View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Infoscience - École polytechnique fédérale de Lausanne

A Visual Interface for Critiquing-based Recommender Systems

Jiyong Zhang
Human Computer Interaction Group, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
jiyong.zhang@epfl.ch

Nicolas Jones
Human Computer Interaction Group, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
nicolas.jones@epfl.ch

Pearl Pu
Human Computer Interaction Group, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
pearl.pu@epfl.ch

ABSTRACT

Critiquing-based recommender systems provide an efficient way for users to navigate through complex product spaces even if they are not familiar with the domain details in e-commerce environments. While recent research has mainly concentrated on methods for generating high quality compound critiques, to date there has been a lack of comprehensive investigation on the interface design issues. Traditionally the interface is *textual*, which shows compound critiques in plain text and may not be easily understood. In this paper we propose a new *visual* interface which represents various critiques by a set of meaningful icons. Results from our real-user evaluation show that the visual interface can improve the performance of critique-based recommenders by attracting users to apply the compound critiques more frequently and reducing users' interaction effort substantially when the product domain is complex. Users' subjective feedback also shows that the visual interface is highly promising in enhancing users' shopping experience.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*Human factors, Human information processing*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology*.

General Terms

Design, Human Factors, Experimentation.

Keywords

Visual interface, compound critiquing, recommender system, recommendation performance, real-user study.

1. INTRODUCTION

Many types of recommender systems have been developed in recent years to help online consumers find their desired products on e-commerce websites, from the very successful collaborative systems [16] to the more recent content-based conversational systems [1]. When the product domain is complex, very often users are not familiar with the details of each product, or may not fully understand and appreciate the trade-offs that exist between different product features. It is unlikely that users are able to input all their preferences precisely at one time. Thus the recommender systems need to interact with users so that they can construct their preferences gradually with a sequence of recommendation cycles [6]. During each cycle, one or more products are recommended based on some evolving model of the user's requirements, and the user has the opportunity to provide feedback in order to steer the recommender in the direction of the desired product.

Different forms of feedback can be used in recommender systems and they assume different degrees of domain expertise and require different levels of user effort [10]. For example, *value elicitation*, where users indicate a precise feature value – "I want a camera with 512MB of storage" – assumes that users have detailed domain knowledge and that they are willing to indicate the precise requirements on a feature by feature basis. In contrast, *preference-based* feedback asks the user to only indicate a preference for one suggestion over another [9].

In this paper we are interested in *critiquing-based recommender systems*, which adopt the form of feedback known as *critiquing* [2]. Critiquing can be viewed as a compromise between the detail provided with value elicitation and the ease of obtaining feedback associated with the preference-based methods. To critique a product a user indicates a directional change to a specific feature. For example, a digital camera shopper might ask for a camera that is *more expensive* than the current suggestion; this is a critique over the *price* feature. More recently a variation of critiquing has appeared, known as *dynamic critiquing*, which involves the automatic generation of *compound critiques* at the time of recommendation [13]. Compound critiques are collections of individual feature critiques which allow the user to indicate a richer form of feedback. For example, a user might indicate that he is interested in a digital camera with a *higher resolution* and a *lower price* than the current recommendation by selecting a *lower price, higher resolution* compound critique. Recently various methods for generating compound critiques have been proposed and their performances have been measured in some previous research [13, 8, 17, 4, 14, 15].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
EC'08, July 8–12, 2008, Chicago, Illinois, USA.
Copyright 2008 ACM 978-1-60558-169-9/08/07 ...$5.00.

In critiquing-based recommender systems, it is important to encourage users to apply compound critiques frequently. In our previous work [14, 15] we have found that most users prefer the more detailed critiquing interface, rather than a simplified one. This result shows that the design of the user interface is a very important issue for the system's performance. However, to date there has been a lack of comprehensive investigation on the impact of interface design issues for critiquing-based recommender systems.

In this paper we are seeking ways to improve the performance of critiquing-based recommender systems from the interface design level. Traditionally, compound critiques are represented textually with sentences [14, 15]. If the product domain is complex and has many features, it often requires too much effort for users to read the whole sentence of each compound critique. We believe that such textual interfaces hamper the users' experience during the recommendation process. Aiming to solve this, we propose a new *visual* design of the user interface, which represents compound critiques via a selection of value-augmented icons. We further develop an online shopping prototype system in both laptop and digital camera domains, and carry out a real-user study to compare the performance of these two designs. We hypothesize that the visual interface can attract users to apply the compound critiques more frequently and reduce the users' interaction efforts substantially compared to the traditional textual interface. Also, the visual interface can lead users to being more confident in finding their desired products. To the best of our knowledge, this is the first work on enhancing the performance of critique-based recommender systems through visual techniques on the user interface design.

The rest of this paper is organized as follows. We first provide a brief review of the related work about critiquing techniques. Then the two interface designs for critiquing-based recommender systems are introduced. Next we describe the setup of the real-user study and report the evaluation results. Finally we present the discussion and conclusions of our work.

