Evaluating Critiquing-based Recommender Agents

Li Chen and Pearl Pu

Human Computer Interaction Group, School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
{li.chen, pearl.pu}@epfl.ch

Abstract

We describe a user study evaluating two critiquing-based recommender agents based on three criteria: decision accuracy, decision effort, and user confidence. Results show that user-motivated critiques were more frequently applied and the example critiquing system employing only this type of critiques achieved the best results. In particular, the example critiquing agent significantly improves users' decision accuracy with less cognitive effort consumed than the dynamic critiquing recommender with system-proposed critiques. Additionally, the former is more likely to inspire users' confidence of their choice and promote their intention to purchase and return to the agent for future use.

Introduction

As online e-commerce has evolved to its second generation where products are becoming more complex, with higher financial risk and increasingly descriptive features, the task of locating a desired choice appears to be too daunting for the average customer. Thus, more effort has been made to develop intelligent agents to assist users in making an informed and accurate decision. As a result, critiquing-based recommender agents have emerged and been broadly recognized as an effective feedback mechanism guiding users to find their ideal products.

The critiquing-based recommender agent simulates an artificial salesperson that recommends options based on users' current preferences and then elicits users' feedback in the form of critiques such as "I would like something cheaper" or "with faster processor speed". These critiques help the agent improve its accuracy in predicting users' needs in the next recommendation cycle. For a user to finally reach her ideal product, a number of such cycles are often required. Due to the fact that people are unlikely to state all of their preferences up front, especially for products that are unfamiliar to them, the critiquing agent is an effective way to help users incrementally construct their preference model and refine it as they see more options.

To our knowledge, the critiquing idea was first mentioned in RABBIT systems (Williams and Tou 1982) as a new interface paradigm for formulating queries to a database. In recent years, the critiquing-based systems have 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 recommendation (termed system-proposed critiquing in this paper). This approach has been adopted in FindMe systems (Burke, Hammond, and Young 19 97) and the more recently proposed *dynamic critiquing* agents to create compound critiques (McCarthy et al. 2005).

An alternative approach has focused on showing examples and stimulating users to make self-motivated critiques (termed user-motivated critiquing), such as the unit critiques employed by the dynamic critiquing system. Our *example critiquing* agent is a purely user-motivated critiquing system, since it allows users to freely combine unit and compound critiques, termed simple and complex tradeoff navigations in (Pu and Kumar 2004). It has been shown that example critiquing systems enable users to achieve much higher decision accuracy, mainly due to the tradeoff support that such systems provide, relative to non critiquing-based systems such as a ranked list (Pu and Kumar 2004, Pu and Chen 2005).

Critiquing is the main intelligent component in both types of recommenders. Evaluating how respective systems' interaction design succeeds in motivating users to benefit from this functionality is highly relevant to the artificial intelligence (AI). We thus have decided to compare the example critiquing agent with the system-proposed critiquing agent. We also believe that by evaluating both agents side by side, we could potentially improve some specific aspects of the interaction design of both approaches. We have chosen the dynamic critiquing system as the representative of system-proposed critiquing systems because its advantages over other similar systems have been established (McCarthy et al. 2005).

The contribution of this work is therefore an in-depth within-subjects user study comparing the performance of the user-motivated *example critiquing* and system-proposed *dynamic critiquing* systems. Selecting the criteria for evaluation is a crucial issue. According to (Bettman, Johnson, and Payne 1990), individuals typically settle for imperfect accuracy of their decisions in return for a reduction in effort consumed. However, earlier research also indicated that online users would be willing to make more effort if they perceived more benefits from the

decision aids (Spiekermann and Parachiv 2002). Therefore, we were interested in investigating how much accuracy users could achieve with the two critiquing agents, and the corresponding effort they were willing to expend. More specifically, the two systems were to be evaluated by users' objective performance (in terms of their decision accuracy, task completion time, and interaction effort) and subjective perceptions (in terms of their perceived cognitive effort, decision confidence, and trusting intentions).

In the following sections, we will first introduce the differences between the two critiquing-based recommender agents. Then we will describe our evaluation criteria and experiment design in more detail, followed by an analysis of the results and discussion. Finally, we will conclude our work and indicate its future direction.

Critiquing-based Recommender Agents

The differences between the example critiquing and dynamic critiquing agents can mainly be clarified based on the following four dimensions.

