface (see Figure 1) since that function is provided by the example critiquing support.



To find similar products with better values than this one

Canon PowerShot S2 IS Digital Camera
$424.15
Canon, 5.3 M pixels, 12x optical zoom, 16 MB memory, 1.8 in screen size, 2.97 in thickness, 404.7 g weight. detail

We have the following

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

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

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

Figure 4: One design of a hybrid critiquing interface with system-proposed critiques and user self-motivated critiquing facility.

After each critiquing process, a set of tradeoff alternatives that best match users' critiques will be returned by the hybrid critiquing system for users to compare. The search algorithm is accordingly chosen to adapt to the type of critiques users posted, for example, it applies similarity and compatibility selection measures if the dynamic proposed critique is picked, and employs EBA plus WADD ranking mechanism if the user specifies self-motivated critiques. Among the recommended items, users can choose one as their final choice and finish the session, or select one as the reference product (i.e. near-target) to start the next round of critiquing.

USER EVALUATION

In order to understand whether the hybrid critiquing interface can achieve a high level of decision accuracy and high subjective opinions from the users, as well as how the user self-motivated example critiquing support affects users' performance in the hybrid system, we have conducted an empirical user study to evaluate the hybrid critiquing interface (henceforth DC+EC) by comparing it with the dynamic critiquing interface (henceforth DC).

Evaluation Framework and Hypotheses

We first established an evaluation framework on which the comparison of the two interfaces was based. Indeed, identifying the appropriate criteria for evaluating the real benefits of a recommender system is a challenging issue. Related works have mostly focused on the evaluation of users' objective interaction effort with the system, such as their interaction cycles [10] and task completion time, and less regarded what actual decision accuracy users can eventually achieve and how much cognitive effort users perceive to be exerted. In addition, to appraise whether the system can convince its users to purchase a product, which is especially important in the e-commerce environments, and stimulate them to return to the system for future use, it is quite meaningful to measure users' subjective opinions on the interface about both their intention to purchase and intention to return. The two intentions are essentially identified as trusting intentions in online environments [6].

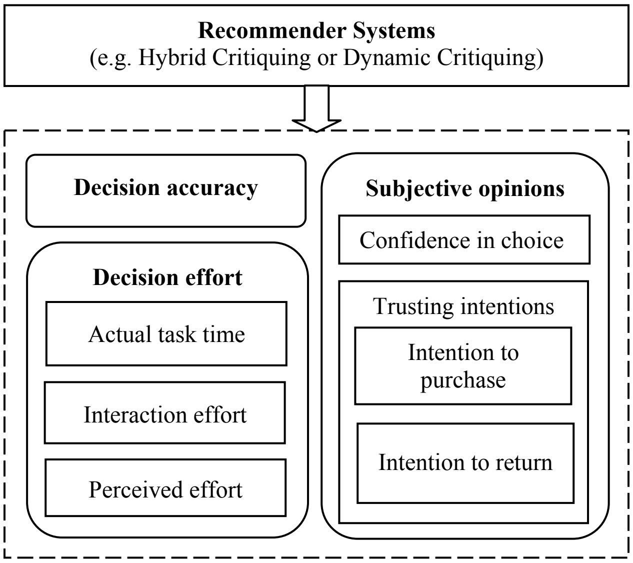


Figure 5: Our evaluation framework for recommender systems.

These requirements have led us to develop an evaluation framework containing all of the important factors. More concretely, it is primarily made up of three components: decision accuracy, decision effort and users' subjective opinions (see Figure 5).

The *decision accuracy* is quantitatively measured by the fraction of participants that switched to a different, better option than the one chosen using the system when they were asked to view all alternatives in the database. A lower switching fraction means that the system allows higher decision accuracy since most users found their target choice using it. This method was also applied by researchers in marketing science to measure decision quality [7].

Decision effort is mainly measured by two aspects: one is users' objective effort consumed including their task completion time and interaction effort; another is users' perceived cognitive effort to indicate the amount of subjective effort they exerted.

Subjective opinions include users' confidence in their choice made with the recommender system and their trusting intentions in terms of intention to purchase the chosen product and intention to return to the system for future use.