2. RELATED WORK

Critiquing was first introduced as a form of feedback for recommender interfaces as part of the FindMe recommender systems [2, 3], and is perhaps best known for the role it played in the Entrée restaurant recommender. During each cycle Entrée presents users with a fixed set of critiques to accompany a suggested restaurant case, allowing users to *tweak* or critique this case in a variety of directions; for example, the user may request another restaurant that is *cheaper* or *more formal*, for instance, by critiquing its *price* and *style* features. In this section we provide a brief review of the critiquing techniques.

2.1 Unit Critique and Compound Critique

The simplest form of critiquing is a *unit critique* which allows users to give feedback (eg. increase or decrease) on a single attribute or feature of the products at a time[3]. It is a mechanism that gives direct control to each individual dimension. The unit critique can be readily presented as a button alongside the associated product feature value and it can be easily selected by the user. In addition, it can be used by users who have only limited understanding of the product domain. However, unit critiques are not very efficient: if a user wants to express preferences on two or more attributes, multiple interaction cycles between the user and the system are required and big jumps in the data space are not possible in one operation.

To make the critiquing process more efficient, an alternative strategy is to consider the use of what we call *compound critiques* [2, 13]. Compound critiques are collections of individual feature critiques and allow the user to indicate a richer form of feedback, but limited to the presented selection. For example, the user might indicate that they are interested in a digital camera with a higher resolution and a lower price than the current recommendation by selecting a *lower price, higher resolution* compound critique.

Obviously, compound critiques have the potential to improve recommendation efficiency because they allow users to focus on multiple feature constraints within a single cycle. Initially the compound critiques were hard-coded by the system designer resulting in the users being presented with a fixed set of compound critiques in each recommendation cycle. These compound critiques may, or may not, be relevant depending on the products that remain at a given point in time. Recently, several dynamic compound critique generation algorithms have been proposed. For example, the *Apriori* approach uses a data-mining algorithm to discover patterns in the types of products remaining, then converts these patterns into compound critiques [13]. In this paper, we adopt an alternative method called the *MAUT* approach. This approach takes the multi-attribute utility theory (MAUT) [7] to model users' preferences. Then it identifies the most suitable products for users and converts them into compound critiques. In the following we recall this method in detail.

2.2 Preference-based Compound Critiquing Generation

With the MAUT approach, in each interaction cycle the system determines a list of products via the user's preference model, and then generates compound critiques by comparing them with the current reference product. The system adaptively maintains a model of the user's preference model based on his critique actions during the interaction process, and the compound critiques are determined according to the utilities they gain.

For a given user, this approach uses the weighted additive utility function to calculate the utility of a given product $\langle x_1, x_2, ..., x_n \rangle$ as follows:

$$U(\langle x_1, \cdots, x_n \rangle) = \sum_{i=1}^{n} w_i V_i(x_i) \quad (1)$$

where n is the number of attributes that the products may have, the weight $w_i (1 \leq i \leq n)$ is the importance of the attribute i, and V_i is a value function of the attribute x_i which can be given according to the domain knowledge during the design time.

The system constructs a preference model which contains the weights and the preferred values for the product attributes to represent the user's preferences. When the user selects a compound critique, the corresponding product is assigned as the new reference product, and the user's preference model is updated based on this critique selection. For each attribute, the attribute value of the new reference product is assigned as the preference value, and the weight of each attribute is adaptively adjusted according to the dif-

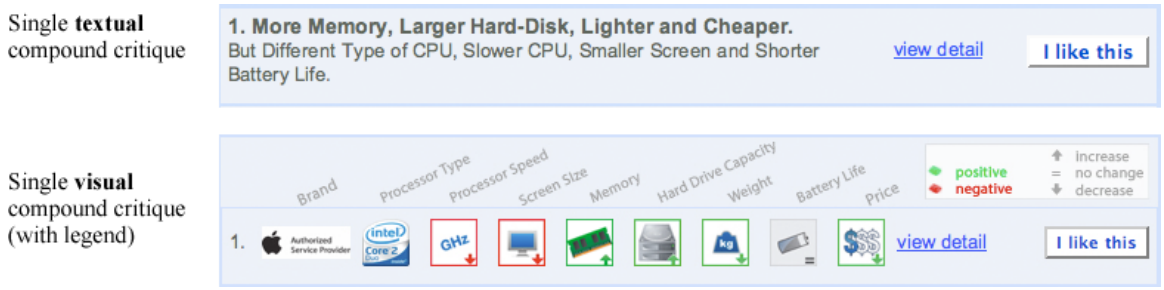

Figure 1: An example of a single compound critique from both the textual interface and the visual interface.

ference between the old preference value and the new preference value. Based on the new reference product and the updated preference model, the system is able to recommend another set of compound critiques. A more in-depth explanation of this approach to generating compound critiques is contained in [17].