Critiquing Generation

Users' interaction with critiquing-based recommenders usually starts with specifying an initial set of preferences and then obtaining a list of recommendations computed based on the initial preferences. At this point, the critiquing agent will stimulate users to make critiques on the recommendations and use these critiques to recommend solutions closer to users' final target in the following cycles. Therefore, the critical concern of the critiquing agent is the generation of critiques and the manner in which they will be presented to users.

As mentioned above, there are principally two approaches to the generation of critiques. The *system-proposed critiquing* approach generates critiques according to its knowledge of the product domain (also called assisted browsing in FindMe systems). For example, the RentMe (Burke, Hammond, and Young 1997) accompanied one suggested apartment with several static critiques, e.g. cheaper, bigger, and nicer. In addition to the use of so-called unit critiques that constrain a single feature at a time, the dynamic critiquing method (McCarthy et al. 2005) employs a set of compound critiques. The latter are dynamically generated by discovering the recurring sets of unit differences between the current recommended item and the remaining cases using the Apriori algorithm. As an example, one compound critique can be "Different Manufacture, Lower Resolution and Cheaper" (see Figure 1).

As a different critiquing mechanism, the *user-motivated critiquing* approach does not propose pre-computed critiques, but provides a facility to motivate users to identify a single or a set of features to improve or compromise by themselves. In our newest version of the example critiquing interface, a "Value Comparison" button, highlighted with an explanation "Find similar products with better values", was located along with each recommended item so that users could click on it to activate the critiquing panel. In the critiquing panel (see Figure 2), three radio buttons are next to each feature, respectively under "Keep" (default), "Improve" and "Take any suggestion", thus allowing users to critique one feature by improving the feature's current value or accepting a compromised value suggested by the system. This critiquing interface also allows users to combine critiques on multiple features simultaneously.

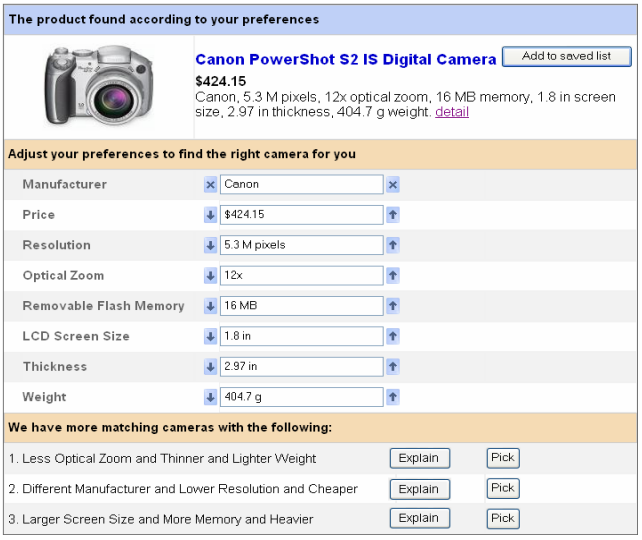

Figure 1. The dynamic critiquing interface (modified for a consistent "look" with the example critiquing interface).

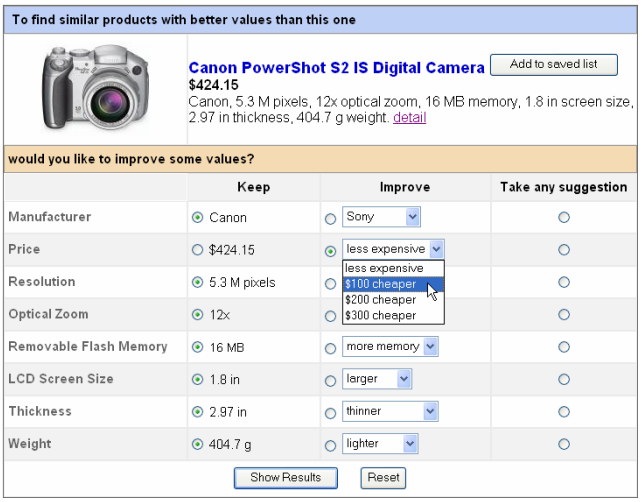

Figure 2. The example critiquing interface.