Therefore, based on the framework, the user evaluation of the two critiquing-based recommender systems can tell us whether the hybrid critiquing interface could enable users to achieve higher decision accuracy while requiring less decision effort, and stimulate users to possess a higher level of subjective opinions. More specifically, we were interested in clarifying the following hypotheses:

Hypothesis 1: relative to the system-proposed critiquing interface, the hybrid critiquing interface with user self-motivated example critiquing support can improve users' decision accuracy;

Hypothesis 2: the hybrid critiquing interface can reduce users' decision effort;

Hypothesis 3: the hybrid critiquing interface can improve users' subjective opinions in terms of their confidence in choice and trusting intentions.

Materials and Participants

Both the hybrid critiquing and dynamic critiquing interfaces were developed for two product catalogs: tablet PCs and digital cameras. The tablet PC catalog is comprised of 55 products, each described by 10 main features (manufacturer, price, processor speed, weight, etc.), and the digital camera catalog comprises 64 products characterized by 8 main features (manufacturer, price, resolution, optical zoom, etc.). All products were extracted from a real e-commerce website.

The entries to the two interfaces are identical, comprised of a preference specification page to obtain users' initial preferences. Then, in the dynamic critiquing interface, the item that best matches users' initial preferences is shown in the beginning, accompanied by a set of unit critiques and three system-proposed compound critiques on the same screen (see Figure 1). Once a critique is selected, a new item will be recommended with updated proposed critiques.

In the hybrid critiquing interface (see Figure 4), this page is modified to combine the example critiquing panel. The item currently critiqued (i.e. the reference product) is displayed with three system-proposed compound critiques and a self-motivated critiquing area. Users can pick the proposed compound critiques or produce critiques on their own. Once critiques are posted, a set of matching items will be returned for users to compare with the reference product. If a user finds her target choice among these items, she can proceed to check out. Otherwise, if she likes one product but wants something improved, she can come back to the critiquing page (by clicking the "Value Comparison" button along with the product) to resume a new critiquing cycle.

In both interfaces, users can view the product's detailed specification with the "detail" link. Users can also save all near-target solutions in their consideration set (i.e. saved list) to facilitate comparing them before checking out.

A total of 36 (6 females) volunteers participated in the user evaluation for a reward valued at approximately 10 CHF per user. Most of them are students in the university (age between 20 and 30), but they are from a variety of different countries (France, India, Switzerland, China, etc.), studying varied subjects (computer science, mechanics, manufacturing, etc.) and pursuing different levels of educational degrees (bachelor, master, or Ph.D.). Among the participants, 29 have online shopping experience.

Experiment Design and Procedure

The user study was conducted in a between groups design. All participants were randomly and evenly divided into two groups, and each group was assigned one interface (either DC or DC+EC) to evaluate. In addition, every participant was randomly assigned one product domain (tablet PC or digital camera) to search.

The user study was conducted at locations convenient for the participants (office, home, cafeteria, etc.) with the help of a provided notebook or desktop computer. An online procedure containing the instructions, evaluated interfaces and questionnaires was implemented so that the users could easily follow, and we could also record all of their actions in a log file. The same administrator supervised the experiment for all of the participants.

At the beginning of each session, the participant was first debriefed on the objective of the experiment and the upcoming tasks. In particular, she was asked to evaluate a product search interface to determine whether it is effective in helping her to make a confident and accurate purchase decision. Thereafter, a short questionnaire was to be filled out about her demographics, e-commerce experience and product knowledge.

The participant would then start evaluating the interface by imagining herself as a potential buyer. The main user task was to "find a product you would purchase if given the opportunity" with the assigned critiquing interface. After the choice was made, the participant was asked to fill in a post-study questionnaire about her perceived cognitive effort, decision confidence, and trusting intentions regarding the interface. Then the interface's decision accuracy was measured by revealing all products to the participant to determine whether she prefers another product in the catalog or stands by the choice made using the critiquing interface.

RESULTS ANALYSIS

Critiquing Application

We first measured how users reacted to the user self-motivated example critiquing support in the hybrid critiquing interface, and whether their application of system-proposed critiques would change due to the appearance of EC (relative to in the dynamic critiquing interface). The results show that among the participants who used DC+EC, 88.9% applied EC during average 76.1% of their critiquing cycles. In addition, around 42% of their critiquing with EC was compound critiques that involve changes on maximal 7 features at a time. The results infer that when the self-motivated critiquing support is enabled in the interface, users will quite frequently apply it to compose critiques (including compound in addition to unit critiques) by themselves.