3. INTERFACE DESIGN

One of the main focusses of this study is on the interface design for critiquing-based recommender systems. In [14, 15] we have implemented an online shopping system on the product domains of both digital cameras and laptops. It is designed in a way that allows users to concentrate on the utilization of both unit critiques and compound critiques as the feedback mechanism. The interface layout is composed of three main elements: a product panel, a *unit critique* panel and a *compound critique* panel. The product panel shows the current recommended product which best matches the user's preferences. In the unit critique panel, each feature is surrounded by two small buttons, which allow users to increase or decrease a value in the case of numeric features, and to change a value in categorical features such as the brand or processor type. In the compound critique panel, a list of compound critiques is shown (as textual sentences). Users can perform a compound critique by clicking the button "I like this" on its right-hand side. These three elements make up the main shopping interface and are always visible to end-users.

We are interested in getting a better perception of the role of the interface's design in the whole interaction process. We are in particular motivated by the frequent observation that people find the compound critiques too complex and admit to not actually reading all the information provided. In this context, we decide to create a visual representation of the compound critiques and to compare it with the traditional textual format through a real-user study. In the rest of this section, the design of the two interfaces are introduced in detail.

3.1 Textual Interface

The textual interface is the standard way to represent compound critiques and was used in previous work [14, 15]. As an example, a typical compound critique will say that this product has "more memory, more disk space, but less battery" than the current best match item. A direct mapping is applied from the computed numerical values of the critique, to decide if there is more or less of each feature. In accordance with [15], here we adopt the detailed interface where users are capable of seeing the product detail behind each compound critique. In addition, for each compound critique, the positive critiques are listed in bold on the first line, while the negative ones follow on the second line in a normal font-weight. Figure 1 contains an example of a single compound critique from the textual interface, and a more detailed example is given in Figure 9.

3.2 Visual Interface

The visual interface used in this study was developed in several phases. The initial idea was to propose a graphical addition in order to complement the textual critiques, but this rapidly evolved into a complete alternative to a textual representation of the critiques. Three main solutions were considered: using icons, providing a graph of the different attributes or using text-effects such as tag-clouds. The first two solutions were kept and selected to build paper prototypes. The first test revealed that the icons were perceived as being closer in meaning to the textual representation, and they were hence chosen for this study.

Icons pose the known challenge that whilst being small they must be readable and sufficiently self-explanatory for users to be able to benefit from them. One difficult task was to create a set of clear icons for both datasets. We refined them twice after small pilot-studies to make them uniform and understandable. They were then *augmented* such as to represent the critiques: the icon *size* was chosen as a mechanism to represent the change of value of the considered parameter. For each parameter of a compound critique, we know if the raw value is bigger, equal or smaller. We used this to adapt the size of the iconized object thus creating an immediate visual impression of what were the features increasing or decreasing.

Whilst designing these icons we were concerned about two major issues. First of all, it rapidly appeared that changing the size of icons was insufficiently clear or even confusing at times. The original idea was that different sizes would create an immediate visual *map* of the proposed product. Unfortunately, this scheme made small icons unreadable and had to be adjusted. Furthermore, indicating an increase in value is not always a *positive* action: an increase in weight is a negative fact (for both cameras and laptops). This is a well known issue with icon design. Secondly we rapidly understood that all the icons would have to be displayed for each compound critique, as a grid layout. The textual critiques only indicate the parameters that change, but doing so with the icons would have resulted in lines of different lengths, making them hard to compare through this alignment prob-

lem. These two potential issues lead us to further extend the icons with additional labelling.

Consequently we decided to add a token to the corner of each icon: an up arrow, a down arrow or an equal sign, to further indicate if the critique was respectively increasing, decreasing or equal to the current best match. At the same time we gave colors to the border and the token of each icon such as to indicate if the change in value was positive, negative or equal. Green was chosen for positive, red for negative and grey for the status quo. For those features without value change, the corresponding icons were shown in light gray. Thus all compound critiques had an equal number of icons and the potential alignment problem was avoided. More importantly, these lines of aligned icons form a comparison matrix and they are decision supportive: a user can quickly decide which compound critique to apply by counting the number of positive or negative icons.

During our pilot user study we found that the visual interface required from users a learning effort. Two measures were taken to tune down this effect. Firstly, a miniature legend of the icons was included at the top of the compound critique panel. Secondly, in our user study we provided an instructions page to users with explanations of the meaning of icons and some icon examples. Figure 1 provides a quick comparison of the textual compound critiques and our visual design (including legend). A more detailed example of the visual interface for compound critiques is given in Figure 10.

Table 1: Post-Stage Assessment Questionnaire

ID	Statement
S1	I found the compound critiques easy to understand.
S2	I didn't like this recommender, and I would never use it again.
S3	I did not find the compound critiques informative.
S4	I am confident that I have found the laptop (or digital camera) that I like.
S5	Overall, it required too much effort to find my desired laptop (or digital camera).
S6	The compound critiques were relevant to my preferences.
S7	I am not satisfied with the laptop (or digital camera) I found using this system.
S8	I would buy the selected laptop (or digital camera), given the opportunity.
S9	I found it easy to find my desired laptop (or digital camera).
S10	I would use this recommender in the future to buy other products.
S11	I did not find the compound critiques useful when searching for laptops (or digital cameras).
S12	Overall, this system made me feel happy during the online shopping process.