Critiquing Modality

We identify three types of modality for the critiques that a user is likely to make. The first one is the *similarity-based critiquing* such as "Find some camera similar to this one." This type of feedback is called preference-based feedback in (Smyth and McGinty 2003), and has been regarded as

the least demanding approach in terms of user effort, domain expertise and interface complexity. In the example critiquing interface, users can perform this similarity-based critiquing by keeping all current values (the default option "Keep") and clicking on the "Show Results". The second is the *quality-based critiquing* such as "Find a similar camera, but cheaper." This type of critiquing is suitable for users who desire feature improvement, but are unable to specify the exact amount to be improved. It was enabled in the example critiquing interface by an option, e.g. "less expensive", in the drop down menu under the "Improve" column. Finally, there is the *quantity-based critiquing* such as "Find something similar to this camera, but at least $100 cheaper." When users have concrete value preferences, this kind of critiquing would be more efficient for them to filter out all irrelevant items. Options, like "$100 cheaper" in the pull-down menu, facilitate this type of critiquing in the example critiquing interface (see Figure 2).

The FindMe and dynamic critiquing agents mainly focus on proposing quality-based critiques, for example, "cheaper," "bigger," or "Different Manufacture, Lower Resolution and Cheaper." These critiques are pre-generated based on the systems' product knowledge and the qualitative differences between items. Dynamic critiquing systems actually viewed their critiques as a compromise between the detail provided by value elicitation and the ease of feedback associated with preference-based methods (McCarthy et al. 2005).

Critiquing Unit

The third dimension characterizing the critiquing agent is the minimum unit allowed for users to critique simultaneously. The traditional system-proposed critiquing agent, such as FindMe (Burke, Hammond, and Young 1997), concentrated on stimulating users to express critique over a single feature at a time, called *unit critiques* in (McCarthy et al. 2005). Dynamic critiquing, on the other hand, presents combinations of critiques, i.e. *compound critiques*, to users as feedback options. The total number of recommendation cycles was shown to decrease from 29 to 6 when users actively selected compound critiques. However, no experiments have been performed so far on how dynamic critiquing improves decision accuracy, which is another fundamental criterion for recommender systems.

The user-motivated critiquing agent does not limit the critiques a user can manipulate during each cycle since users are essentially building critiques on their own. The focus here is to assist users in making tradeoffs, which is a process shown to improve decision accuracy (Pu and Chen 2005). By nature, the tradeoff navigation involves finding products with more optimal values on one or several attributes, while accepting compromised values for other attributes. That is why the example critiquing interface prompts users to "Improve" some feature values, while also enabling them to make compromises on other features by "Take any suggestion". The compound critiques proposed by the dynamic critiquing agent can also be regarded as tradeoff suggestions, such as "Different Manufacture, Lower Resolution and Cheaper" that improve on the price but make sacrifices on the manufacturer and resolution. However, these tradeoffs are predetermined for the user, and thus may not be acceptable for all users.

Critiquing Coverage

Finally we discuss the coverage of example products to be presented to users after each critiquing process. The example critiquing agent displays 7 items to users during each recommendation cycle (see (Faltings, Torrens, and Pu 2004) for the optimal number of solutions to display). In the first cycle, these items match closest to users' initial preferences. After the critiques have been specified, the system generates 7 tradeoff alternatives to the current candidate, which maximally satisfy users' critiques. The searching algorithm is based on the weighted additive sum rule (WADD) from multi-attribute utility theory. Users' preferences are structured as a set of (attribute value, weight) pairs. A major reason to show multiple alternatives is to facilitate products comparison (see the importance of comparison matrix in (Haubl and Trifts 2000)).

The dynamic critiquing agent displays one item after each critiquing, which not only satisfies users' critiques, but also is the most similar to previous recommendation. This simple display strategy has the advantage of not overwhelming users with too much information, but it bears the risk of engaging users in longer interaction cycles.

User Evaluation

Evaluation Criteria

The example critiquing and dynamic critiquing agents were evaluated in a comparative user study. We were interested in knowing how often people applied user-motivated and system-proposed critiques in both systems, and how differently users performed when using the two critiquing agents to make a decision. The user's performance was concretely evaluated in terms of the following aspects.

Decision Accuracy. The foremost criterion of evaluating a recommender agent should be the actual decision accuracy it enables users to eventually achieve. In our experiment, this criterion was quantitatively measured by the fraction of participants that switched to a different, better option when they were asked to view all alternatives in the database. A lower switching fraction means that the interface allows higher decision accuracy since most of users are able to find their target choice with it. Contrarily, a higher switching fraction implies that the recommender is not accurate in predicting what users want. For expensive products, inaccurate tools could cause both financial damage and emotional burden to the decision maker. This method was also applied by researchers in

marketing science to measure decisions (Haubl and Trifts 2000).