As for the system-proposed critiques, the percentage of users who applied them decreased from 83.3% (on DC) to 44.4% (on DC+EC), and the average application frequency per user also dropped from 3.2 times to 1.1 ($p = 0.05$). However, the application of system-proposed critiques in DC+EC was found to be significantly correlated with a higher frequency of users' application of EC ($p = 0.06$). That is, the system-proposed critiques can motivate users to create critiques on their own more often. Another phenomenon is that 83.3% of users using DC+EC ended their session by making self-motivated EC, which implies that the system-proposed critiques were mostly employed before users considered making their final choice.

Thus, the above data indicates that due to the appearance of the EC interface, users less frequently picked the proposed critiques, but chose to self build critiques with EC more actively. It also infers that the system-proposed critiques are likely applicable in the earlier cycles when users are less certain about their preferences and have a superficial understanding of the product domain. Later on, once users obtain a certain degree of product knowledge and what they want, they will be more likely to perform self-motivated critiques that will ultimately lead to their final choice.

For the next step of analysis, we will see whether such a frequent application of EC can result in a positive influence on users' decision performance and their subjective opinions on the interface.

Decision Accuracy and Decision Effort

The decision accuracy of the hybrid critiquing interface, as defined above, was 66.7%, since 12 out of 18 participants found their target choice using the interface. Comparatively, DC had a lower decision accuracy of 33.3% (6 out of 18) as the remaining users switched to a different, better choice when they were given the opportunity to view all of the products in the catalog. The difference was proven to be significant by t-test (t = -2.06, $p < 0.05$).

The higher decision accuracy obtained by the hybrid critiquing interface was further examined in respect to the relevant users' critiquing application (see Figure 6). The result shows that 50% of the decision accuracy was contributed from users who applied both system-proposed DC and self-motivated EC, 41.67% from only applying EC, and 8.33% from users whose choice was the first recommendation according to their initial preferences (i.e. without critiquing process). This phenomenon exhibits a significant distribution ($p = 0.03$). Thus, in total 91.67% of the decision accuracy comes from the 88.9% participants who have applied the self-motivated EC while using the hybrid critiquing interface.

Furthermore, the users who applied EC more frequently (more than once) during their interaction with DC+EC achieved 77.7% decision accuracy, versus 55.6% from the users applying EC less than or equal to once. Combined with the result that the system-proposed critiques can motivate users to be more frequently self-motivated to produce critiques, it can be implied that this process could potentially guide users to make a more accurate decision.

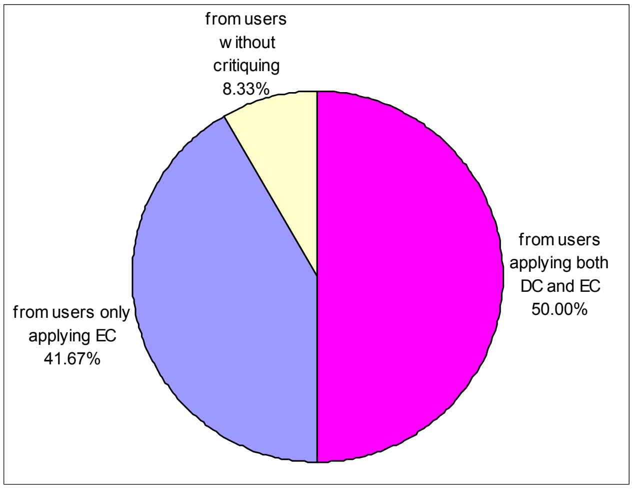


Figure 6: The distribution of users who found their target choice using the hybrid critiquing interface.

We also tracked how much effort users actually expended to achieve the corresponding accuracy. As shown in the evaluation framework, decision effort was measured both from users' objective performance and their subjective perception. Regarding the objective task completion time, the participants who used DC spent average 5 minutes in locating their choice, while the other group using DC+EC consumed slightly more time (5.5 minutes) on average. However, this difference is not significant (t = -0.48, *p* = 0.63).