4. REAL-USER EVALUATION

We conducted a real-user evaluation to compare the performance of the two interfaces. In this section we first present the performance evaluation criteria, then we outline the setup of the evaluation and introduce the datasets and participants.

4.1 Evaluation Criteria

There are two types of criteria for measuring the performance of a critiquing-based recommender system: the objective criteria from the interaction logs and the subjective criteria from users' opinions. In this real-user evaluation we mainly concentrate on the following objective criteria: the average interaction length, the application frequency of compound critiques, and the recommendation accuracy. Participants' subjective opinions include understandability, usability, confidence to choose, intention to purchase, etc. They are obtained through several questionnaires, which will be introduced later in this section.

4.2 Evaluation Setup

For this user-study we extended the online shopping system developed in previous studies [15] such as to support both interfaces. In addition, the MAUT approach was applied to generate compound critiques dynamically in all situations. We adopted a within-subjects design of the real-user evaluation where each participant is asked to evaluate the two different interfaces in sequence and finally compare them directly. The interface order was randomly assigned so as to equilibrate any potential bias. To eliminate the learning effect that may occur when evaluating the second interface, we adopted two different datasets (laptops and digital cameras) so that the user was facing different domains each time. As a result, we had four (2×2) conditions in the experiment, depending on the factor of interface order (visual first vs. textual first) and product dataset order (digital camera first vs. laptop first). For each user, the second stage of evaluation is always the opposite of the first so that he or she may not take the same evaluation twice.

We implemented a wizard-like online web application containing all instructions, interfaces and questionnaires so that subjects could remotely participate in the evaluation. The general online evaluation procedure consists of the following steps.

Step 1. The participant is asked to input his/her background information.

Step 2. A brief explanation of the critiquing interface and how the system works is shown to the user.

Step 3. The user participates the first stage of the evaluation. The user is instructed to find a product (either laptop or camera, randomly determined) he/she would be willing to purchase if given the opportunity. The user is able to input his/her initial preferences to start the recommendation (see figure 8), and then he/she can play with both unit critiques and compound critiques to find a desired product to select. Figure 11 illustrates the online shopping system with the visual interface and the laptop dataset.

Step 4. The user is asked to fill in a post-stage assessment questionnaire to evaluate the system he/she has just tested. He/she can indicate the level of agreement for each statement on a five-point Likert scale, ranging from -2 to $+2$, where -2 means "strongly disagree" and $+2$ is "strongly agree". We were careful to provide a balanced coverage of both positive and negative statements so that the answers are not biased by the expression style. The post-stage questionnaire is composed of twelve statements as listed in table 1.

Step 5. Recommendation accuracy is estimated by asking the participant to compare his/her chosen product to the full

Table 2: Final Preference Questionnaire

ID	Questions
Q1	Which system did you prefer?
Q2	Which system did you find more informative?
Q3	Which system did you find more useful?
Q4	Which system had the better interface?
Q5	Which system was better at recommending products (laptops or cameras) you liked?
S13	I understand the meaning of the different icons in the visual interface.

Table 3: Demographic characteristics of participants

Characteristics		Users (83 in total)
Nationality	Switzerland	36
	China	13
	France	12
	Ireland	6
	Italy	4
	Other Countries	12
Age	<20	6
	20-24	30
	25-29	40
	≥30	7
Gender	female	15
	male	68
Online Shopping Experience	Never	2
	≤ 5 times	38
	>5 times	43

list of products to determine whether or not he/she prefers another product. In our practice, the datasets are relatively large, and revealing all of these products to the user at once during the accuracy test would lead the user to confusion. To deal with this, we designed the accuracy test interface to show 20 products in one page at a time, and we provided the function of allowing users to sort the products by different attributes. Such interfaces are called *Rankedlists* and have been used as baseline in earlier research such as [12].

Step 6 – 8. These are steps for the second stage of evaluation which are almost identical to the steps 3 – 5, except that this time the user is facing the system with a different interface/dataset combination (i.e. for avoiding bias).

Step 9. After completing both stages of evaluation, a final preference questionnaire is presented to the user to compare both systems he/she has evaluated. The user needs to indicate which interface (textual or visual) is preferred in terms of various criteria such as overall preference, informativeness, interface etc. The questions are listed in table 2. This final preference questionnaire also contains an extra statement ($S13$) to evaluate if the icons that we have designed are easy to understand.

4.3 Datasets and Participants

The datasets used in this experiment were updated one week before the beginning of the experiment, resulting in them containing the most recent products currently available on the market. The laptop dataset contains 610 different items. Each laptop product has 9 features: *brand, processor type, processor speed, screen size, memory, hard drive, weight, battery life, and price.* The second one is the digital camera dataset consisting of 96 cases. Each camera is represented by 7 features: *brand, price, resolution, optical zoom, screen size, thickness and weight.* Besides, each product has a picture and a detail description.