Decision Effort. Another important criterion is the amount of decision effort users expend to make the choice. We not only measured how much objective effort users actually consumed on the two interfaces based on their task completion time and interaction effort, but also measured their perceived cognitive effort ("How much effort do you perceive of processing information to arrive at the decision?"), which we hope would indicate the amount of subjective effort people exerted..

Confidence. The third criterion is users' confidence in their decision ("How confident are you that the product you just chose is really the best choice for you?"). In addition, we also measured their trusting intentions (Grabner-Kräuter and Kaluscha 2003) in terms of intention to purchase ("Do you intend to purchase the product that you just chose if given the opportunity?"), and intention to return ("Do you intend to return to the recommender agent in the future to search for a product?"). These factors basically reveal users' subjective opinion on the agent.

Materials and Participants

Both the example critiquing (henceforth EC) and dynamic critiquing (henceforth DC) agents were developed for two product catalogs, resulting in a 2x2 configuration. The tablet PC catalog comprises 55 products, each described by 10 main features (manufacturer, price, processor speed, weight, etc.). 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 recommenders are identical with a preference specification page to get users' initial preferences. Then in the example critiquing interface, the top 7 most matching items will be returned. If a user finds her target choice among the 7 items, she can proceed to check out. However, if she likes one product (called the reference product) but wants something improved, she can come to the critiquing interface (by clicking the "Value Comparison" button along with the recommendation) to produce one or multiple critiques based on the reference product (see Figure 2). Afterwards, a new set of items will be recommended and the user can compare them with the reference product. In the dynamic critiquing interface, the item that most closely matches users' initial preferences is shown in the beginning, accompanied by a set of self-motivated 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 both agents' 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 (5 females) volunteers participated in the user evaluation for a reward that costs around 10 CHFs. They are from 13 different countries (Switzerland, USA, China, etc.) and have different educational backgrounds (high school, bachelor, master and doctor). Among the participants, 29 have online shopping experiences.

Experiment Design and Procedure

The user study was conducted in a within-subjects design. Each participant evaluated the two critiquing-based recommenders one after the other. In order to avoid any carryover effect, we developed four (2x2) experiment conditions. The manipulated factors are recommenders' order (EC first vs. DC first) and product catalogs' order (digital camera first vs. tablet PC first). Participants were evenly assigned to one of the four experiment conditions, resulting in a sample size of 9 subjects per condition cell.

The experiment was implemented as an online procedure containing all instructions, interfaces and questionnaires. Users could easily follow it, and all of their actions were automatically recorded in the log file. The same administrator supervised the experiment for all of the participants. The main user task was to "find a product you would purchase if given the opportunity" for each product catalog with a respective recommender agent. 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 recommender. Then the recommender'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 agent. After the participant evaluated the two recommenders, a final post-question was asked about her preference over which critiquing agent she would like to use for future search.

Results Analysis

Critiquing Application

Results of our user study showed that 88.9% of the participants performed self-motivated critiques while using the example critiquing agent, and on average 49% of their critiquing cycles were used for unit critiques with the remaining 51% of cycles for compound critiques. While interacting with the dynamic critiquing agent, 83.3% of the participants applied unit critiques during average 72% of their critiquing cycles and picked system-proposed compound critiques in the remaining time. This finding indicates that most users performed around an equal amount of unit and compound critiques when they were self-motivated, but chose to make up to 44% more unit critiques ($p < 0.05$) if the compound critiques were proposed by the system. It seems that the user-motivated unit critiquing was more popular in DC, indicating that the system-proposed critiques may have a poor prediction on users' choice for compound critiques.

As for the critiquing modality, similarity-based critiquing, quality-based critiquing and quantity-based

critiquing were actively accessed by respectively 33.3%, 41.7% and 47.2% of the participants when they used EC. This indicates that many users were apt to choose other types of critiques besides quality-based critiquing, the single type of critiquing modality allowed by the dynamic critiquing agent, when they were given the opportunity to self compose critiques. It also indicates that the critiquing agent should more flexibly adapt to users who come with different degrees of preference certainty and critiquing demands.

Decision Accuracy

The decision accuracy of the example critiquing agent was shown to be significantly different ($p < 0.01$, $t = 3.39$) from it of the dynamic critiquing recommender. 86.1% of the participants actually found their target choice using EC. However, DC allowed a relative lower decision accuracy of 47.2%, since the remaining 52.8% of users switched to a different, better choice when they were given the opportunity to view all of the products in the catalog.