Moreover, users' actual interaction cycles (i.e. critiquing cycles) indicate a highly significant reduction (of up to 64%) due to the integration of EC in the hybrid critiquing interface (4.61 with DC+EC vs. 11.06 cycles with DC, t = 2.61, *p* = 0.01). Figure 7 shows the association of such a large reduction in interaction cycles with approximately 200% improvement on users' decision accuracy. Therefore, it indicates that the hybrid critiquing interface can allow users to reach higher decision accuracy while requiring them to be involved in fewer interaction cycles.

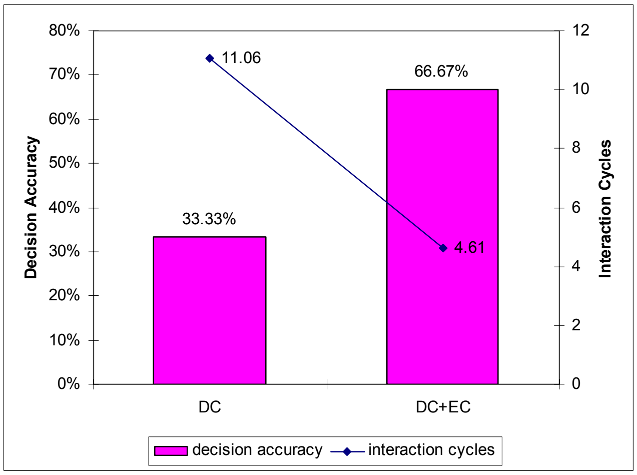


Figure 7: Decision accuracy improvement and reduction in interaction cycles for the two interfaces.

In addition to the above-mentioned objective measurement, we asked users to respond to two interrelated post-study questions (respectively on a 5-point Likert scale ranging from 1 "strongly disagree" to 5 "strongly agree") to determine their perceived cognitive effort (see Table 2 for concrete questions and statistics of user responses). The lower mean rate represents a smaller amount of subjective effort an average user perceived during her interaction with the corresponding interface. As a result, the overall cognitive effort was perceived as significantly lower (t = 2.23, *p* = 0.03) on the hybrid critiquing interface (see Figure 8; mean = 2.06 vs. mean = 2.67 on DC).

Therefore, in addition to the actual reduction in interaction cycles, users also perceived the hybrid critiquing interface as requiring less effort in obtaining and processing information to arrive at their decision, although the actual time spent was slightly (but not significantly) more given users would probably consume more time in making self-motivated critiques.

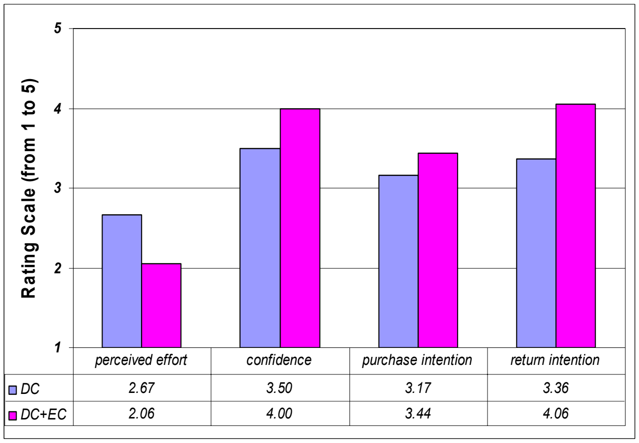


Figure 8: Users' mean responses to the post-questions about their perceived effort, decision confidence and trusting intentions.

Subjective Opinions

Like the perceived effort, users' subjective opinion on the assigned interface was also measured through the post-study questionnaire. The three main aspects contained in the evaluation framework (see Figure 5) were respectively measured by asking users whether they were confident they made the best choice with the interface (i.e. *confidence in choice*), whether they intended to purchase the chosen product once given the opportunity (i.e. *intention to purchase*) and whether they would return to the interface for future use (i.e. *intention to return*). Table 2 lists the concrete questions. Each question was also required to respond on a 5-point Liker scale ranging from 1 "strong disagree" to 5 "strongly agree".