Table 4: Design of the real-user evaluation

Group	First stage		Second stage	
	Interface	Dataset	Interface	Dataset
I (20 users)	Textual	Camera	Visual	Laptop
II (20 users)	Textual	Laptop	Visual	Camera
III (23 users)	Visual	Camera	Textual	Laptop
VI (20 users)	Textual	Laptop	Textual	Camera

To attract users to participate in our user study, we set an incentive of 100 EUR and users were informed that one of those who had completed the user study would have a chance to win it. The user study was carried out over two weeks. Users participated in the user study remotely without any supervision. Finally we obtained 83 users in total who completed the whole evaluation process. Their demographic information is shown in table 3. The participants were evenly assigned to one of the four experiment conditions, resulting in a sample size of roughly 20 subjects per condition cell. Table 4 shows the details of the user study design.

5. EVALUATION RESULTS

5.1 Recommendation Efficiency

To be successful, a recommender system must be able to efficiently guide a user through a product-space and, in general, short recommendation sessions are to be preferred. For this evaluation, we measure the length of a session in terms of recommendation cycles, i.e. the number of products viewed by users before they accepted the system's recommendation. For each recommendation interface and dataset combination we averaged the session lengths across all users. It is important to remember that any sequencing bias was eliminated by randomizing the presentation order in terms of interface type and dataset.

Figure 2 presents the results of the average session lengths with different interfaces. The visual interface appears to be more efficient than the baseline textual interface. For the laptop dataset, the visual interface can reduce the interaction cycles substantially from 11.7 to 5.5, a reduction of 53%. The difference between these two results is significant ($p = 0.03$, with ANOVA test in this paper). For the camera dataset, the visual interface can reduce the average interaction cycle from 9.7 to 7.3, a reduction of 25% (not significant, $p = 0.31$).

We also look into the detail of each interaction session to see how often the compound critiques had actually been applied. Previous studies have shown that frequent usage of compound critiques is correlated with shorter sessions. Higher application frequencies would indicate that users find the compound critiques more useful. Figure 3 shows application frequency of compound critiques for both systems.

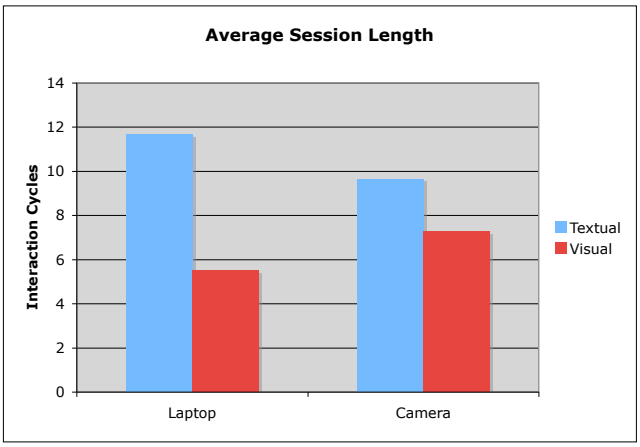

Figure 2: Average session lengths for both user interfaces.

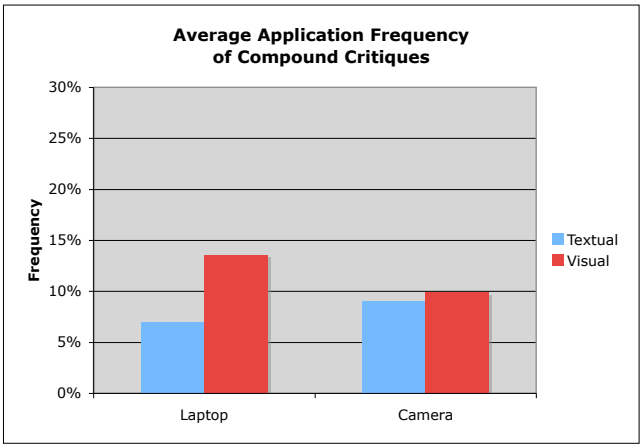

Figure 3: Average application frequency of the compound critiques for both user interfaces.

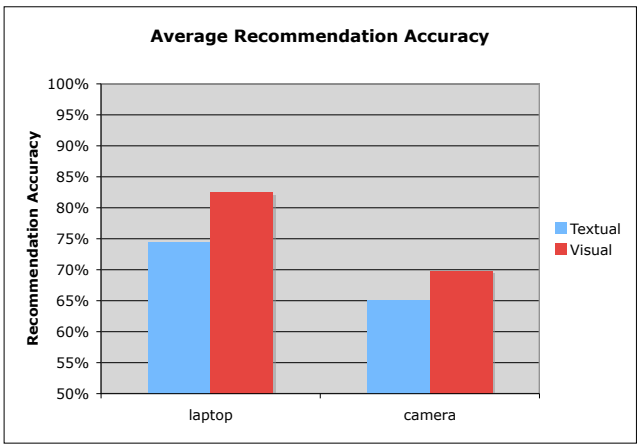

Figure 4: Average recommendation accuracy for both user interfaces.