Figure 3 illustrates the relationship between critiquing application frequency and decision accuracy on a per cycle basis. Note that in the first cycle, the example critiquing agent already achieved 41.7% decision accuracy, resulting from 80.6% application of self-motivated critiques. Later on, EC's decision accuracy gradually increased and reached its final accuracy of 86.1% in the 8th cycle. In the dynamic critiquing interface, users made more critiques, but unfortunately did not succeed in obtaining higher decision accuracy. Its maximal accuracy of 47.2% was achieved in the 24th cycle.

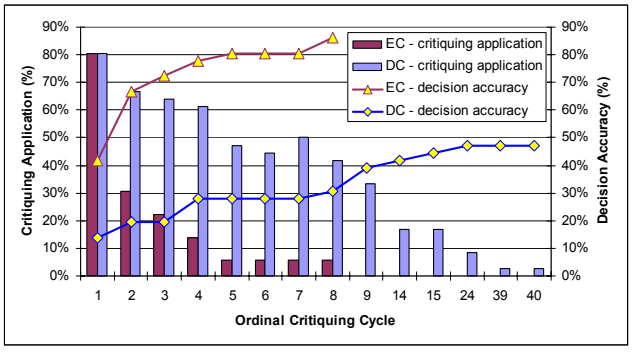

Figure 3. Relationship between decision accuracy and critiquing application on a per cycle basis.

Decision Effort

It was then interesting to know how much effort users actually expended to achieve the corresponding accuracy. As introduced before, the effort was measured by two aspects: the objective effort in terms of task completion time and interaction effort, and the subjective effort psychologically perceived by users.

The average task completion time was 4.2 minutes with EC versus 3.9 minutes with DC, but this slight difference is not significant ($p = 0.4$, $t = 0.84$). The interaction effort was further detailed as follows:

- **Critiquing effort** refers to how many times users consulted with the critiquing agent to refine their preferences. Results showed that the participant was on average involved in 2.1 critiquing cycles with EC, compared to 7.6 cycles with DC ($p < 0.001$).
- **The number of products viewed** is however higher with EC (22 items on average versus 9 with DC, $p < 0.001$). We believe that this is mainly due to the difference between their critiquing coverage (7 items vs. 1 item returned during each recommendation cycle).
- **Consideration set size** indicates how much effort users expended in seriously comparing the items stored in their saved list for final selection. On average, more items (1.53) were put in EC's saved list, compared to 1.33 items in DC's, although the difference did not reach a significant level ($p = 0.18$).
- **Investigation of a product's detailed information** reveals the effort made to look into interesting products' detailed specification pages that provide more information than just the main features. More items were examined in detail in EC than DC (1.11 versus 0.47, $p < 0.05$)

The valuable finding is therefore that with slightly more time spent in EC than DC, users actually consumed less effort in critiquing, but viewed more than twice the number of items, seriously compared more items in their saved list, and carefully examined more items' detailed specifications.

More interesting is that although the objective effort is slightly higher with EC (except critiquing effort), users perceived it to be less demanding (with a lower cognitive effort of 2.14 versus 2.47 for DC, $p < 0.1$; see Figure 4). That is, users did not perceive that they spent more effort in viewing products and performing in-depth examination of some of them while using EC.

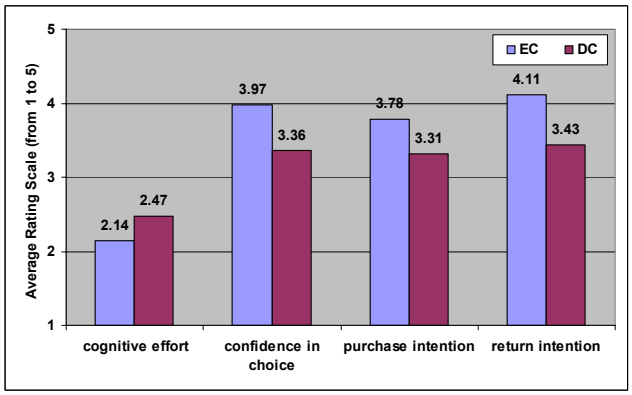

Figure 4. Subjective perceptions with EC and DC.

Confidence and Trusting Intentions

In addition, participants were more confident that they made the best choice with EC (3.97 against 3.36 on a 5-

point Likert scale, $p < 0.01$; see Figure 4), implying that they truly perceived EC to provide a higher level of decision accuracy. The confidence in choice was further proved to be significantly correlated with the participants' perceived cognitive effort (correlation = -0.464, $p < 0.01$). This means that once users experienced more accurate benefit from the recommender agent, they would likely perceive less cognitive effort consumed on it even though more objective effort was actually spent in making the choice.