Analysis of users' answers shows that both groups of participants indicated a high degree of agreement with these statements for both interfaces, but the hybrid critiquing interface gained relatively higher scores regarding all of the three aspects. More concretely, participants possessed significantly more confidence in

Measured variables	**Questions related to the variables (each responded on a 5-point Likert scale)**	**Mean (St.d.)**		**Median**	
		DC	DC+EC	DC	DC+EC
Perceived cognitive effort	I easily found the information I was looking for.	2.67 (0.97)	1.94 (0.73)	2	2
	Looking for a product using this interface required too much effort (*reverse scale*).	2.67 (1.19)	2.17 (0.92)	2.5	2
Confidence in choice	I am confident that the product I just "purchased" is really the best choice for me.	3.5 (0.62)	4 (0.49)	4	4
Intention to purchase	I would purchase the product I just chose if given the opportunity.	3.17 (0.86)	3.44 (0.86)	3	4
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.	3.39 (0.98)	3.83 (1.15)	3.5	4
	I don't like this interface, so I would not use it again (*reverse scale*).	3.33 (1.03)	4.28 (0.89)	3.5	4.5

Table 2: Concrete questions to measure users' subjective perceptions and descriptive statistics of their answers.

their choice made with DC+EC (4 against 3.5 with DC, $p = 0.01$; see Figure 8), implying that they truly perceived the hybrid critiquing interface to provide a higher level of decision accuracy.

Additionally, the group using DC+EC expressed a higher level of intention to purchase the product they chose (3.44 against 3.17 using DC, $p = 0.34$), and significantly a higher level of intention to return to the hybrid critiquing interface for future use (4.06 versus 3.36 to DC, $p = 0.03$; see Figure 8). These results imply that the hybrid critiquing interface can potentially convince the user to buy a product more effectively, and establish a stronger long-term relationship with the user since she will be more likely to use it again.

Discussion

So far, most of our hypotheses are well supported by the user evaluation. Participants using the hybrid critiquing interface that combines the user self-motivated example critiquing support with system-proposed critiques achieved much higher decision accuracy than the participants using the standard system-proposed critiquing interface. In addition, the former group of users went through fewer interaction cycles and spent less subjective effort on average to reach a higher level of decision accuracy, although slightly more actual task time was consumed. In addition, they indicated a higher level of subjective opinions on the hybrid critiquing interface, in terms of their confidence in choice, intention to purchase and intention to return. Thus, the integration of the example critiquing aid has been proven to perform quite effectively in enabling the system to improve its users' decision performance and subjective perceptions.

This finding strongly implies our overall suggestion to improve the current critiquing-based recommender systems. That is, in addition to proposing a set of pre-computed critiques for users to choose [2,9,10,22], it is beneficial to provide a facility that allows users to build and combine critiques on their own. This is based on the fact that users actively reacted to the self-motivated critiquing support when it was present, and frequently applied it especially in the last cycles that led to their final choice. As for the system-proposed critiques, although they were less frequently employed in the hybrid critiquing interface, their application can motivate users to more frequently compose critiques by themselves.

CONCLUSION

We proposed a novel critiquing-based recommender interface, the hybrid critiquing interface that integrates the user self-motivated critiquing facility in order to compensate for the limitations of system-proposed critiques. To measure the hybrid critiquing interface's effectiveness, especially the role of the integrated example critiquing support, we conducted a user study to compare it with the system-proposed dynamic critiquing interface. The two interfaces were concretely evaluated based on the evaluation framework we have established involving the major standards for recommender systems. The user study shows that the hybrid critiquing interface was able to significantly motivate users to consider applying critiques and more than doubled the average decision accuracy achieved by dynamic critiquing interfaces with less cognitive effort. Moreover, results show that users built a higher level of confidence in their choice with the hybrid critiquing system and increased their trusting intentions to the system.

In conclusion, it infers that an effective critiquing-based recommender system should not only pre-actively generate a set of critiques that users might accept to improve the current recommendation, but also provide a user self-motivated critiquing facility, such as the example critiquing interface, which can allow users to freely define tradeoff criteria by themselves. Extending the conclusion to a more general scope, both our previous and current research suggest that preference-based recommender systems must always respect users' initiatives and give them the maximum control for constructing preferences and critiques [4,20].