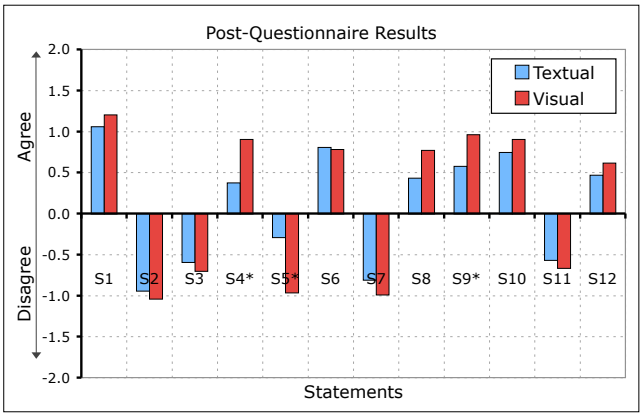

Figure 5: Results from the post-stage assessment questionnaire.

For the system with textual interface, the average application frequencies are respectively 7.0% (for laptops) and 9.0% (for cameras). For the system with visual interface, the average application frequency is nearly doubled to 13.6% for the laptop dataset (significant different, $p = 0.01$). For the camera dataset the application frequency is 9.9%, a 9.5% increase compared to the baseline textual interface (not significant, $p = 0.70$). Since for both systems we are using exactly the same algorithm to generate the compound critiques, the results shows that the visual interface can attract more users to choose the compound critiques during their decision process. Also, compared to the two systems with different datasets, it seems to show that the visual interface can be more effective when the domain is more complex.

5.2 Recommendation Accuracy

Recommenders should also be measured by the *quality* of the recommendations over the course of a session [11]. One factor for estimating recommendation quality is the recommendation accuracy, which can be measured by letting users to review their final selection with reference to the full set of products (see [12]). Formally, here we define recommendation *accuracy* as the percentage of times that users choose to stick with their selected product. If users consistently select a different product the recommender is judged to be not very accurate. The more people stick with their selected best-match product then the more accurate the recommender is considered to be.

Figure 4 presents the average accuracy results for both interfaces on both datasets. The system with textual interface performs reasonably well, achieving an accuracy of 74.4% and 65.0% on the laptop and camera datasets respectively. By comparison, the system with visual interface achieves 82.5% accuracy on the laptop dataset and 70.0% on the camera dataset, which have been increased 10% and 7% respectively. It appears that the visual interface produces more accurate recommendations. However, these improvements are not significant ($p = 0.378$ for laptop dataset, and $p = 0.648$ for camera dataset).

5.3 User Experience

In addition to the above objective evaluation results we were also interested in understanding the quality of the user experience afforded by the two interfaces. As we have mentioned earlier, a post-stage assessment questionnaire was

given when each system had been evaluated. The twelve statements are listed in table 1. A summary of the average responses from all users is shown in figure 5.

From the results we can see that both systems with different interfaces received positive feedback from users in terms of their ease of understanding, usability and interfacing characteristics. Users were generally satisfied with both systems (see $S2$ and $S7$) and found them quite efficient (see $S5$). We also noticed that overall, the visual interface has received higher absolute values than the baseline textual interface on all these statements. It is especially worth pointing out that there are three statements where the visual interface has significant improvements: S4 ($p = 0.001$), S5 ($p < 0.01$) and S9 ($p = 0.014$). These results show that the visual interface is significantly better than the textual interface in the criteria of efficiency, ease of use and leading to a more confident shopping experience.

The final preference questionnaire asked each user to vote on which interface (textual or visual) had performed better. The detail of the final preference questionnaire is shown in table 2, and the results are shown in Figure 6. The results show that overall users feel that the visual interface is better than the textual interface in all given criteria. For instance, 51% of all users may prefer the visual interface compared to 25% of whom prefer the textual interface (see Q1). Also, more than 55% of users think the visual interface is better (see Q4). Furthermore, although the two systems have exactly the same algorithm to generate compound critiques, the visual interface can enhance users' perception on the recommendation quality (see Q5). These results show that the visual interface has gained a much stronger support from end-users during the online shopping process.

In the final questionnaire we provided one extra statement (S13) for users to evaluate if the icons in the visual interface are understandable. Again users were asked to score this statement from -2 (strongly disagree) to 2 (strongly agree). The overall average score is 1.23, which shows that the icons are quite understandable and have been well designed.

We examined those users who had stated that they understood the meaning of the different icons in the visual interface(see $S13$), to see which system they preferred. The results are shown in Figure 7. We can see that amongst this subset of users, a much higher percentage of them prefer the visual interface to the textual interface. For example in $Q1$, 61% of those users voted for the visual system, while 13% of them voted for the textual one. These results suggest that if users understand the meaning of different icons, they they are even more likely to prefer the visual interface.

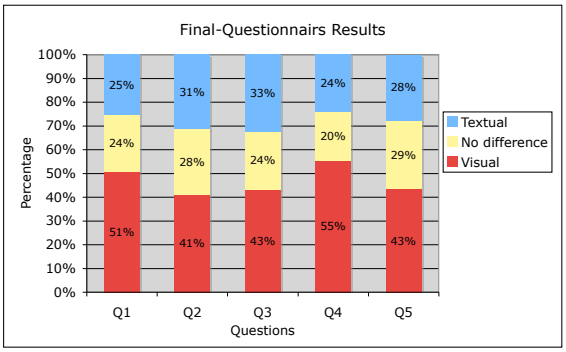

Figure 6: Results from the final preference questionnaire.