Moreover, participants indicated on average a 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, $p < 0.01$) and to return to EC for future use (4.11 versus 3.43, $p < 0.001$; see Figure 4). These results imply a potential long term relationship established between the user and recommender.

Discussion

The responses to the final post-question on users' preference over which critiquing-based recommender they would use in the future show that most participants (63.9%) subjectively preferred the EC to DC. The pro arguments are that EC enables them to have more freedom in making different types of critiques, and thus provides a higher degree of control. In addition, it is more flexible and effective to compare near-satisfying products with the EC agent. Accordingly, the negative aspects of DC are that it lacks more intelligent comparison facility, it is too rigid in the proposed critiques, and it does not allow combining critiquing criteria by users themselves.

From the comments of the remaining users (36.1%) who favored DC, we can also see the advantages of DC. The users thought it was more intuitive and easier to refine preferences. They found the proposed compound critiques matched the tradeoffs that they were prepared to make. By selecting them, they accelerated their decision making.

Conclusion and Future Work

We have investigated the differences between two approaches of critiquing-based recommender agents: user-motivated and system-proposed. We described an in-depth user study evaluating the performance of the user-motivated *example critiquing* and system-proposed *dynamic critiquing* agents in terms of participants' decision accuracy, the respective interaction effort and subjective confidence. The results indicate that users can achieve a higher level of decision accuracy with less cognitive effort expended using the example critiquing recommender mainly due to its facility to enable users to freely combine unit and compound critiques. In addition, the confidence in choice made with the example critiquing agent is higher, resulting in users' increased intention to purchase the product they have found and return to the agent in the future. Combined with previous evaluations, it is possible to conclude that the example critiquing recommender agent is a more effective tool for finding complex products compared to both non critiquing-based and system-proposed critiquing systems.

Our future work includes improving system-proposed critiquing agents to more precisely predict the tradeoffs users are prepared to make, for example by using a priori data. Such improvement makes it feasible to integrate system-proposed critiques into the user-motivated recommender agent so that the hybrid system both effectively exposes the domain knowledge via the proposed critiques and allows users to freely choose unit and compound critiques.

References

Bettman, J. R.; Johnson, E. J.; and Payne, J. W. 1990. A Componential Analysis of Cognitive Effort in Choice. *Organizational Behavior and Human Decision Making Processes* 45:111–139.

Burke, R.; Hammond, K.; and Young, B. 1997. The FindMe Approach to Assisted Browsing. *IEEE Expert: Intelligent Systems and Their Applications* 12(4):32-40.

Faltings, B.; Torrens, M.; and Pu, P. 2004. Solution Generation with Qualitative Models of Preferences. *International Journal of Computational Intelligence and Applications* 20(2):246-264.

Grabner-Kräuter, S., and Kaluscha, E. A. 2003. Empirical Research in On-line Trust: A Review and Critical Assessment. *International Journal of Human-Computer Studies* 58(6):783-812.

Haubl, G., and Trifts, V. 2000. Consumer Decision Making in Online Shopping Environments: the Effects of Interactive Decision Aids. *Marketing Science* 19(1):4-21.

McCarthy, K.; Reilly, J.; McGinty, L.; and Smyth, B. 2005. Experiments in Dynamic Critiquing. In *Proceedings of the Tenth International Conference on Intelligent User Interfaces*, 175-182, New York: ACM Press.

Pu, P., and Chen, L. 2005. Integrating Tradeoff Support in Product Search Tools for E-Commerce Sites. In *Proceedings of the Sixth ACM Conference on Electronic Commerce*, 269-278, ACM Press.

Pu, P., and Kumar, P. 2004. Evaluating Example-Based Search Tools. In *Proceedings of the Fifth ACM Conference on Electronic Commerce*, 208-217, ACM Press.

Smyth, B., and McGinty, L. 2003. An Analysis of Feedback Strategies in Conversational Recommenders. In *the Fourteenth Irish Artificial Intelligence and Cognitive Science Conference (AICS 2003)*.

Spiekermann, S., and Parachiv, C. 2002. Motivating Human-Agent Interaction: Transferring Insights from Behavioral Marketing to Interface Design. *Journal of Electronic Commerce Research* 2(3):255-285.

Williams, M. D., and Tou, F. N. 1982. RABBIT: An Interface for Database Access. In *Proceedings of the ACM '82 Conference*, 83-87, ACM Press.