In the future, we will recruit more ordinary subjects with a higher degree of diversity in their age groups, their professions, nationalities and educational backgrounds to further evaluate the hybrid critiquing-based recommender system. Based on current and future data collected from real users, we will also investigate whether the results (e.g. users' decision performance and subjective perceptions) vary among people with vs. without online shopping experience, and moreover among users who are novices and experts in a specific product domain (e.g. digital camera).

ACKNOWLEDGMENTS

We thank the Swiss National Science Foundation for sponsoring the reported research work. We are grateful to the participants of our user studies for their patience and time. We also thank Jennifer Graetzel and Nicolas Jones for their suggestions in improving the quality of the paper.

REFERENCES

1. Agrawal, R., Imielinski, T. and Swami, A. Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD 1993*, 207-216.
2. Burke, R., Hammond, K. and Young, B. The FindMe approach to assisted browsing. *IEEE Expert: Intelligent Systems and Their Applications* 12, 4 (1997), 32-40.
3. Chen, L. and Pu, P. Trust building in recommender agents. In *Workshop on ICETE 2005*, 135-145.
4. Chen, L. and Pu, P. Evaluating critiquing-based recommender agents. In *Proc. AAAI 2006*, 157-162.
5. Faltings, B., Torrens, M., and Pu, P. Solution generation with qualitative models of preferences. *International Journal of Computational Intelligence and Applications 20*, 2 (2004), 246-264.
6. Grabner-Kräuter, S. and Kaluscha, E.A. Empirical research in on-line trust: a review and critical assessment. *International Journal of Human-Computer Studies 58*, 6 (2003), 783-812.
7. Haubl, G. and Trifts, V. Consumer decision making in online shopping environments: the effects of interactive decision aids. *Marketing Science 19*, 1 (2000), 4-21.
8. Keeney, R. and Raiffa, H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* Cambridge University Press, 1976.
9. McCarthy, K., McGinty, L., Smyth, B. and Reilly, J. A live-user evaluation of incremental dynamic critiquing. In *Proc. ICCBR 2005*, 339-352.
10. McCarthy, K., Reilly, J., McGinty, L. and Smyth, B. Experiments in dynamic critiquing. In *Proc. IUI 2005*, ACM Press (2005), 175-182,.
11. Payne, J.W., Bettman, J.R. and Johnson, E.J. *The Adaptive Decision Maker*. Cambridge University Press, 1993.
12. Pu, P. and Chen, L. Integrating tradeoff support in product search tools for e-commerce sites. In *Proc. ACM EC 2005*, ACM Press (2005), 269-278.
13. Pu. P. and Chen, L. Trust building with explanation interfaces. In *Proc. IUI 2006*, 93-100.
14. Pu, P. and Faltings, B. Decision tradeoff using example critiquing and constraint programming. *Special Issue on User-Interaction in Constraint Satisfaction, CONSTRAINT: an International Journal* 9, 4 (2004).
15. Pu, P. and Kumar, P. Evaluating example-based search tools. In *Proc. ACM EC 2004*, ACM Press (2004), 208-217.
16. Reilly, J., McCarthy, K., McGinty, L. and Smyth, B. Dynamic critiquing. In *Proc. ECCBR 2004,* 763-777.
17. Reilly, J., McCarthy, K., McGinty, L. and Smyth, B. Explaining compound critiquing. In *Workshop on UKCBR 2004*, 12-20.
18. Torrens, M., Faltings, B. and Pu, P. SmartClients: constraint satisfaction as a paradigm for scaleable intelligent information systems. *International Journal of Constraints 7*, 1 (2002), 49-69.
19. Torrens, M., Weigel, R. and Faltings, B. Java constraint library: bringing constraints technology on the Internet using the Java language. In W*orkshop on AAAI 1997*, 10-15.
20. Viappiani, P., Faltings, B. and Pu, P. Preference-based search using example-critiquing with suggestions. to appear in *Journal of Artificial Intelligence Research*.
21. Williams, M.D. and Tou, F.N. RABBIT: an interface for database access. In *Proc. ACM 1982 Conference*, ACM Press (1982), 83-87.
22. Zhang, J. and Pu, P. A comparative study of compound critique generation in conversational recommender systems. In *Proc. AH 2006*, 234-243.