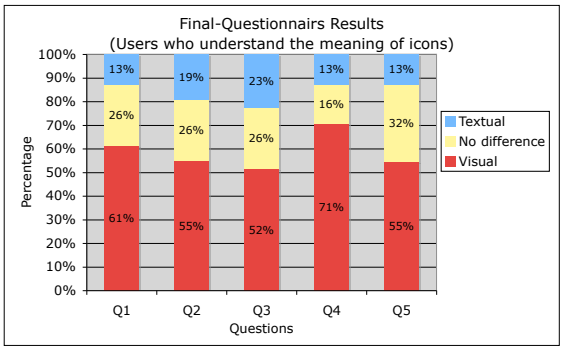

Figure 7: Results from the final preference questionnaire (Users who understand the meaning of the icons).

6. DISCUSSION

It is interesting to notice that in the user study results, while the visual interface performed better than the textual interface with both laptop and camera datasets, the visual interface has achieved higher performance improvements with the laptop dataset than with the camera dataset. The main difference between the two datasets is that the laptop assortment is more complex. It contains more products and each have more features than the cameras. When the product domain is rich, the textual interface will generate very long strings of text to describe the compound critiques, which are not easy for users to read. By comparison, the visual interface could provide an intuitive and effective way for users to make decisions (for example by simply counting the number of positive and negative icons). We believe that as the system gets more complex, the long textual descriptions become more complicated to read, and the synthetic nature of the visual solution becomes a real tool with tremendous advantages.

It is worth pointing out that the visual interface can be applied in mobile-commerce environment. On mobile internet devices, the textual interface could provide difficulties for users to read the whole compound critiques with small screens. However, the visual interface could possibly leverage the same amount of information in a much more compact manner so that the compound critiques can still be applied efficiently.

While a large proportion of users prefer the visual interface for the critiquing-based recommender system, we also noticed that there is still a small number of users who insist on the textual interface. After all, it requires some additional learning effort to understand the meaning of various icons at the beginning. A few methods we could apply to satisfy this part of users in future include adding some detailed instructions and illustrative examples to educate new users, or in our system we could provide both textual and visual interfaces and let the users choose the preferred interfaces adaptively by themselves.

During the user study several users commented on the fact that our system lacked some additional functions that currently exist in other normal websites. For example some users wanted to have a flexible search function by specifying

preference values on multiple features during the interaction process. We do believe that by integrating such additional functions in the critiquing-based system, a higher overall satisfaction level can be reached. For example, it has been shown that a hybrid system is able to achieve higher overall performance [5]. However, in this user study, the main purpose was to learn about the performance of critiquing techniques for recommender systems. Our current system was deliberately designed to exclude those functions in order to make sure that users would focus on the function of unit critiquing and compound critiquing that had been automatically recommended by the system. It will be our future work to find proper ways to integrate more functions into the current critiquing-based recommender system.

7. CONCLUSIONS

User interface design is an important issue for critiquing-based recommender systems. Traditionally the interface is *textual*, which shows compound critiques as sentences in plain text. In this paper we propose a new *visual* interface which represents various critiques by a set of meaningful icons. We developed an online web application to evaluate this new interface using a mixture of objective criteria and subjective criteria. Our comparative real-user study showed that the visual interface is more effective than the textual interface. It can significantly reduce users' interaction efforts and attract users to apply the compound critiques more frequently in complex product domains. Users' subjective feedback also showed that the visual interface is highly promising in enhancing users' shopping experience.

8. ACKNOWLEDGMENTS

Funding for this research was provided by the Swiss National Science Foundation under grant 200020-111888 and 200021-116454. We thank James Reilly for co-developing the earlier versions of the real-user evaluation platform.

9. REFERENCES

[1] D. Aha, L. Breslow, and H. Munoz-Avila. Conversational case-based reasoning. *Applied Intelligence*, 14:9–32, 2001.

[2] R. Burke, K. Hammond, and B. Young. Knowledge-based navigation of complex information spaces. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 462–468. AAAI Press, 1996. Portland, OR.

[3] R. Burke, K. Hammond, and B. Young. The FindMe approach to assisted browsing. *IEEE Expert*, 12(4):32–40, 1997.

[4] L. Chen and P. Pu. Evaluating critiquing-based recommender agents. In *AAAI*. AAAI Press, 2006.

[5] L. Chen and P. Pu. Hybrid critiquing-based recommender systems. In *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*, pages 22–31, New York, NY, USA, 2007. ACM Press.

[6] B. Faltings, P. Pu, M. Torrens, and P. Viappiani. Designing example-critiquing interaction. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 22–29, Island of Madeira (Portugal), January 2004. ACM.

[7] R. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley and Sons, New York, 1976.

[8] K. McCarthy, J. Reilly, L. McGinty, and B. Smyth. Experiments in dynamic critiquing. In *Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI '05)*, pages 175–182. ACM Press, 2005. San Diego, CA, USA.

[9] L. McGinty and B. Smyth. Evaluating preference-based feedback in recommender systems. In *Proceedings of the Thirteenth Irish International Conference on Artificial Intelligence and Cognitive Science (AICS-2002)*, volume 2464 of *Lecture Notes in Computer Science*, pages 209–214. Springer, 2002. Limerick, Ireland.

[10] L. McGinty and B. Smyth. Tweaking critiquing. In *Proceedings of the Workshop on Intelligent Techniques for Personalization as part of The Eighteenth International Joint Conference on Artificial Intelligence, (IJCAI-03)*, pages 20–27, 2003. Acapulco, Mexico.

[11] D. McSherry. Similarity and compromise. In K. Ashley and D. Bridge, editors, *Proceedings of the 5th International Conference on Case-Based Reasoning (ICCBR 2003)*, volume 2689 of *Lecture Notes in Computer Science*, pages 291–305. Springer, 2003. Trondheim, Norway, June 23-26, 2003.

[12] P. Pu and P. Kumar. Evaluating example-based search tools. In *Proceedings of the ACM Conference on Electronic Commerce (EC'04)*, pages 208–217, New York, USA, 2004.

[13] J. Reilly, K. McCarthy, L. McGinty, and B. Smyth. Dynamic critiquing. In *Proceedings of the Seventh European Conference on Case-Based Reasoning (ECCBR-04)*, volume 3155 of *Lecture Notes in Computer Science*, pages 763–777. Springer, 2004. Madrid, Spain.

[14] J. Reilly, J. Zhang, L. McGinty, P. Pu, and B. Smyth. A comparison of two compound critiquing systems. In *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*, pages 317–320, New York, NY, USA, 2007. ACM Press.

[15] J. Reilly, J. Zhang, L. McGinty, P. Pu, and B. Smyth. Evaluating compound critiquing recommenders: a real-user study. In *EC '07: Proceedings of the 8th ACM conference on Electronic commerce*, pages 114–123, New York, NY, USA, 2007. ACM Press.

[16] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pages 175–186, Chapel Hill, North Carolina, 1994. ACM.

[17] J. Zhang and P. Pu. A comparative study of compound critique generation in conversational recommender systems. In *Proceedings of the Fourth International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2006)*, volume 4018 of *Lecture Notes in Computer Science*, pages 234–243. Springer, 2006. Dublin, Ireland, June 21-23, 2006.

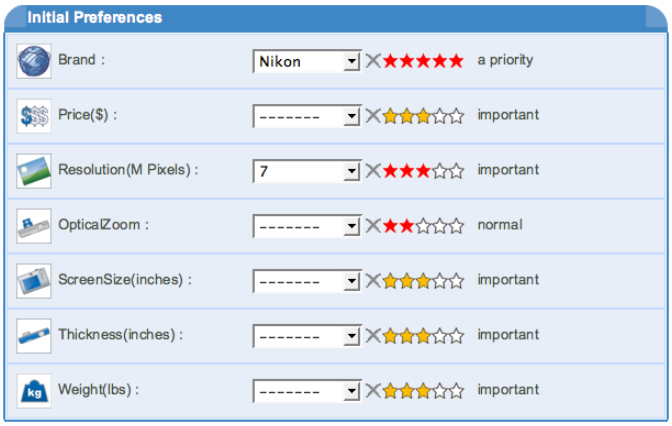

Figure 8: Screenshot of the interface for initial preferences (with digital camera dataset).

Figure 9: Screenshot of the interface for textual compound critiques (with laptop dataset).

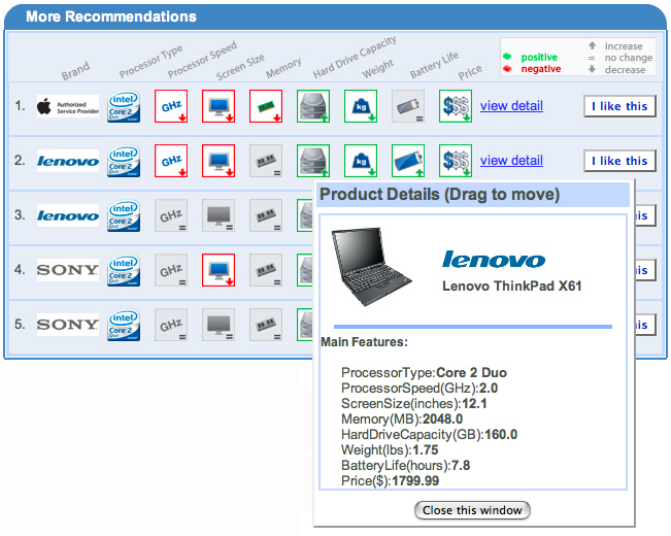

Figure 10: Screenshot of the interface for visual compound critiquing (with laptop dataset).

Figure 11: Screenshot of the visual interface for the online shopping system (with laptop dataset).